Podstawy klasycznej (nienadzorowanej) analizy danych

J. Tabor

22 grudnia 2017

Spis treści

ZASADY: Ocena końcowa będzi obliczana jako średnia z ćwiczeń, egzaminu pisemnego i ustnego. Osoby które średnią z ćwiczeń i pisemnego będą miały minimum 4, są zwolnione z ustnego.

Egzamin pisemny będzie się składał z 3 godzinnych kolokwiów (na wykładzie).

I wyklad

- zmienna losowa - zakres, normalizacja do zakresu 0-1 - średnia, mediana - odchylenie standardowe

- wektor losowy - srednia - funkcja kosztu - wyprowadzenie średniej, mediany jednowymiarowej - uogólnienie na wielowymiarową

- współczynnik korelacji - macierz kowariancji

- histogram - motywacja

zadania 1. wzór na zmianę średniej, etc pod wpływem liniowych transformacji 2. normalizacja tak by średnia zero, odchylenie standardowe 1 3. policzyć współczynnik korelacji, macierz kowariancji, etc

PLANY: dokończyć histogram, gęstość normalizacja więcej-wymiarowa

Rozdział 1

Dane skalarne

1.1 Podstawowe charakterystyki danych

Zaczniemy od przypomnienia podstawowych charakterystyk dla danych X (czyli próbki). Najbardziej podstawową jest zakres danych, czyli minimalny przedział zawierający dane:

$$[\min X, \max X].$$

Tych wartości używa się w jednej z naturalnych form preprocessingu danych jest normalizacja. Tego typu preprocessing zazwyczaj wykonuje się osobno dla każdej współrzędnej, w celu zrównoważenia wagi różnych współrzędnych w danych (pomaga w metodach klasyfikacji). Często stosuje się normalizację względem zakresu $X \to Y$, gdzie:

$$x_i \to y_i = \frac{x_i - \min X}{\max X - \min X}.$$

Wtedy zakres danych zostaje przerzucony do przedzialu $[0, 1]$.

Średnia z próbki $X = (x_i)$ to

$$\mathrm{mean}X = \tfrac{1}{N}(x_1 + \ldots + x_N).$$

Średnia ma pewne minusy:

- jest bardzo czuła na błędy (pojawienie się outliersów),
- zwraca zazwyczaj wynik który może nie być reprezentowany w zbiorze danych.

Outliersy, czyli wartości oddalone mogą być spowodowane przez wiele, powodów, najsławniejszym jest zawartość żelaza w szpinaku[1]. Bardzo dobrym przykładem mogą to być zarobki w dziesięcioosobowej firmie, w które szef zarabia 12 tys, a pracownicy po 2 tys. Wtedy średnia wynosi 3 tys, a nikt nie ma takich zarobków i większość zarabia poniżej średniej, co wywołuje frustrację.

W związku z tym rozważa się drugi miernik, a mianowicie *medianę*, który oznacza wartość m dzielącą próbkę na dwie „połówki”:

$$P(X \leq m) \geq 1/2 \text{ oraz } P(X \geq m) \geq 1/2.$$

gdzie $X_{\leq m} = \{x \in X : x \leq m\}$ oraz $P(A) = \mathrm{card}A/\mathrm{card}X$. Proszę zauważyć, że względem powyższej definicji mediana jest przedziałem (w praktyce jedyność jest wtedy gdy zbiór danych

[1] Szpinak - cytowanie

ma nieparzystą ilość elementów, zaś w przypadku gdy zbiór ma parzystą liczbę elementów za medianę przyjmuje się dowolnego reprezentanta tego przedziału):

$$\text{median}(x_1, \ldots, x_N) = [x_{\lfloor (N+1)/2 \rfloor}, x_{\lceil (N+1)/2 \rceil}] \text{ o ile dane są posortowane: } x_1 \leq \ldots \leq x_N.$$

Jak łatwo widać, mediana jest reprezentowana przez realna wartość ze zbioru danych, a co więcej jest relatywnie nieczuła na outliersy. Pokażę, że oba te pojęcia są konsekwencją wyboru funkcji kosztu jaki rozpatrujemy przy zastępowaniu danych.

Gdy chcemy zmierzyć zmienność wyników, która pozwala nam sprawdzić swoje zaufanie do wyników, rozważamy odchylenie standardowe $\sigma(X)$ (pierwiastek z wariancji $\text{Var}(X)$), które mierzy średni błąd:

$$\sigma(X)^2 = \text{Var}X = \tfrac{1}{N} \sum_i (x_i - \text{mean}X)^2.$$

Przykład 1.1. Rozważmy dwie osoby, które mierzyły stół. Jedna uzyskała wyniki $X_1 = \{0.5, 1.0, 1.5\}$, a druga $X_2 = \{0.99, 1.00, 1.01\}$. Pozornie można przyjąć, że ponieważ srednie są równe, wyniki są takie same, ale oczywiście widać, że pierwsza osoba mierzyła ten stół znacznie mniej dokładnie niż druga, co dobrze pokazuje właśnie odchylenie standardowe:

$$\sigma(X_1) = \frac{1}{2\sqrt{3/2}} \approx 0.41 \text{ a } \sigma(X_2) = \frac{1}{100\sqrt{3/2}} \approx 0.008.$$

Pokażemy najprostsze własności wariancji.

Obserwacja 1.1. *Mamy*

$$\text{Var}(x_i) = \text{mean}(x_i^2) - (\text{mean}(x_i))^2.$$

Dowód. Mamy

$$\frac{1}{N} \sum_i (x_i - \text{mean}X)^2 = \frac{1}{N} \sum_i x_i^2 - 2\frac{1}{N} \sum_i x_i \text{mean}X + (\text{mean}X)^2$$

$$= \frac{1}{N} \sum x_i^2 - (\text{mean}X)^2 = \text{mean}(x_i^2) - (\text{mean}(x_i))^2.$$

□

Następna obserwacja będzie pozwalała stosować zrównoleglać liczenie wariancji (lub pozwalać na up-date on-line).

Obserwacja 1.2. *Niech* $X = (x_1, \ldots, x_N)$, $Y = (y_1, \ldots, y_K)$. *Niech* $X \cup Y = (x_1, \ldots, x_N, y_1, \ldots, y_K)$ *oraz*

$$p_X = \tfrac{N}{K+N}, p_Y = \tfrac{K}{K+N}.$$

Wtedy

$$\text{mean}(X \cup Y) = p_X \text{mean}X + p_Y \text{mean}Y \tag{1.1}$$

oraz

$$\text{Var}(X \cup Y) = p_X \text{Var}X + p_Y \text{Var}Y + p_X p_Y (\text{mean}X - \text{mean}Y)^2.$$

Dowód. Mamy z poprzedniej obserwacji oraz (1.1)

$$\mathrm{Var}(X \cup Y) = p_X \mathrm{mean} X^2 + p_Y \mathrm{mean} Y^2 - (p_X \mathrm{mean} X + p_y \mathrm{mean} Y)^2$$
$$= p_X(\mathrm{mean} X^2 - (\mathrm{mean} X)^2) + p_Y(\mathrm{mean} Y - (\mathrm{mean} Y)^2) + p_X p_Y (\mathrm{mean} X - \mathrm{mean} Y)^2.$$

□

Zadanie 1.1. *Proszę wyliczyć wzory na* $\mathrm{mean}((x_i + y_i)_i)$ *oraz* $\mathrm{mean}(cX)$.

Zadanie 1.2. *Analogicznie do normalizacji względem zakresu używa się normalizacji względem współczynników rozkładu:*

$$x_i \to y_i = \frac{x_i - \mathrm{mean}(X)}{\sigma(X)}.$$

Proszę sprawdzić, że po dokonaniu tej procedury dostajemy rozkład o średniej zero i odchyleniu jeden.

Zadanie 1.3. *Zbiór* X *zawiera* 20 *punktów o średniej* 2 *i wariancji* 6. *Dorzucono do danych liczbę* 3. *Policz nową średnią i wariancję.*

Zadanie 1.4. *Proszę policzyć medianę* $X = \{1, 2, 5, 2, 3\}$.

1.2 Funkcja kosztu

Pokażemy, że wybór średniej czy mediany jako reprezentanta danych X jest konsekwencją wyboru funkcji kosztu. Chcemy dokonać reprezentacji danych za pomocą jednego punktu (zbliżone do kompresji).

Problem 1.1. *Zastąpić grupę punktów/danych* $X = \{x_i\}_{i=1..k} \subset \mathbb{R}^d$ *za pomocą jednego* v, *tak by był minimalny błąd.*

Dwa pytania:

- co rozumiemy przez błąd?
- jak znaleźć ten jeden punkt (minimum)?

Błąd można rozumieć na wiele sposobów. Dla przykładu możemy rozpatrzeć

$$\max_i |x_i - v|.$$

Wtedy jak widzimy optymalne v to środek zakresu $[\min X, \max X]$. Zazwyczaj jednak chcemy by błąd się sumował po całym zbiorze, w związku z tym pojawiają się dwa najczęściej stosowane błędy:

$$\sum_i |x_i - v| \text{ oraz } \sum_i (x_i - v)^2 = \sum_i |x_i - v|^2.$$

Zajmiemy się najpierw tym drugim, który jest łatwiejszy w analizie (różniczkowalność). Ma natomiast tą wadę, że jest bardzo czuły na zaburzenia. Jest to tak zwany *błąd kwadratowy* (SE: squarred error) popełniany przy zastąpieniu każdego punktu z zestawu danych X przez jeden punkt v:

$$\mathrm{SE}(X, v) = \sum_i |x_i - v|^2.$$

Łatwo widać, że

$$\mathrm{SE}(X, v) = N[v^2 - 2(\tfrac{1}{N} \sum_i x_i)v + \tfrac{1}{N} \sum_i x_i^2].$$

Dygresja 1.1. *PRZYPOMNIENIE: Funkcja $ax^2 + bx + c$, gdzie $a > 0$, osiąga minimum w punkcie $-b/(2a)$. Wartość minimalna wynosi $-\Delta/(4a)$.*

W konsekwencji otrzymujemy, że minimum jest uzyskiwane dla v równego średniej:

$$v = \mathrm{mean}(X) = \tfrac{1}{N}\sum_i x_i,$$

Błąd dany przez sumę modułów. Rozważmy błąd dany przez:

$$v \to \sum_i |x_i - v|.$$

Lemat 1.1. *Załóżmy dodatkowo, że X jest posortowany, to znaczy $x_1 \leq x_2 \leq \cdots \leq x_{k-1} \leq x_k$. Rozpatrzmy funkcję*

$$f : v \to \sum_{i=1}^{k} |x_i - v|.$$

Wtedy $f'(v) = 2i - k$ dla $v \in (x_i, x_{i+1})$ (gdzie x_0 interpretujemy jako $-\infty$ a x_{k+1} jako $+\infty$).

Dowód. Zauważmy, że funkcja $v \to |x_i - v|$ ma pochodną w punkcie v równą -1 o ile $x_i > v$ i 1 o ile $x_i < v$. Oznacza to, że pochodna funkcji f w punkcie v jest równa

$$\mathrm{card}\{i : x_i < v\} - \mathrm{card}\{i : x_i > v\} = \mathrm{card}\{i : x_i < v\} - (k - \mathrm{card}\{i : x_i < v\}),$$

co daje tezę. □

Wniosek 1.1. *Przy założeniu jak wyżej (zbiór X posortowany), funkcja f ma następujące własności:*

- *k parzyste: silnie maleje na przedziale $(-\infty, x_{k/2}]$; jest stała na przedziale $[x_{k/2}, x_{k/2+1}]$; silnie rośnie na przedziale $[x_{k/2+1}, \infty)$.*
- *k nieparzyste: silnie maleje na przedziale $(-\infty, x_{(k+1)/2}]$; silnie rośnie na przedziale $[x_{(k+1)/2}, \infty)$.*

W ten sposób wyprowadziliśmy definicję mediany – jest to przedział na którym nasza funkcja f osiąga minimum.

Konkludując widzimy, że zarówno mediana jak i średnia są konsekwencjami wyboru funkcji kosztu.

Zadanie 1.5. *Policzyć średnią i medianę dla 3 lub więcej liczb, drastycznie zaburzyć jedną – zobaczyć jaki jest wpływ zaburzenia (wartości oddalonych - outliersów) na średnią i medianę. Proszę sformułować odpowiedni wniosek dla mediany.*

Zadanie 1.6. *Znamy liczność zbioru X, i jego średnią i wariancję. Mamy dane v. Wylicz*

$$\mathrm{SE}(X, v).$$

1.3 Histogram

Dużo więcej informacji niesie nam histogram, który pozwala w miarę dobrze graficznie pokazać częstość występownia danych. Idea jest bardzo prosta – dzielimy zbiór liczb rzeczywistych na rozłączne pudłka [odcinki] jednakowej długości, i zliczamy ilość wystąpień elementów zbioru w każdym pudełku. Można na to patrzeć jak na kompresję danych (pamiętamy tylko z pewną dokładnością)

Zadanie 1.7. *Nasz zbiór danych to* $X = \{0.1, 0.6, 0, 9, 1, 1.5, 3.1\}$. *Proszę zrobić histogram bazujący na podziale* $[0, 1), [1, 2), [2, 3), [3, 4)$.

Rozdział 2

Dane wektorowe

2.1 Współczynnik korelacji

Zajmiemy się teraz sytuacją gdy nasz zestaw danych jest na płaszczyźnie. Czyli mamy zestaw

$$X = \{(x_i, y_i)\} \subset \mathbb{R}^2.$$

Przykładowo może być badanie pacjenta, w którym mierzymy BMI i poziom cukru. Najbardziej podstawowym pytaniem, jest to czy te zmienne od siebie zależą (w przypadku pytania o BMI i poziom cukru oczywiście tak jest). Pytanie o niezależność jest trudne (jeszcze się nim zajmiemy), i choć teoretycznie możemy go rozpatrywać, nie ma dobrych praktycznych współczynników które się stosuje. W związku z tym zajmujemy się prostszym i bardziej zrozumiałym pytaniem o zależność liniową między zmiennymi (współrzędnymi wyniku):

$$y_i \approx a_1 x_i + b_1 \text{ lub dualnie } x_i \approx a_2 x_i + b_2.$$

Czyli czy

$$y = a_1 x + b \text{ lub } x = a_2 y + b \text{ gdzie } x = (x_1, \ldots, x_N), y = (y_1, \ldots, y_N) \in \mathbb{R}^N. \tag{2.1}$$

Chcemy sprawdzić, czy powyższe wektory są w zależności liniowej. Trochę przeszkadza b, więc przesuwamy do zera (odejmując od obu średnią) rozpatrując wektory

$$\hat{x} = x - \text{mean}x = (x_i - \text{mean}x), \hat{y} = y - \text{mean}y = (y_i - \text{mean}y).$$

Łatwo wtedy sprawdzić, że (2.1) jest równoważne stwierdzeniu, że wektory v_1, v_2 są współliniowe:

$$\hat{y} = a_1 \hat{x} \text{ bądź } \hat{x} = a_2 \hat{y}.$$

Aby wyprowadzić, korelację, indeks który bada zależność liniową między zmiennymi, będziemy potrzebowali następujące przypomnienie z algebry liniowej.

Dygresja 2.1. *Załóżmy, że mamy dwa wektory $x, y \in \mathbb{R}^N$. Chcemy umieć sprawdzić, czy są one współliniowe, czyli czy istnieje α_1 bądź (równoważnie) α_2 takie*

$$y = a_1 x \textit{ lub } x = a_2 y.$$

Dodatkowo chcemy, aby indeks który to mierzy zwracał nam też informację, jak blisko jesteśmy współliniowości (a nie dla przykładu 1 jeżeli współliniowe, a zero jak nie).

Powszechnie stosowany indeks do mierzenia tej współliniowości jest określony przez kąt (a precyzyjniej jego cosinus) między wektorami v, w. Otóż wektory są współliniowe, jeżeli kąt pomiedzy nimi jest równy 0 bądź π (czyli jego cosinus to ± 1). Im dalej od kąta zero, tym mniejsza jest współliniowość, a najmniejsza jest dla kąta $\pi/2$ (cosinus kąta wtedy wynosi zero), kiedy wektory są prostopadłe. Jak wiemy, cosinus kąta można policzyć dzieląc iloczyn skalarny przez iloczyn długości wektorów:

$$\cos(\angle x, y) = \frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}.$$

Przypominam, że długość wektora (norma), to jego odległość od zera i z tw. pitagorasa wynosi $\|x\| = \sqrt{x_1^2 + \ldots + x_N^2} = \langle x, x\rangle$ a $\langle x, y\rangle = x_1 y_1 + \ldots + x_N y_N$.

Tak więc sprawdzenie zależności liniowej sprowadza się do wyliczenia

$$\rho = \frac{\sum_i (x_i - \mathrm{mean}x)(y_i - \mathrm{mean}y)}{\sqrt{\sum_i (x_i - \mathrm{mean}x)^2}\sqrt{\sum_i (y_i - \mathrm{mean}y)^2}} = \frac{\mathrm{cov}(x, y)}{\sigma(x)\sigma(y)},$$

gdzie przez $\mathrm{cov}(x, y)$ oznaczamy uśredniony iloczyn skalarny pomiędzy $(x_i - \mathrm{mean}x)$ i $(y_i - \mathrm{mean}y)$:

$$\mathrm{cov}(x, y) = \frac{1}{N}\sum_i (x_i - \mathrm{mean}x)(y_i - \mathrm{mean}y).$$

Zadanie 2.1. *Zakładamy, że x, y spełniają* (2.1)*. Kładziemy*

$$\hat{x} = x - \mathrm{mean}x = (x_i - \mathrm{mean}x), \hat{y} = y - \mathrm{mean}y = (y_i - \mathrm{mean}y).$$

Pokaż, że

$$\hat{y} = a_1\hat{x} \text{ bądź } \hat{x} = a_2\hat{y}.$$

Zadanie 2.2. *Proszę wyliczyć korelacje między pierwwszą i drugą współrzędną dla a) $X = \{(l, 2l + 1)\}$, $l = 1..10$, b) $X = \{(\cos(2\pi k/6), \sin(2\pi k/6)) : k = 0..5\}$.*

2.2 Dane wektorowe

Zajmiemy się teraz rozpatrzeniem podstawowych współczynników opisujących dane wektorowe X w $\mathbb{R}^D$.

Najprostszym i najczęściej stosowanym jest oczywiście średnia:

$$\mathrm{mean}X = \frac{1}{N}\sum_i x_i,$$

którą się definiuje analogicznie jak w przypadku danych jednowymiarowych. Oczywiście, jeżeli $X = (X_1, \ldots, X_D)$ (czyli X^l oznacza l-tą współrzędną X), to

$$\mathrm{mean}X = (\mathrm{mean}X_1, \ldots, \mathrm{mean}X_D). \tag{2.2}$$

Pokażemy, że analogicznie jak w przypadku jednowymiarowym średnia minimalizuje funkcję kosztu będącą sumą kwadratów norm:

$$\mathrm{mean}X = \operatorname*{argmin}_v \mathrm{SE}(X, v) \text{ gdzie } \mathrm{SE}(X, v) = \sum_i \|x_i - v\|^2.$$

Jest to bezpośredni wniosek z następującej oberwacji.

Obserwacja 2.1. *Własność:*

$$\mathrm{SE}(X, v) = \mathrm{SE}(X, \mathrm{mean}X) + |X| \cdot \|v - \mathrm{mean}_X\|^2,$$

wynika z bezpośredniego rozpisania wzorów:

$$\mathrm{SE}(X, v) = \sum_i \|x_i - v\|^2 = \sum_i \langle x_i, x_i\rangle - 2\langle \sum_i x_i, v\rangle + |X|\langle v, v\rangle$$

$$= \sum_i \langle x_i, x_i\rangle - 2|X|\langle \mathrm{mean}_x, v\rangle + |X|\langle v, v\rangle,$$

a podstawiając w powyższym $v = \mathrm{mean}_X$ *dostajemy*

$$\mathrm{SE}(X, \mathrm{mean}X) = \sum_i \langle x_i, x_i\rangle - |X|\langle \mathrm{mean}_X, \mathrm{mean}_X\rangle.$$

Po odjęciu otrzymujemy to co chcieliśmy.

W związku z powyższym stosujemy oznaczenie

$$\mathrm{SE}(X) = \mathrm{SE}(X, \mathrm{mean}X)$$

jako minimalny możliwy błąd popełniany przy zastąpieniu całego zbioru jednym punktem.

Dygresja 2.2. *Gdybyśmy chcieli zdefiniować analog wielowymiarowy mediany, należałoby rozważyć minima funkcji*

$$v \to \operatorname*{argmin}_v \sum_i \|x_i - v\|.$$

Okazuje się, że są efektywne metody szukania tego minimum, ale nie istnieje jawny wzór tak jak w przypadku jednowymiarowym.

Macierz kowariancji definuje się biorąc kowariancje każdych współrzędnych:

$$\mathrm{cov}X = [\mathrm{cov}(X^l, X^k)]_{lk},$$

gdzie X^l to zbiór składający się z l-tej współrzędnej X. Proszę zauważyć, że macierz kowariancji to macierz symetryczna, która na głównej przekątnej ma wariancje kolejnych współrzędnych.

Obserwacja 2.2. *Mamy*

$$\mathrm{cov}X = \frac{1}{N}\sum_i (x_i - \mathrm{mean}X)(x_i - \mathrm{mean}X)^T.$$

Dowód. Niech

$$A = \mathrm{cov}X \text{ oraz } B = \frac{1}{N}\sum_i (x_i - \mathrm{mean}X)(x_i - \mathrm{mean}X)^T.$$

Wtedy

$$A_{lk} = \mathrm{cov}(X^l, X^k) = \frac{1}{N}\sum_i (x_i^l - \mathrm{mean}(X^l))(x_i^k - \mathrm{mean}(X^k)),$$

Oczywiście

$$[vw^T]_{lk} = \begin{bmatrix} v_1w_1 & v_1w_2 & \cdots \\ v_2w_1 & v_2w_2 & \cdots \\ \cdots & & \end{bmatrix}_{lk} = v_lw_k,$$

czyli

$$B_{lk} = \frac{1}{N}\sum_i (x_i^l - \mathrm{mean}(X)^l)(x_i^k - \mathrm{mean}(X)^k).$$

Na podstawie (2.2) mamy $(\mathrm{mean}X)^l = \mathrm{mean}(X^l)$, co daje tezę obserwacji. □

Okazuje się, że suma kwadratów możemy obliczyć z kowariancji.

Wniosek 2.1. *Mamy*

$$\mathrm{SE}(X) = |X| \cdot \mathrm{tr}(\mathrm{cov}X)$$

Dowód. Korzystając z tego, że $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, mamy

$$|X| \cdot \mathrm{tr}(\mathrm{cov}X) = \mathrm{tr}(\sum_i (x_i - \mathrm{mean}X)(x_i - \mathrm{mean}X)^T)$$

$$= \sum_i \mathrm{tr}((x_i - \mathrm{mean}X)^T(x_i - \mathrm{mean}X)) = \sum_i \|x - \mathrm{mean}X\|^2.$$

□

Zadanie 2.3. *Proszą pokazać, że dla liniowego A mamy*

$$\mathrm{mean}(AX + b) = A\mathrm{mean}X + b.$$

Zadanie 2.4. *Proszę policzyć średnią i macierz kowariancji dla zbioru*

$$X = \{(1,0), (0,2), (1,2), (-1,0), (0,0)\}.$$

Następnie na podstawie macierzy kowariancji i ilości elementów wyliczyć $\mathrm{SE}(X)$.

Zadanie 2.5. *Proszą pokazać, że dla liniowego A mamy*

$$\mathrm{cov}(AX + b) = A\mathrm{cov}XA^T.$$

Zadanie 2.6. *Proszę pokazać, że*

$$\mathrm{cov}X = \frac{1}{N}\sum_i x_i x_i^T - (\mathrm{mean}X)(\mathrm{mean}X)^T.$$

Wsk.: proszę powtórzyć rozumowanie dla skalarów.

Zadanie 2.7. *Niech $X = (x_1, \ldots, x_N)$, $Y = (y_1, \ldots, y_K)$. Znamy średnie i macierze kowariancji X oraz Y. Proszę wyliczyć średnią i macierz kowariancji dla $X \cup Y = (x_1, \ldots, x_N, y_1, \ldots, y_K)$.*

Wsk.: proszę powtórzyć rozumowanie dla skalarów.

2.3 Macierze symetryczne, nieujemnie i dodatnio określone

Aby przeprowadzić normalizację dla danych wektorowych, będziemy musieli przypomnieć podstawowe informacje dotyczące macierzy symetrycznych.

Zacznijmy od przypomnienia wektora własnego i wartości własnej macierzy A: mówimy, że $v \neq 0$ jest wektorem własnym odpowiadającej wartości własnej $\lambda \in \mathbb{C}$, jeżeli

$$Av = \lambda v.$$

Załóżmy, że $V = [v_1, \ldots, v_D]$ jest bazą złożoną z wektorów własnych odpowiadających wartościom własnym $(\lambda_1, \ldots, \lambda_D)$. Ponieważ $Av_i = \lambda_i v_i$, dostajemy

$$AV = V\Lambda \text{ dla } \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D),$$

gdzie przez $\mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ oznaczam macierz diagonalną mającą na głównej przekątnej wartości $\lambda_1, \ldots, \lambda_D$.

Mówimy, że macierz A jest symetryczna, jeżeli

$$A = A^T.$$

Twierdzenie 2.1. *FAKT. Niech A będzie macierzą symetryczną. Wtedy*

- *wartości własne λ_i macierzy A są rzeczywiste,*
- *można znaleźć takie wektory własne $V = [v_1, \ldots, v_D]$ macierzy A odpowiadające wartościom własnym λ_i które tworzą bazę ortonormalną.*

Przypominam, że wektory (v_i) są ortonormalne, jeżeli

$$\langle v_i, v_j \rangle = v_i^T v_j = \delta_{ij},$$

co jest równoważne temu, że

$$V^T V = I \text{ lub } V^{-1} = V^T \text{ dla } V = [v_1, \ldots, v_D].$$

Ponieważ $AV = V\Lambda$, otrzymujemy w konsekwencji, że

$$A = V\Lambda V^{-1} \text{ lub równoważnie } A = V\Lambda V^T.$$

Dla macierzy symetrycznych możemy zdefiniować odpowiedniki dowolnych funkcji rzeczywistych (o ile są określone na wartościach własnych). W szczególności dla f takiego, że $\lambda_1, \ldots, \lambda_D \in dom\, f$ kładziemy

$$f(A) = V\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_D))V^{-1}.$$

Ważną klasę macierzy symetrycznych tworzą macierze nieujemnie i dodatnio określone:

- macierz symetryczna A jest nieujemnie określona jeżeli $x^T Ax \geq 0$ dla każdego x,
- macierz symetryczna A jest dodatnio określona jeżeli $x^T Ax \geq 0$ dla każdego $x \neq 0$.

FAKT: Można pokazać, że macierz symetryczna jest nieujemnie (dodatnio) określona wtw gdy wartości własne są nieujemne (dodatnie).

Łatwo można pokazać, że suma macierzy nieujemnie (dodatnio) określonych, jest nieujemnie (dodatnio) określona.

Zgodnie z powyższym, potęga o współczynniku s dla macierzy symetrycznej nieujemnie określonej (jeżeli $s > 0$) / dodatnio określonej (dla dowolnego s) definiowana jest wzorem

$$A^s = V\mathrm{diag}(\lambda_1^s, \ldots, \lambda_D^s)V^{-1}.$$

Wprost z powyższej definicji proszę sprawdzić, że $A^s A^t = A^{s+t}$, i w szczególności:

$$\sqrt{A} \cdot \sqrt{A} = A \text{ gdzie } \sqrt{A} = A^{1/2}.$$

Proszę zwrócić uwagę, że do policzenia ujemnych potęg potrzebujemy dodatniej określoności macierzy.

Algorytm 2.1. Algorytm liczenia pierwiastka z dodatnio określonej macierzy symetrycznej A:

1. Policz wartości własne $(\lambda_1, \ldots, \lambda_D)$ i wektory własne $V = [v_1, \ldots, v_D]$ (zakładamy, że V jest układem ortonormalnym, czyli $V^T V = I$).

2, Końcowy wzór:

$$\sqrt{A} = V\Lambda^{1/2}V^T \text{ gdzie } \Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_D^{1/2}).$$

Macierze symetryczne (nieujemnie określone) są ważne między innymi z tego powodu, że macierze kowariancji są symetryczne i nieujemnie określone.

Stwierdzenie 2.1. *Niech $\Sigma = \mathrm{cov}X$ dla $X \subset \mathbb{R}^D$. Wtedy:*

1. *Σ jest macierzą symetryczną nieujemnie określoną,*
2. *Σ jest dodatnio określoną macierzą wtw gdy $\mathrm{lin}(X - \mathrm{mean}(X)) = \mathbb{R}^D$,*
3. *jeżeli Σ nie jest dodatnio określona, to*

$$\mathrm{lin}(X - \mathrm{mean}(X)) = \mathrm{lin}(v_i : \lambda_i > 0) = \mathrm{Range}(\Sigma),$$

gdzie v_i, λ_i to kolejne wektory i wartości własne Σ.

Jeżeli zachodzi 3, to wtedy po prostu redukujemy wymiar danych zawężając się do odpowiedniej przestrzeni rozpiętej na wektorach $v_1, \ldots, v_k$ odpowiadających dodatnim wartościom własnym. W praktyce robimy to za pomocą operacji

$$\mathbb{R}^D \supset X \in x \to (\langle x - \mathrm{mean}(X), v_i\rangle)_{i=1..k} \in \mathbb{R}^k.$$

Dowód. ad 1. Pokażę, że vv^T jest macierzą symetryczną nieujemnie określoną (macierz kowariancji jako suma takich macierzy też będzie).

Korzystając z tego, że $(AB)^T = B^T A^T$ mamy

$$(vv^T)^T = (v^T)^T(v)^T = vv^T.$$

Także

$$x^T vv^T x = (x^T v)(x^T v)^T = \langle x, v\rangle^2 \geq 0 \text{ dla dowolnego } x,$$

gdyż $\langle x, v\rangle = x^T v$.

Punkty 2 i 3 pozostawiam bez dowodu. □

Zadanie 2.8. *Proszę policzyć pierwiastek z macierzy*

$$A = \begin{bmatrix} 10 & -6 \\ -6 & 10 \end{bmatrix}.$$

Zadanie 2.9. *Niech A będzie macierzą symetryczną, a V bazą ortonormalną składającą się z wektorów własnych A.*

a) Pokazać, że dla $x = x_1 v_1 + \ldots + x_D v_D \in \mathbb{R}^D$ mamy

$$x^T A x = \lambda_1 x_1^2 + \ldots + \lambda_D x_D^2,$$

b) Korzystając z a), pokazać, że macierz symetryczna jest nieujemnie (dodatnio) określona wtw gdy wartości własne są nieujemne (dodatnie).

2.4 Whitening, odległość Mahalanobisa

Jedną z możliwości preprocessingu to normalizacja po każdej z współrzędnych z osobna. Niestety, jeżeli dla przykładu dane są bardzo skupione wokół jakiejś podprzestrzeni, to taka operacja niewiele zazwyczaj zmieni. My zaś chcemy, by dane były w miarę możliwości równomiernie ułożone w przestrzeni.

Z punktu widzenia wielu metod nauczania maszynowego, najlepiej jeżeli dane są znormalizowane, czyli jeśli

1. średnia jest zero,
2. współrzędne mają odchylenie standardowe równe 1,
3. nie ma między współrzędnymi zależności liniowej.

Uzyskanie tego, by wartość oczekiwana była zero, jest łatwe - po prostu przesuwamy do środka ciężkości:

$$Y = (y_i)_i = (x_i - \mathrm{mean} X)_i.$$

Jak widać, powyższe warunki 2-3 są równoważne temu, że macierz kowariancji jest równa identyczność. Natomiast powstaje oczywiście pytanie, w jaki sposób zmodikować próbkę, by współrzędne były liniowo niezależne.

Zawęzimy się do operacji liniowych. Interesuje nas w konsekwencji następujący problem:

Problem 2.1. *Mamy dane $X \subset \mathbb{R}^D$. Czy (i ew. kiedy) można znaleźć taką macierz odwracalną A, że*

$$\mathrm{cov}(AX) = I?$$

Ponieważ $\mathrm{cov}(AX) = A\mathrm{cov}(X)A^T$, powyższe się sprowadza do:

Problem 2.2. *Niech $\Sigma = \mathrm{cov} X$ dla pewnego zbioru danych X. Czy (i ew. kiedy) można znaleźć taką macierz odwracalną A, że*

$$A\Sigma A^T = I?$$

Łatwo widać, że aby dało się odpowiedzieć na to pytanie pozytywnie, macierz kowariancji musi być odwracalna.

I teraz powstaje pytanie jak dobrać A by powstała nam w wyniku operacji po prawej stronie identyczność, co jak widać jest równoważne

$$A^T A = \mathrm{cov}X^{-1}.$$

Ponieważ kowariancja jest macierzą symetryczną, aby zagwarantować jednoznaczność zawężamy się do szukania A w klasie odwzorowań symetrycznych $A = A^T$. I wtedy jak wiemy z poprzedniej sekcji rozwiązanie jest dane przez pierwiastek:

$$A = \sqrt{\mathrm{cov}X^{-1}} = (\mathrm{cov}X)^{-1/2}.$$

Konkludując, whitening dany jest przez

$$\phi_X : x \to (\mathrm{cov}X)^{-1/2}(x - \mathrm{mean}X).$$

Z pojęciem whiteningu jest blisko powiązane pojęcie zwane *metryką Mahalanobisa*. Otóż zobaczmy, jak by wyglądało gdbyśmy mierzeli odległość dwóch punktów po whiteningu:

$$\|((\mathrm{cov}X)^{1/2}(x_i - \mathrm{mean}X)) - ((\mathrm{cov}X)^{1/2}(x_j - \mathrm{mean}X))\|^2 = (x_i - x_j)^T \mathrm{cov}X^{-1}(x_i - x_j).$$

Wprowadźmy teraz oznaczenie na normę Mahalanobisa:

$$\|x\|_\Sigma^2 = x^T \Sigma^{-1} x \text{ dla } \Sigma = \mathrm{cov}X.$$

Wtedy mamy

$$\|\phi_X(x_i) - \phi_X(x_j)\| = \|x_i - x_j\|_\Sigma.$$

Ogólnie metrykę Mahalnobisa definiuje się dla dowolnej macierzy dodatnio określonej Σ. Zauważmy, że norma Mahalanobisa jest zadana przez iloczyn skalarny Mahalanobisa dany wzorem

$$\langle x, y \rangle_\Sigma = x \Sigma^{-1} y.$$

Jak widzimy, aby policzyć odległość Mahalanobisa, mamy dwie równoważne możliwości - albo transformujemy dane, i używamy zwykłej metryki euklidesowej, albo zostawiamy dane i modyfikujemy metrykę (w uproszczeniu pierwsze podejście prowadzi do representation learning, a drugie do metric learning).

Metryki Mahalanobisa się używa domyślnie w dużej ilości problemów, gdyż oryginalna często jest zła i nieodpowiednio dopasowana do danych – jednostki nie są optymalnie ustawione (przyklad z wzrostem i butami).

Zadanie 2.10. *Niech*

$$\Sigma = \begin{bmatrix} 10 & -6 \\ -6 & 10 \end{bmatrix}.$$

Proszę policzyć $\|x\|_\Sigma$ *dla a)* $x = (1, 2)$*, b)* $x = (0, 1)$ *Uwaga: zapis* (x, y) *oznacza w zapisie wektorowym punkt* $\begin{bmatrix} x \\ y \end{bmatrix}$.

Zadanie 2.11. *Mamy*

$$X = \{(1, 0), (-1, 0), (0, 4), (0, -4)\}.$$

Proszę a) policzyć macierz kowariancji b) policzyć odległość każdego punktów od zera w metryce Mahalanobisa c) dokonać whiteningu.

Zadanie 2.12. *Metryka Mahalanobisa jest niezmiennicza na transformacje afiniczne w następującym sensie*

$$\|x_i - x_j\|_{\mathrm{cov}X} = \|\phi(x_i) - \phi(x_j)\|_{\mathrm{cov}\phi(X)}$$

dla dowolnych $x_i, x_j \in X$ *i dowolnej odwracalnej transformacji afinicznej* $\phi(x) = Ax + b$*.*

Rozdział 3

k-means

k-means jest tak naprawdę metodą kompresji/dyskretyzacji. Ale jest używany do klastrowania. Ogólny cel to znalezienie dla danego zbioru X punktów $v_1, \ldots, v_k$ i przyporządkowania[1] $j : X \to \{1, \ldots, k\}$, że kwadratowy błąd popełniany przez zastąpienie x przez $v_{j(x)}$:

$$\sum_i \|x_i - v_{j(i)}\|^2$$

jest minimalny. Funkcja j nam mówi, o numerze klastra do którego dany punkt jest przyporządkowany. Aby pokazać metodę szukania minimum, rozbijamy na dwa podproblemy - gdy mamy dane punkty v_i (i tylko pytamy się który jest najlepszy), i gdy mamy przyporządkowanie, ale szukamy punktów.

3.1 Zafiksowane centra $v_1, \ldots, v_k$

Problem 3.1. *Postawienie problemu: Mamy dany zestaw możliwych punktów których używamy do dyskretyzacji (kompresji)* $V = \{v_1, \ldots, v_k\} \subset \mathbb{R}^N$.

Chcemy znaleźć, dla zestawu danych $X \subset \mathbb{R}^N$*, przyporządkowanie punktom indeksu*

$$X \ni x \to j(x) \in \{1, \ldots, k\}$$

tak by zminimalizować całkowity (kwadratowy) błąd popełniony przy dyskretyzacji

$$\mathrm{SE}(X, j) = \sum_i \|x_i - v_{j(x_i)}\|^2.$$

Widać, że wystarczy nam się zająć tym, którym punktem ze zbioru V należy przybliżyć x, aby błąd był możliwie najmniejszy:

$$j(x) = \operatorname*{argmin}_{j \in \{1,\ldots,k\}} \|x - v_j\|.$$

Zobaczmy, że w konsekwencji oznacza to, że minimalizujemy wartość

$$\mathrm{SE}(X, \{v_1, \ldots, v_k\}) = \sum_i d^2(x_i, \{v_1, \ldots, v_k\}),$$

gdzie $d(x, V)$ to odległość x od zbioru V.

[1] UWAGA: gdy $X = (x_i)$ używam dla prostoty oznaczenia $j(i)$ na $j(x_i)$.

Diagram Voronoi. Chcemy patrzeć, gdzie wpadnie nowy punkt – podział przestrzeni. Przybliżamy x najbliższym elementem ze zbioru V. Pokażemy, że powyższe oznacza, na podstawie wcześniejszych wyliczeń, że płaszczyzna (przestrzeń) rozbija się na wielokąty (wielościany), reprezentujące zbiory punktów dla których dany element $v \in V$ jest najbliższy – to jest tak zwany *diagram Voronoi*. Wynika to z następującej obserwacji:

Obserwacja 3.1. *Rozpatrzmy punkty $v, w \in \mathbb{R}^N$. Wtedy zbiór punktów na płaszczyźnie równo odległych od v i w to jest dokładnie hiperpłaszczyzna przechodząca przez $(v+w)/2$ i prostopadła do wektora $w - v$.*

Co więcej

- *punkt x jest bliżej w o ile $\langle x - \frac{v+w}{2}, w - v\rangle > 0$;*
- *punkt x jest bliżej v o ile $\langle x - \frac{v+w}{2}, w - v\rangle < 0$.*

Dowód. Mamy

$$\begin{aligned}
&\{x \in \mathbb{R}^N : \|x - w\| < \|x - v\|\} = \{x : \langle x - w, x - w\rangle < \langle x - v, x - v\rangle\} \\
&= \{x : \langle x, x\rangle - 2\langle x, w\rangle + \langle w, w\rangle < \langle x, x\rangle - 2\langle x, v\rangle + \langle v, v\rangle\} \\
&= \{x : 2\langle x, w - v\rangle > \langle w, w\rangle - \langle v, v\rangle\} = \{x : 2\langle x, w - v\rangle > \langle w + v, w - v\rangle\} \\
&= \{x : \langle x, w - v\rangle > \langle \tfrac{w+v}{2}, w - v\rangle\} = \{x : \langle x - \tfrac{w+v}{2}, w - v\rangle > 0\}.
\end{aligned}$$

Dla równości oczywiście analogicznie dostajemy

$$\{x \in \mathbb{R}^N : \|x - w\| = \|x - v\|\} = \{x : \langle x - \tfrac{w+v}{2}, w - v\rangle = 0\} = \{x : (x - \tfrac{w+v}{2}) \perp (w - v)\},$$

co opisuje żądaną hiperpłaszczyznę. □

Czyli (na płaszczyźnie) diagram Voronoi dla dwóch punktów to dwie półpłaszczyzny oddzielone prostą rozdzielającą. Diagram Voronoi dla większej ilości punktów można zbudować przecinając odpowiednio te półpłaszczyzny, czyli dostajemy wielokąty wypukłe: Diagram Voronoi można zobaczyć w: `http://alexbeutel.com/webgl/voronoi.html`

3.2 Zafiksowana funkcja indeksująca $j : X \to \{1, \ldots, k\}$

Problem 3.2. *Postawienie problemu: Mamy daną funkcję indeksującą j. Chcemy znaleźć, dla zestawu danych $X \subset \mathbb{R}^N$, zestaw możliwych punktów których używamy do dyskretyzacji (kompresji) $V = \{v_1, \ldots, v_k\} \subset \mathbb{R}^N$ tak by zminimalizować całkowity (kwadratowy) błąd popełniony przy dyskretyzacji*

$$\mathrm{SE}(X, V) = \sum_i \|x_i - v_{j(x_i)}\|^2.$$

Pytamy się, jak przy zafiksowanej funkcji indeksującej, dobrać centra $v_1, \ldots, v_k$ aby nastąpiła minimalizacja funkcji kosztu (która w naszym przypadku oznacza błąd przybliżenia).

Niech X_l oznacza podzbiór X składający się z punktów które mają indeks l (czyli wszystkie te punkty będą przybliżane za pomocą jednej wartości):

$$X_l = \{x \in X : j(x) = l\}.$$

I teraz interesuje nas, by znaleźć taki punkt v_l, który by minimalizował

$$v_l = \operatorname*{argmin}_v \mathrm{SE}(X_l, v).$$

Ale my już wiemy jakie jest rozwiązanie! Po prostu

$$v_l = \mathrm{mean} X_l.$$

3.3 Ogólny problem

Natomiast wyobraźmy sobie, że możemy dobrać V mające k punktów dowolnie. Prowadzi nas to do

Problem 3.3. *Chcemy znaleźć, dla zestawu danych $X \subset \mathbb{R}^N$, zestaw możliwych punktów których używamy do dyskretyzacji (kompresji) $V = \{v_1, \ldots, v_k\} \subset \mathbb{R}^N$ oraz funkcję indeksującą j tak by zminimalizować całkowity (kwadratowy) błąd popełniony przy dyskretyzacji*

$$\mathrm{SE}(X, j, V) = \sum_i \|x_i - v_{j(x_i)}\|^2.$$

Okazuje się, że powyższy problem nie daje się efektywnie rozwiązać (w informatyce mówi się, że jest NP-trudny). Znajduje się więc lokalne minima tego problemu. Idea polega na szukaniu minimów lokalnych funkcji dwóch zmiennych:

IDEA. Załóżmy, że mamy skomplikowaną funkcję $s(x, y)$ dwóch zmiennych x i y, której chcemy znaleźć minimum. A przy tym, mając zafiksowane $\bar{x}$ potrafimy znaleźć minimum $y \to s(\bar{x}, y)$, oraz mając zafiksowane $\bar{y}$ potrafimy znaleźc minimum $x \to s(x, \bar{y})$. Wtedy jedna z metod minimalizacji, będzie polegała, na szukaniu tego minimum poruszając się naprzemiennie wzdłuż współrzędnych x i y:

1. Fiksujemy na początek dowolny warunek początkowy $\bar{x}$ dla x
2. kładziemy $i = 0$, $x_0 = \bar{x}$, $y_0 = \operatorname{argmin}_y s(x_0, y)$.
3. Definiujemy
$$x_{i+1} = \operatorname*{argmin}_x s(x, y_i) \text{ oraz } y_{i+1} = \operatorname*{argmin}_x s(x_{i+1}, y)$$
4. wracamy do punktu 3, o ile spadła nam istotnie wartość $f(x_{i+1}, y_{i+1})$ w stosunku do $f(x_i, y_i)$, w przeciwnym razie wychodzimy z pętli.

Są metody które szukają lokalnego rozwiązania *k-means*.
Metoda Lloyda:

1. początkowo (kładziemy $l = 0$) jako $V^l = \{v_1^l, \ldots, v_k^l\}$ wybieramy losowe/dowolne elementy zbioru X;
2. dokonujemy dyskretyzacji X za pomocą V^l, wtedy X rozdziela się nam na podzbiory X_j^l punktów które będą zastąpione (inaczej mówiąc którym najbliżej do) przez v_j^l;
3. zauważmy, że z tego co pokazaliśmy wcześniej, błąd kwadratowy zmiejeszymy, jeżeli zamiast dyskretyzacji X_j^l przez v_j^l zastąpimy go przez jego średnią, czyli kładziemy $v_j^{l+1} = E(X_j^l)$ i $V^{l+1} = \{v_1^{l+1}, \ldots, v_k^{l+1}\}$;
4. zwiększamy l o jeden, i o ile zmieniło się choć jedno v_j (w stosunku do poprzedniego kroku), skaczemy do punktu 2, w przeciwnym razie kończymy procedurę.

Widać, że powyższa procedura za każdym krokiem w sposób gwarantowany minimalizuje nam błąd kwadratowy. Nie mamy oczywiście natomiast żadnej gwarancji, że znajdziemy w ten sposób globalne minimum (aby zwiększyć szanse by tak było, zazwyczaj startuje się wielokrotnie wybierając różne punkty początkowe na start).

Inicjalizacja początkowych punktów:

- zupełnie losowo wybrane punkty z danych
- wybieramy jeden, potem następny jak najdalej, itd
- k-means++ najpierw jeden, potem następny zgodnie z rozkładem prawdopodobieństwa proporcjonalnym do kwadratu odległości

k-means++ algorytm:

1. Choose one center uniformly at random from among the data points. For each data point x, compute $D(x)$, the distance between x and the nearest center that has already been chosen.
2. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D^2(x)$.
3. Repeat Steps 2 and 3 until k centers have been chosen.

Now that the initial centers have been chosen, proceed using standard k-means clustering.

Zadanie 3.1 (Python)**.** *Napisać k-means korzystający z metody Lloyda.*

3.4 Podejście Hartigana

Potencjalnie ważne inne podejście – Hartigana (jeżeli da się zastosować, to jest szybsze i lepsze, znajduje lepsze optima). Zamiast modyfikować klastry, iteruje po kolejnych punktach (inicjalizacja każdy punkt początkowo wrzucamy do losowego klastra):

1. w każdym punkcie mamy „wajchę" którą potencjalnie przełączamy wtedy gdy po sumaryczna funkcja kosztu (w naszym przypadku suma kwadratów) się zmniejszy
2. musimy umieć szybko przeliczać jak się zmieni całościowy koszt po dołączeniu/odłączeniu jednego punktu

W takim razie zajmijmy się podejściem Hartigana. Idea: w każdym punkcie mamy „dźwignię" którą przełączamy przynależność punktu, i to pozwala nam sprawdzić gdzie się opłaca przełączyć, by maksymalnie obniżyć błąd (coś w rodzaju przewidywania przyszłości). Udaje się zastosować jedynie w tej sytuacji, gdy łatwo jest dokonać update'u energii, to znaczy potrafimy wyliczyć dla $X \cup \{x\}$ i $X \setminus \{x\}$.

Przypominam oznaczenia, $\mathrm{SE}(X, v)$ to błąd wynikający z zastąpienia wszystkich punktów ze zbioru X przez punkt v, a $\mathrm{SE}(X)$ to najmniejszy możliwy błąd realizowany przez zastąpienie wszystkich elementów z X przez średnią $\mathrm{mean}X$ (dla prostoty oznaczam przez m_X):

$$\mathrm{SE}(X, v) = \sum_i \|x_i - v\|^2 \text{ oraz } \mathrm{SE}(X) = \mathrm{SE}(X, m_X).$$

Przez $|X|$ oznaczam liczność zbioru X.

Obserwacja 3.2.

$$\mathrm{SE}(X_1 \cup X_2) = \mathrm{SE}(X_1) + \mathrm{SE}(X_2) + \frac{|X_1||X_2|}{|X_1| + |X_2|}\|m_{X_1} - m_{X_2}\|^2.$$

Dowód. Dowód wynika bezpośrednio z Obserwacji 2.1 oraz z faktu, że

$$m_{X_1 \cup X_2} = \frac{|X_1|}{|X_1| + |X_2|} m_{X_1} + \frac{|X_2|}{|X_1| + |X_2|} m_{X_2},$$

co oznacza, że

$$\mathrm{SE}(X_1 \cup X_2, m_{X_1 \cup X_2}) = \mathrm{SE}(X_1, m_{X_1 \cup X_2}) + \mathrm{SE}(X_2, m_{X_1 \cup X_2})$$

$$= \mathrm{SE}(X_1) + |X_1| \cdot \|m_{X_1 \cup X_2} - m_{X_1}\|^2 + \mathrm{SE}(X_2) + |X_2| \cdot \|m_{X_1 \cup X_2} - m_{X_2}\|^2$$

$$= \mathrm{SE}(X_1) + \mathrm{SE}(X_2) + |X_1| \cdot (\tfrac{|X_2|}{|X_1|+|X_2|})^2 \|m_{X_1} - m_{X_2}\|^2 + |X_2| \cdot (\tfrac{|X_1|}{|X_1|+|X_2|})^2 \|m_{X_1} - m_{X_2}\|^2$$

$$= \mathrm{SE}(X_1) + \mathrm{SE}(X_2) + \tfrac{|X_1||X_2|}{|X_1|+|X_2|} \|m_{X_1} - m_{X_2}\|^2.$$

□

Bezpośrednio z powyższej obserwacji dostajemy:

Obserwacja 3.3. *1. Mamy*

$$\mathrm{SE}(X \cup \{x\}) = \mathrm{SE}(X) + \frac{|X|}{|X|+1} \|x - m_X\|^2, m_{X \cup \{x\}} = \frac{|X|}{|X|+1} m_X + \frac{1}{|X|+1} x,$$

2. Oraz

$$\mathrm{SE}(X \setminus \{x\}) = \mathrm{SE}(X) - \frac{|X|}{|X|-1} \|x - m_X\|^2, m_{X \setminus \{x\}} = \frac{|X|}{|X|-1} m_X - \frac{1}{|X|-1} x.$$

Wniosek 3.1. *Jeżeli mamy klastry X_i i X_j, oraz punkt $x \in X_i$, to będzie się nam opłacało go przełączyć do j wtw gdy:*

$$\mathrm{SE}(X_i \setminus \{x\}) + \mathrm{SE}(X_j \cup \{x\}) < \mathrm{SE}(X_i) + \mathrm{SE}(X_j),$$

czyli na podstawie powyższego gdy:

$$-\frac{|X_i|}{|X_i|-1} \|x - m_i\|^2 + \frac{|X_j|}{|X_j|+1} \|x - m_j\|^2 < 0,$$

czyli gdy

$$\frac{|X_j|}{|X_j|+1} \|x - m_j\|^2 < \frac{|X_i|}{|X_i|-1} \|x - m_i\|^2. \tag{3.1}$$

Proszę zobaczyć, że to jest trochę zbliżone do diagramu Voronoi, ale mamy nieliniową barierę decyzyjną.

Podejście Hartigana dla k-means.
Na wejściu:

- dane $X = (x_i)_{i=1..n} \subset \mathbb{R}^N$
- początkowa przynależność do klastra $\sigma : X \to \{1, \ldots, k\}$ (wybierana losowo).

Rozważamy klastry

$$X_l = \{x \in X : j(x) = l\}.$$

Dla każdego z tych klastrów wyliczamy jego średnią m_i i ilość elementów N_i [2]

Następnie chodzimy kolejno po punktach zbioru, zmieniając przynależność danego punktu o ile zachodzi warunek (3.1) – wtedy też przeliczamy średnie zbiorów i liczności.

Kończymy procedurę gdy przy pełnym przejściu po zbiorze nie zmieniliśmy przynależności żadnego punktu.

Zadanie 3.2 (Python)**.** *Napisać k-means korzystający z metody Hartigana.*

Zadanie 3.3. *Mamy dane dwa klastry X_1, X_2 na płaszczyźnie, o średnich $(-1, 0)$ i $(3, 2)$ i licznościach 10, 20, odpowiednio. Punkt $x = (1, 2)$ należy do klastra X_1. Czy opłacałoby się w podejściu Hartigana z punktu widzenia minimalizacji funkcji kosztu przeniesienie go do X_2?*

Zadanie 3.4. *Podaj przykład zbioru z podziałem na dwa klastry, tak, że metoda Lloyda ich nie zmodyfikuje, a Hartigana tak.*

[2] Uwaga: wystarczy w tym celu przejść raz zbiór: Dla $i = 1..k$ kładziemy $s_i = 0$, $N_i = 0$. For $l = 1..n$ do $s_{j(l)} += x_l$, $N_{j(l)} ++$. Po przejściu kładziemy $m_i = s_i/N_i$ dla $i = 1..l$.

Rozdział 4

PCA

W podejściu rozważanym w k-means celem było zastąpienie zbioru danych za pomocą k punktów $\{v_1, \ldots, v_k\}$ tak by błąd popełniony przy zastąpieniu punktem najbliższym był najmniejszy. Czyli celem była minimalizacja

$$\mathrm{SE}(X, V) = \sum_i d^2(x_i, V) \text{ dla } V \subset \mathbb{R}^D, \mathrm{card} V = k.$$

Oznacza to, że zastępujemy zbiór X dyskretnym zbiorem składającym się z k-punktów.

Okazuje się, że równie ważne, jak nie ważniejsze jest podejście polegające na zastąpieniu X przez podprzestrzeń k-wymiarową. Czyli wtedy nas cel to

$$\mathrm{SE}(X, V) = \sum_i d^2(x_i, V) \text{ dla } V \text{ podprzestrzeń liniowa, } \dim V = k.$$

Ponieważ chcemy mieć możliwość dokonywania translacji, szukamy w klasie przestrzeni afinicznych:

$$\mathrm{SE}(X, W) = \sum_i d^2(x_i, W) \text{ dla } W \text{ podprzestrzeń afiniczna, } \dim W = k.$$

Podrzestrzeń afiniczna to liniowa przesunięta, czyli

$$W = v + V.$$

4.1 Rzutowania ortogonalne

Do dalszych rozważań, przypomnę podstawowe informacje o rzutowaniach ortogonalnych. Jeżeli mamy podprzestrzeń $V \subset \mathbb{R}^D$, to dla każdego $x \in \mathbb{R}^D$ istnieje dokładnie jeden najbliższy $x_V \in V$ do x:

$$x_V = \operatorname*{argmin}_{v \in V} \|x - v\|.$$

Odwzorowanie $x \to x_V$ oznaczam przez p_V. Okazuje się, że p_V jest odwzorowaniem liniowym. Jeżeli $v_1, \ldots, v_k$ jest bazą ortonormalną V, to rzutowanie jest dane wzorem

$$p_V(x) = \langle x, v_1 \rangle v_1 + \ldots + \langle x, v_k \rangle v_k. \tag{4.1}$$

Łatwo zauważyć (ćw), że wzór na projekcję macierzowo dany jest wzorem

$$p_V = V V^T.$$

Jeżeli do kompresji używamy podprzestrzeni afinicznej $W = v + V$, to wtedy oczywiście rzut jest dany wzorem

$$p_W(x) = v + \langle (x - v), v_1 \rangle v_1 + \ldots + \langle (x - v), v_k \rangle v_k,$$

co oznacza, że jeżeli chcemy to przedstawić w układzie współrzędnych o środku w v i bazie v_i to dostajemy współrzędne

$$x \to (\langle (x - v), v_1 \rangle, \ldots, \langle (x - v), v_k \rangle) \in \mathbb{R}^k.$$

Przypominam, że $x \perp y$ o ile $\langle x, y \rangle = 0$. Mówimy, że x jest prostopadłe do V, co zapisujemy $x \perp V$, o ile

$$x \perp v \text{ dla każdego } v \in V.$$

Wtedy $p_V(x)$ to jedyny punkt taki, że $x - p_V(x) \perp V$.

Dla każdej przestrzeni $V \subset \mathbb{R}^D$ możemy rozważać przestrzeń ortogonalną

$$V^\perp = \{x \in X : x \perp V\}.$$

Mamy

$$x = p_V(x) + p_{V^\perp}(x) \text{ dla } x \in \mathbb{R}^D. \tag{4.2}$$

W konsekwencji

$$d(x, V) = \|x - p_V(x)\| = \|p_{v^\perp}(x)\|.$$

Zadanie 4.1. *Korzystając z* (4.1) *proszę wyliczyć wzór na rzutowanie ortogonalne na*

$$V = \{(x, y, z) : x + y + z = 0\}.$$

Proszę podać macierz tego rzutowania.

Zadanie 4.2. *Korzystając z* (4.2) *proszę wyliczyć wzór na rzutowanie ortogonalne na*

$$V = \{x = (x_1, \ldots, x_D) \in \mathbb{R}^D : \sum_i x_i = 0\}.$$

Wsk.: proszę zauważyć, że $w = (1, \ldots, 1)$ *jest prostopadły do* V.

4.2 Optymalne położenie środka

Załóżmy, że podprzestrzeń liniową V mamy zafiskowaną, i modyfikujemy tylko przesunięcie v. Zaczniemy od pokazania, że co nie zaskakujące, optymalnie v to środek X:

$$\mathrm{mean}X = \operatorname*{argmin}_v d^2(X, v + V).$$

Obserwacja 4.1. *Niech będzie dana podprzestrzeń wektorowa* V *przestrzeni* $\mathbb{R}^D$. *Wtedy*

$$d^2(X, v + V) = \mathrm{SE}(p_{V^\perp}(X)) + |X| \cdot \|p_{V^\perp}(\mathrm{mean}X - v)\|^2 \tag{4.3}$$

i w konsekwencji wartość ta jest minimalizowana gdy $v = \mathrm{mean}X$ *(środek ciężkości* x*).*

Dowód. Oczywiście

$$d^2(x; v+V) = d^2(x-v; V) = \|p_{V^\perp}(x-v)\|,$$

Czyli na podstawie Obserwacji 2.1

$$d^2(X, v+V) = \sum_i d^2(x_i, v+V) = \sum_i \|p_{V^\perp}(x_i) - p_{V^\perp}(v)\|^2$$

$$= \sum_i d^2(p_{V^\perp}(x_i), p_{V^\perp}(v)) = \mathrm{SE}(p_{V^\perp}(X), p_{V^\perp}(v))$$

$$= \mathrm{SE}(p_{V^\perp}(X)) + |X| \cdot \|\mathrm{mean}(p_{V^\perp}(X)) - p_{V^\perp}(v)\|^2.$$

Ponieważ z liniowości rzutowań, mamy $\mathrm{mean}(p_{V^\perp}(X)) = p_{V^\perp}(\mathrm{mean}X)$, dostajemy tezę. □

W konsekwencji oznacza to, że jeżeli mamy możliwość wyboru translacji przestrzeni, zawsze wybieramy środek.

4.3 Sytuacja jednowymiarowa

Teraz zajmiemy się przypadkiem jednowymiarowym. Rozpatrzmy X, i zajmiemy się szukaniem takiego v, że X jest optymalnie kompresowany przez przestrzeń afiniczną zaczepioną w średniej X i rozpiętej na v, czyli $\mathrm{mean}X + \mathbb{R}v$. Czyli szukamy minimalizacji

$$\operatorname*{argmin}_v \mathrm{SE}(X - \mathrm{mean}X, \mathbb{R}v).$$

Możemy oczywiście założyć, że v ma normę jeden, wtedy

$$\mathrm{SE}(X - \mathrm{mean}X, \mathbb{R}v) = \sum_i d^2(x_i - \mathrm{mean}X, \mathbb{R}v) = \sum_i \|(x_i - \mathrm{mean}X) - p_{\mathbb{R}v}(x_i - \mathrm{mean}X)\|^2.$$

Niech $z_i = x_i - \mathrm{mean}X$. Mamy oczywiście $p_{\mathbb{R}v} z_i = \langle z_i, v\rangle v$ oraz

$$\|z_i - p_{\mathbb{R}v} z_i\|^2 = \|z_i - \langle z_i, v\rangle v\|^2 = \|z_i\|^2 - 2\langle z_i, v\rangle\langle z_i, v\rangle + \|\langle z_i, v\rangle v\|^2$$

$$= \|z_i\|^2 - 2\langle z_i, v\rangle^2 + \langle z_i, v\rangle^2 = \|z_i\|^2 - \langle z_i, v\rangle^2.$$

Teraz minimalizacja

$$\sum_i (\|z_i\|^2 - \langle z_i, v\rangle^2) = \sum_i \|z_i\|^2 - \sum_i \langle z_i, v\rangle^2$$

jest oczywiście równoważna maksymalizacji

$$\sum_i \langle z_i, v\rangle^2.$$

Konkludując mamy problem znalezienia

$$\mathrm{argmin}\{\sum_i \langle z_i, v\rangle^2 \mid v : \|v\| = 1\}.$$

Ponieważ

$$\langle v, w\rangle = v^T w = w^T v,$$

to

$$\sum_i \langle z_i, v\rangle^2 = \sum_i v^T z_i z_i^T v = v^T (\sum_i z_i z_i^T) v$$

$$= v^T (\sum_i (x_i - \mathrm{mean}X)(x_i - \mathrm{mean}X)^T) v = v^T |X| \mathrm{cov}X v.$$

W konsekwencji sprowadziliśmy do problemu

$$\mathrm{argmax}\{v^T \mathrm{cov}X v \,|\, v : \|v\| = 1\}.$$

Problem 4.1. *Mamy daną macierz symetryczną nieujemnie określoną Σ. Należy znaleźć*

$$\mathrm{argmax}\{v^T \Sigma v \,|\, v : \|v\| = 1\}.$$

Stwierdzenie 4.1. *Maksimum jest realizowane przez wektor własny odpowiadający największej wartości własnej Σ.*

Dowód. Zmieniamy bazę na taką ortonormalną która diagonalizuje Σ, czyli mamy taką bazę ortonormalnę $F = [f_1, \ldots, f_D)]$ (patrz Twierdzenie 2.1), że Σ się diagonalizuje, czyli

$$\Sigma = F \Lambda F^{-1} = F \Lambda F^T,$$

gdzie Λ to macierz diagonalna mająca na diagonali uporządkowane malejąco wartości własne.

Wtedy dla

$$v = \alpha_1 f_1 + \ldots + \alpha_D f_D = F\alpha \text{ dla } \begin{bmatrix} \alpha_1 & \vdots & \alpha_D \end{bmatrix}$$

mamy

$$v^T \Sigma v = \alpha^T \Lambda \alpha = \lambda_1 \alpha_1^2 + \ldots + \lambda_D \alpha_D^2.$$

Interesuje nas w takim razie szukanie maksimum

$$\lambda_1 \alpha_1^2 + \ldots + \lambda_D \alpha_D^2$$

przy warunku $\|v\|^2 = \alpha_1^2 + \ldots + \alpha_D^2 = 1$. Ponieważ λ_1 jest największe, mamy

$$\lambda_1 \alpha_1^2 + \ldots + \lambda_D \alpha_D^2 \leq \lambda_1 (\alpha_1^2 + \ldots + \alpha_D^2) = \lambda_1.$$

Czyli w konsekwencji maksimum jest osiągane dla

$$\lambda_1 = 1, \lambda_2 = \ldots = \lambda_D = 0,$$

co oznacza, że jako v bierzemy wektor własny odpowiadający największej wartości własnej Σ. □

Uwaga 4.1 (interpretacja geometryczna). Załóżmy, że chcemy wyznaczyć kierunek najbardziej reprezentatywny dla zestawu danych X (zakładamy, że średnia jest zero).

Weźmy jeden punkt $x \in \mathbb{R}^N$ i rozpatrzmy $\langle y, x\rangle$ (proszę narysować poziomice). Oczywiście, największe (przy danej normie) jest w x, ale najmniejsze w $-x$. Ponieważ interesuje nas prosta przechodząca przez zarówno x jak i $-x$, jeżeli weźmiemy $\langle y, x\rangle^2$ dostaniemy formę kwadratową, dla której kierunek największego wzrostu będzie dokładnie wyznaczał zarówno x jak i $-x$.

Dla danych x po prostu sumujemy te funkcje kwadratowe, dostając:

$$\mathbb{R}^N \ni y \to \sum_i \langle y, \mathrm{x}^i\rangle^2,$$

i po przeliczeniu dostajemy, że powyższe odwzorowanie dane jest wzorem

$$y \to y' \Sigma_{\mathrm{x}} y.$$

W konsekwencji nasza intuicja jest taka, by wybrać w formie kwadratowej zdefiniowanej przez Σ_{x} kierunek największego wzrostu, i to będzie najlepsze przybliżenie. Pokażemy, że tak jest.

4.4 Sytuacja wyżej-wymiarowa

Zaczniemy od pokazania wzoru który pozwala wyliczyć sumę kwadratów tylko przy pomocy kowariancji.

Stwierdzenie 4.2. *Niech $X = (x_i)_{i=1..n}$ zbiór danych, $v = [v_1, \ldots, v_k]$ baza ortonormalna pewnej podprzestrzeni $V \subset \mathbb{R}^D$. Wtedy*

$$\mathrm{SE}(X - \mathrm{mean}X, V) = \mathrm{SE}(X) - |X|\mathrm{tr}(v^T \mathrm{cov}X v).$$

Dowód. Niech $z_i = x_i - \mathrm{mean}X$, $Z = (z_i)$.

Interesuje nas wartość

$$\mathrm{SE}(Z, V) := \sum_{i=1}^{n} \|z_i - p_V z_i\|^2 = \sum_{i=1}^{n} (\|z_i\|^2 - \|p_V z_i\|^2)$$

$$= \sum_{i=1}^{n} \|z_i\|^2 - \sum_{i=1}^{n} \|p_V z_i\|^2.$$

Oczywiście

$$\sum_{i=1}^{n} \|z_i\|^2 = \mathrm{SE}(X).$$

Z drugiej strony, mamy

$$\|p_V z\|^2 = \sum_{j=1}^{k} \langle z, v_j \rangle^2 = \sum_{j=1}^{k} (v_j^T z) \cdot (z^T v_j) = \sum_{j=1}^{k} v_j^T (z z^T) v_j = \mathrm{tr}(V^T z z^T V),$$

czyli

$$\sum_{i=1}^{n} \|p_V z_i\|^2 = \sum_i \mathrm{tr}(v^T z_i z_i^T v) = \mathrm{tr}\big(v^T (\sum_i z_i z_i^T) v\big).$$

Ponieważ $|X|\mathrm{cov}X = \sum_i z_i z_i^T$ dostajemy tezę. □

Proszę zauważyć, że

$$\mathrm{SE}(X - \mathrm{mean}X, V) = \mathrm{SE}(X) - |X|\mathrm{tr}(v^T \mathrm{cov}X v) = |X| \cdot (\mathrm{tr}(\mathrm{cov}X) - \mathrm{tr}(v v^T \mathrm{cov}X))$$

$$= |X| \cdot \mathrm{tr}((I - v v^T)\mathrm{cov}X)) = |X| \cdot \mathrm{tr}(p_{V^\perp} \mathrm{cov}X),$$

gdzie jak przypominam $p_{V^\perp}$ to rzutowanie ortogonalne na przestrzeń prostopadłą do V.

Teraz jesteśmy już w stanie sformułować główne twierdzenie obecnej sekcji, które pozwala nam zminimalizować błąd. Dowód jest podobny do przypadku jednowymiarowego.

Twierdzenie 4.1. *Rozpatrzmy wszystkie q-wymiarowe podprzestrzenie V o bazie ortonormalnej v w przestrzeni p-wymiarowej. Wtedy wartość*

$$\mathrm{tr}(v^T \Sigma v)$$

jest maksymalna, gdy v to pierwsze q-elementów bazy ortonormalnej składającej się z wektorów własnych macierzy Σ ustawionych malejąco po wartościach własnych.

Dowód. Bierzemy bazę $F = [f_1, \ldots, f_p]$ zbudowaną z ortonormalnych wektorów własnych Σ która diagonalizuje Σ (Λ po diagonalizacji, zakładamy jak zwykle, że wartości własne w Λ są ustawione malejąco), tzn.:

$$F\Lambda F^T = \Sigma \text{ lub równoważnie } \Lambda = F^T \Sigma F.$$

Niech $c = [c_1, \ldots, c_q]$ oznaczają współrzędne $v = [v_1, \ldots, v_q]$ w tej nowej bazie F, to jest $c_i = F^{-1}v_i$, czyli $c = F^{-1}v$. Można łatwo sprawdzić, że c_i też jest ortonormalny, co więcej maksymalizacja $v \to \mathrm{tr}(v^T \Sigma v)$ sprowadza się do maksymalizacji

$$c \to \mathrm{tr}((Fc)^T \Sigma (Fc)) = \mathrm{tr}(c^T \Lambda c).$$

Łatwo można sprawdzić, że

$$\mathrm{tr}(c^T \Lambda c) = \sum_{j,k} \lambda_j c_{jk}^2 = \sum_{j=1}^{p} (\sum_{k=1}^{q} c_{jk}^2)\lambda_j = \sum_{j=1}^{p} \lambda_j a_j,$$

gdzie

$$a_j = \sum_{k=1}^{q} c_{jk}^2 (\text{kwadrat normy } j\text{-tego wiersza } c).$$

Ponieważ c to baza ortonormalna, $c^T c = I$ czyli

$$\sum_{jk} c_{jk}^2 = \sum_{j=1}^{p} \left(\sum_{k=1}^{q} c_{jk}^2 \right) = q,$$

czyli

$$\sum_{j=1}^{p} a_j = q.$$

Teraz możemy rozszerzyć c do macierzy c ortogonalnej o wymiarach $p \times p$. Ale teraz wiersze D też są ortogonalne, wiersze mają normę jeden, czyli wiersze c są ograniczone od góry przez jeden, czyli

$$a_j = \sum_{k=1}^{q} c_{jk}^2 \leq 1.$$

W konsekwencji wylądowaliśmy na problemie maksymalizacji

$$\sum_{j=1}^{p} \lambda_j a_j \text{ przy warunkach } a_j \in [0,1], \sum_{j=1}^{p} a_j = q.$$

Widać, że rozwiązanie jest maksymalne gdy

$$a_1 = \ldots = a_q = 1, \text{ oraz } a_{q+1} = \ldots = a_p = 0.$$

Ale to jest realizowane dla c_i będacych kolejnymi elementami bazy kanonicznej, czyli wtedy oczywiście w konsekwencji $v_i = f_i$. □

Ile wymiarów wybrać?
Załóżmy, że już znalezliśmy optymalną bazę. Wtedy mamy

Stwierdzenie 4.3. *Mamy*

$$\sum_i d^2(x_i, V_k) = \mathrm{SE}(X) - |X| \sum_{i=1}^{k} \lambda_i.$$

I wtedy ustalamy jaki procent wariancji chcemy mieć wyjaśniony.

Zadanie 4.3. *Mamy przestrzeń w* $\mathbb{R}^4$ *zadaną przez* $(1,1,0,0), (0,0,1,1)$ *i zbiór o kowariancji* I *i liczności* 100*. Proszę policzyć błąd popełniony przy rzutowaniu.*

Rozdział 5

Zmienne i wektory losowe: DODATEK

Poniższy rozdział powinien być w zasadzie dobrze znany.

Zmienną losową jest na przykład funkcja opisująca wagę lub wzrost ciała wylosowanego z pewnej populacji osobnika. Rozkad zmiennej to miara na zbiorze wartości tej zmiennej. Rozdziela się na zmienne losowe ciągłe i dyskretne, w związku z czym są rozkłady ciągłe i dyskretne.

Wektor losowy odpowiada sytuacji, gdy mierzymy dla wylosowanego obiektu więcej cech rzeczywistych, czyli element z $\mathbb{R}^d$.

5.1 Rozkłady dyskretne

Rozkład jest rozkładem dyskretnym, gdy możemy wylosować tylko skończoną (bądź co najwyżej przeliczalną) liczbę punktów $x_1, \ldots, x_N$. Taki rozkład określa się przez podanie prawdopodobieństw wylosowania każdej z tych liczb, czyli $p_1, \ldots, p_k$.

W praktyce, jeżeli mamy dane pochodzące z tego rozkładu, możemy *wyestymować* te wartości prawdopodobieństw, biorąc coraz większą próbkę z rozkłady, możemy zliczyć $\hat{p}_i$ jak często wylosujemy każdy z elementów x_i. Asymptotycznie przy wielkości próbki zmierzającej do nieskończoności, te częstości zmierzają do prawdopodobieństw p_i które mówią, że nasza zmienna wylosuje x_i. Konkludując, w naszym przypadku dostajemy rozkład prawdopodobieństwa (x_i, p_i).

Zacznijmy od rozkładu dyskretnego który jak pokażemy pozwala generować wszyskie inne. Otóż chodzi nam o skonstruowanie zmiennej losowej $\mathbb{X}$ która przyjmuje z jednolitym prawdopodobieństwem przyjmuje wartości w zbiorze $\mathbb{Z}_M = \{0, \ldots, M-1\}$, gdzie$M$ jest bardzo dużą liczbą (z powodów numeryczych M jest często potęgą dwójki). Czyli chcemy by każde $i \in \mathbb{Z}_M$ było losowane z jednakowym prawdopodobieństwem $1/M$.

Uwaga 5.1. Ponieważ programowo nie da się na komputerze zaimplementować generatora prawdziwych liczb losowych, w związku z czym istnieją sprzętowe generatory liczby losowych które używają fizycznych zjawisk które posiadają własności losowe typu zjawiska kwantowe (korzystają z nich głównie banki dla bezpieczeństwa). To co się robi najczęściej w praktyce z powodu szybkości i potencjalnej powtarzalności doświadczeń to generowanie liczb pseudolosowych (pseudo-random number generator).

Często stosowane generatory liczb losowych zazwyczają polegają na określeniu wartości startowej x (SEED) i funkcji $f : \mathbb{Z}_M \to \mathbb{Z}_M$ tak, że nas ciąg pseudolosowy jest dany przez

$$x_{n+1} = f(x_n).$$

Funkcja f musi być tak dobrana aby nie było oczywistego związku pomiędzy poprzednimi wartościami a następnymi oraz by była szybka w obliczaniu. Najprostsze generatory powyższego typu to LCG (linear congruential generator):

$$x_{n+1} = (ax_n + c) \bmod m.$$

Pomimo tego, że szybkie, nie powinny być używane do zadań wymagających prawdziwej losowości, gdyż nie spełniają wszystkich testów statystycznych sprawdzających losowość.

Uwaga 5.2. Zły dobór parametrów może mieć tragiczne skutki – niesławny tu jest RANDU zaprojektowany w latach 60tych przez IBM-a:

$$x_{j+1} = 65539x_j \bmod 2^{31}.$$

Otóż $x_{k+2} = (2^{16} + 3)x_{k+1} = (2^{16} + 3)^2 x_k$, co oznacza, że

$$x_{k+2} = (2^{32} + 6 \cdot 2^{16} + 9)x_k = [6 \cdot (2^{16} + 3) - 9]x_k = 6x_{k+1} - 9x_k \bmod 2^{31}.$$

W konsekwencji punkty (x_k, x_{k+1}, x_{k+2}) leżą w przestrzeni $\mathbb{R}^3$ na niewielkiej liczbie płaszyzn (jest silna korelacja, nie ma niezależności). W konsekwencji wiele prac fizycznych bazujących na symulacjach losowych używających tego generatora okazało się być nieprawdziwych.

Generatory zbliżone w sensie idei do LCG, a uznawane za dobre, to pracujące na rozwinięciu bitowym danych, dla przykładu można polecić różne udoskonalenia `xorshift`.

Z rozkładów dyskretnych które należy znać, to rozkłady dwumianowy (Bernoulliego) odpowiadający ilości sukcesów w n rzutach monetą (p odpowiada za sukces w pojedynczym doświadczeniu):

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

i rozkład Poissona (opisuje dobrze na przykład liczbę osób w kolejce do kasy):

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda}$$

Zadanie 5.1. *Jak ktoś z Państwa nie zna, proszę doczytać we własnym zakresie.*

5.2 Gęstość

Teraz zajmiemy się przypadkiem ciągłym, gdzie możemy potencjalnie wylosować dowolne liczby rzeczywiste. Przypomnijmy pojęcia histogramu, który jest bardzo zbliżony do koncepcji dyskretyzacji. Otóż ponieważ trudno jest opisać wszystkie możliwe wyniki, dla prostoty i zmniejszenia ilości pamięci możemy podzielić R rozłączne „pudełka" równej długości:

$$P_l = [a + \tfrac{l}{K}(b-a), a + \tfrac{l+1}{K}(b-a)) \text{ dla } l \in \mathbb{Z}.$$

Mając zestaw danych $X = (x_i)$ aby stworzyć histogram zliczamy ile (albo ile procent) danych wpadło danych do każdego pudełka.

Wyobraźmy sobie teraz, że chcemy zrobić dwukrotnie drobniejszy podział, bo na przykład pojawiło się więcej danych. Wtedy każde pudełko rozbije się nam na dwa, i w konsekwencji w nowych pudełkach będzie dwukrotnie mniej danych. Czyli jak będziemy robić coraz drobniejsze podziały, nasz histogram będzie miał coraz mniejszą wysokość.

Aby temu przeciwdziałać, rozważa się histogramy bazujące na *gęstości prawdopodobieństwa* (je także używa się wtedy gdy chcemy używać histogramu o różnej szerokości przedziałów). A mianowicie wysokość h_i danego prostokąta (odpowiadającemu graficznie naszemu pudełku) ustala się tak, by jego pole było równe prawdopodobieństwu wpadania danych do pudełka:

$$h_i \cdot |P_i| = P(\mathbb{X} \in P_i) \approx \tfrac{1}{|X|}\mathrm{card}\{i : x_i \in P_i\}.$$

Biorąc coraz węższe szerokości, otrzymujemy[1] graniczny histogram który nazywamy gęstością $f_{\mathbb{X}}$ zmiennej losowej $\mathbb{X}$. Mianowicie, dla punktu $x \in \mathbb{R}$, rozpatrujemy coraz mniejsze pudełka P_x wokół x, i rozpatrujemy iloraz zawartości prawdopodobieństwa przez szerokość pudełka $|K_x|$:

$$\frac{P(\mathbb{X} \in P_x)}{|P_x|} \to f_{\mathbb{X}}(x) \text{ przy } P_x \to x.$$

Przy założeniu, że $f_{\mathbb{X}}$ jest dobrze zdefiniowaną *gęstością, czyli nieujemną funkcją która całkuje się do jedynki*, możemy odtworzyć prawdopodobieństwo wylosowania punktu ze zbioru A całkując:

$$P(\mathbb{X} \in A) = \int_A f_{\mathbb{X}}(x)dx. \tag{5.1}$$

Fakt, że $\mathbb{X}$ ma gęstość $f_{\mathbb{X}}$ oznaczamy $\mathbb{X} \sim f_{\mathbb{X}}$.

W przypadku wektorów losowych (czyli gdy losujemy dane z $\mathbb{R}^D$), powyższe wzór się nie zmienia, tylko zamiast przedziałów bierzemy jakiekolwiek D-wymiarowe kostki (czy równoległościany) wokół x, i wtedy $|P_x|$ oznacza ich D-wymiarową objętość.

I teraz widzimy na podstawie (5.1), że dwa wektory losowe są utożsamiane jeżeli mają równą gęstość.

5.3 Rozkłady ciągłe

Zajmiemy się rozkładami ciągłymi (posiadającymi gęstość). Najbardziej istotny poza rozkładem normalnym który omówię później jest rozkład równomierny na odcinku $[0, 1]$, oznaczam go przez $\mathrm{unif}_{[0,1]}$. Gęstość takiego rozkładu to funkcja stale równa 1 na odcinku $[0, 1]$, a 0 poza nim. Możemy losować (oczywiście w przybliżeniu) z tego rozkładu biorąc wynik generatora dyskretnego na zakresie $\{0, \ldots, M-1\} \subset \mathbb{N}$ i dzieląc go przez M. Jeżeli M jest duże, a często M jest rzędu 2^{32} czy 2^{64} powyższy wynik jest numerycznie nieodróżnialny od prawdziwego rozkładu równomienie rozłożonego na odcinku $[0, 1]$.

Z rozkładów które koniecznie trzeba znać, to rozkład wykładniczy:

$$\frac{1}{\lambda}\exp(-x/\lambda) \text{ dla } x > 0,\ 0 \text{ w przeciwnym razie,}$$

oraz rozkład normalny $N(m, \sigma^2)$ o średniej m i wariancji σ^2:

$$N(m, \sigma^2)(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\tfrac{1}{2}((x-m)/\sigma)^2).$$

[1] UWAGA: nie zawsze ta granica istnieje!

5.4 Generowanie rozkładów

Pokażemy, że umiejętność losowania danych z $\mathrm{unif}_{[0,1]}$ pozwala na generowanie liczb z dowolnych rozkładów. Najprościej oczywiście wylosować rozkład równomierny na odcinku $[a, b]$ – a mianowicie, jeżeli $\mathbb{X}$ ma rozkład równomierny na odcinku $[0, 1]$ to jak zaraz zobaczymy $a + (b - a)\mathbb{X}$ ma rozkład równomierny na odcinku $[a, b]$. Często się właśnie generuje rozkłady biorąc funkcję ϕ na znanym rozkładzie $\mathbb{X}$.

Dosyć prosto także generować rozkład dyskretny o prawdopodobieństwach p_i: wystarczy losować x_i jeżeli z rozkładu jednostajnego na $[0, 1]$ trafiliśmy do przedziału $[\sum_{j<i} p_j, \sum_{j\leq i} p_i]$. Uogólnienie na rozkłady ciągłe to metoda odwracania dystrybuanty: jeżeli dystrybuanta rozkładu P to ϕ (i ϕ jest odwracalna), to $\phi^{-1}(\mathbb{U})$ ma zadany rozkład P. Nie zawsze da się stosować, bo musi być wzór na odwrotną do dystrybuanty.

W ten sposób można dla przykładu generować rozkład wykładniczy.

INNE METODY: TODO

5.4.1 Gęstość rozkładu zmiennej $\phi(\mathbb{X})$

W poniższych rozważaniach zakładamy zawsze, że ϕ jest funkcją kawałkami różniczkowalną której pochodna jest odwracalna. Zajmiemy się najpierw przypadkiem najprostszym kiedy zmienna losowa $\mathbb{X}$ przyjmuje wartości w przedziale $[a, b]$ a rozpatrywana funkcja ϕ jest różnowartościowa i przekształca przedział $[a, b]$ w sposób jednoznaczny na przedział $[c, d]$. Gęstość $\mathbb{X}$ oznaczamy przez $f_{\mathbb{X}}$.

Spróbujmy więc wyliczyć gęstość zmiennej losowej $\mathbb{Y} = \phi(\mathbb{X})$. Weźmy w tym celu punkt $y \in [c, d]$ i jedyny $x \in [a, b]$ taki, że $y = \phi(x)$. Spróbujemy obliczyć gęstość $\mathbb{Y}$ w punkcie y. Weźmy małe $\delta > 0$ i rozpatrzmy pudełko $K_y^\delta = y + [-\delta, \delta]$ wokół y. Korzystając z różniczkowalności ϕ mamy

$$\phi(x + h) \approx \phi(x) + \phi'(x)h \text{ dla małych } h.$$

Oznacza to, że dla małych δ, kładąc

$$K_x^\delta = x + \frac{1}{\phi'(x)}[-\delta, \delta]$$

dostajemy

$$\phi(K_x^\delta) \approx y + [-\delta, \delta] = K_y^\delta \text{ oraz } |K_y^\delta| = |\phi'(x)| \cdot |K_x^\delta|.$$

I teraz mamy:

$$\frac{P(\mathbb{Y} \in K_y^\delta)}{|K_y^\delta|} \approx \frac{P(\phi(\mathbb{X}) \in \phi(K_x^\delta))}{|\phi'(x)| \cdot |K_x^\delta|} \approx \frac{P(\mathbb{X} \in K_x^\delta)}{|\phi'(x)| \cdot |K_x^\delta|} \to \frac{1}{|\phi'(x)|} f_{\mathbb{X}}(x) \text{ przy } \delta \to 0.$$

Konkludując dostajemy

$$f_{\mathbb{Y}}(y) = \frac{1}{|\phi'(x)|} f_{\mathbb{X}}(x) \text{ gdzie } x = \phi^{-1}(y).$$

Stosując powyższy wzór dla $\mathbb{X} \sim \mathrm{unif}_{[0,1]}$ i $\phi(r) = a + (b - a)r$ dostajemy sposób na generowanie rozkładu $\mathrm{unif}_{[a,b]}$ z rozkładu $\mathrm{unif}_{[0,1]}$.

Zadanie 5.2. *Rozkład wykładniczy ma gęstość daną wzorem:*

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} \text{ dla } x \geq 0, \\ 0 \text{ w przeciwnym przypadku.} \end{cases}$$

Łatwo można pokazaż, że λ to średnia i λ^2 to wariancja (pokazać).

Rozkład wykładniczy możemy wygenerować z jednostajnego $\mathbb{X} \sim \mathrm{unif}_{[0,1]}$ obkładając przez funkcję logarytm, to znaczy

$$\mathbb{Z} = \lambda \ln(\mathbb{X})$$

ma rozkład wykładniczy (pokazać).

Rozpatrzmy teraz sytuację jednowmiarową, gdy nie mamy odwaracalności ϕ. Wtedy y ma potencjalnie wiele przeciwobrazów, to znaczy zbiór

$$\phi^{-1}(y) = \{x : \phi(x) = y\}$$

może mieć więcej niż jeden element. W konsekwencji powtarzając poprzednie rozumowanie, otrzymujemy jedynie, że y może być uzyskany przez więcej x, tak więc i jego prawdopodobieństwo powstaje jako suma po przeciwobrazach:

$$f_{\mathbb{Y}}(y) = \sum_{x:\phi(x)=y} \frac{1}{|\phi'(x)|} f_{\mathbb{X}}(x).$$

Teraz zajmiemy się przypadkiem wielowymiarowym, zakładamy więc, że $\mathbb{X}$ jest wektorem losowym w $\mathbb{R}^D$ o gęstości $f_{\mathbb{X}}$, i rozpatrujemy funkcję różnowartościową $f : \mathbb{R}^D \to \mathbb{R}^D$. Aby powtórzyć rozumowanie jak w przypadku jednowymiarowym, przyda nam się następujące przypomnienie z algebry liniowej.

Dygresja 5.1. *Niech $K \subset \mathbb{R}^D$ będzie kostką bądź równoległościanem i niech $A : \mathbb{R}^D \to \mathbb{R}^D$ będzie odwracalnym odwzorowaniem liniowym. Wtedy*

$$|AK| = |\det A| \cdot |K| \text{ oraz } |K| = |\det A^{-1}||AK|,$$

gdzie $|L|$ oznacza objętość zbioru L a $\det A$ oznacza wyznacznik macierzy A. Oczywiście także $\det A^{-1} = 1/\det A$.

Spróbujemy obliczyć gęstość $\mathbb{Y}$ w punkcie y. Weźmy małe $\delta > 0$ i rozpatrzmy pudełko w kształcie hiper-kostki (sześcianu) $K_y^\delta = y + [-\delta, \delta]^D$ wokół y. Korzystając z różniczkowalności ϕ mamy

$$\phi(x + h) \approx \phi(x) + d_x\phi \cdot h \text{ dla małych } h.$$

gdzie $d_x\phi$ oznacza pochodną $\phi = (\phi_1, \ldots, \phi_D)$ w punkcie x, czyli macierz daną przez pochodne cząstkowe

$$d_x\phi = \left[\frac{\partial \phi_i}{\partial x_j}\right]_{ij}.$$

Oznacza to, że dla małych δ, definiując D-wymiarowy równoległościan

$$K_x^\delta = x + (d_x\phi)^{-1}[-\delta, \delta]^D$$

dostajemy

$$\phi(K_x^\delta) \approx y + [-\delta, \delta]^D = K_y^\delta,$$
$$|K_y^\delta| = |y + d_x\phi(K_x^\delta)| = |\det d_x\phi| \cdot |K_x^\delta|.$$

I teraz mamy:

$$\frac{P(\mathbb{Y} \in K_y^\delta)}{|K_y^\delta|} \approx \frac{P(\phi(\mathbb{X}) \in \phi(K_x^\delta))}{|d_x\phi(K_x^\delta)|} = \frac{P(\mathbb{X} \in K_x^\delta)}{|\det d_x\phi| \cdot |K_x^\delta|} \to \frac{1}{|\det d_x\phi|} f_{\mathbb{X}}(x) \text{ przy } \delta \to 0.$$

Konkludując dostajemy

$$f_{\mathbb{Y}}(y) = \frac{1}{|\det d_x\phi|} f_{\mathbb{X}}(x) \text{ gdzie } x = \phi^{-1}(y).$$

W przypadku gdy ϕ nie jest różnowartościowa, rozumując analogicznie jak w przypadku jednowymiarowym dostajemy wzór

$$f_{\mathbb{Y}}(y) = \sum_{x:\phi(x)=y} \frac{1}{|\det d_x\phi|} f_{\mathbb{X}}(x) \text{ gdzie } x = \phi^{-1}(y).$$

Zadanie 5.3. *$\mathbb{X}$ ma rozkład jednostajny na odcinku $[-1, 1]$. Jaki rozkład ma $\mathbb{X}^2$?*

5.4.2 Rozkład warunkowy

Załóżmy, że mamy wektor losowy $\mathbb{X}$ i interesuje nas jedynie sytuacja gdy wylosowaliśmy $\mathbb{X}$ w danym zbiorze A. Wtedy gęstość się normalizuje do A.

Sposób losowania jest bardzo prosty - losujemy, a jeżeli nie wypadło w A, to odrzucamy.

Przykład 5.1. Jeżeli potrafimy losować z rozkładu jednostajnego na $[0, 1]$, to oczywiście też potrafimy losować z rozkładu jedostajnego na $[0, 1]^D$. W konsekwencji, jeżeli A jest ograniczony, to potrafimy losować z rozkładu jednostajnego na A za pomocą odrzucania.

Typowy przykład to obliczanie π (pola koła - za pomocą rzucania rzutkami).

Obliczanie objętości kuli D-wymiarowej:

$$V_D(R) = \frac{\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}+1\right)} R^D \approx \frac{1}{\sqrt{D\pi}} \left(\frac{2\pi e}{D}\right)^{\frac{D}{2}} R^D.$$

Rzucanie rzutkami do kuli wysoko wymiarowej jest trudne.

Sytuacja ciekawsza - gdy A ma miarę zero (zazwyczaj prosta, czy płaszczyzna). Wtedy zawężamy się do całkowania po tym zbiorze.

5.4.3 Niezależność i rozkłady brzegowe

Ważną własnością jest *niezależność zmiennych losowych* $X, Y : \Omega \to \mathbb{R}$ (nad tą samą przestrzenią probabilistyczną, w sensie interpretacji z czarną skrzynką, to należy na to patrzeć jakby jedno naciśnięcie przyciska zwracało nam zarówno X jak i Y). Otóż zmienne są niezależne gdy

$$P((X, Y) \in I \times J) = P(X \in I) \cdot P(Y \in J) \text{ dla dowolnych przedziałów } I, J \subset \mathbb{R}.$$

Intuicyjnie chodzi o to, że wynik wylosowany na jednej nie wpływa na wynik wylosowany na drugiej. Problem z tym, że nie da się tego łatwo przekształcić na jakiś współczynnik który się da policzyć efektywnie (można kombinować, na przykład dla dywergencji Kullbacka-Leiblera, ale i tak się nie da tego policzyć jak mamy tylko próbkę, trzeba używać estymacji). W związku z tym wprowadza się pojęcie uboższe, ale za to policzalne.

Jeżeli współczynnik korelacji jest różny od zera, to zmienne są zależne, ale nie odwrotnie – mogą być zmienne zależne dla których współczynnik korelacji liniowej jest zero (dla przykładu wylosowanie losowej pary liczb (x, y) na okręgu jednostkowej, oczywiście zmienne są zależne, nawet znamy tą zależność $x^2 + y^2 = 1$, a nie są jak łatwo sprawdzić skorelowane).

5.5 Zasada największej wiarygodności (MLE): TODO

Rozdział 6

Entropia

6.1 Nierówność Krafta

Będziemy korzystać z książki [**?**, Rozdział V] oraz [**?**, Rozdział 4].

Mamy alfabet źródłowy S (o mocy m) i alfabet kodowy $A = \{a_1, \ldots, a_n\}$. W praktyce alfabetem kodowym jest $\{0, 1\}$.

Chcemy przesłać tekst napisany w alfabecie źródłowym, ale nasz kanał informacyjny pozwala na przesyłanie tylko A. Czyli chcemy każdy element z S wyrazić za pomocą słów z A^* (niepuste słowa o skończonej długości).

Definicja 6.1. Przez funkcję kodującą (kodowanie) rozumiem dowolną funkcję $\varphi : S \to A^*$.

Kodowanie nazywamy *nieosobliwym* jeżeli jest iniektywne, to znaczy jeżeli dwa różne elementy kodowane są różnymi kodami (słowamu). Jeżeli mamy wiele, to wtedy oddzielamy znakiem specjalnym (zazwyczaj przecinkiem, spacją bądź średnikiem). Ale to nie jest wygodne, bo musimy używać dodatkowego symbolu, który nie możemy używać do kodowania (w konsekwencji możemy mniej zakodować).

Definicja 6.2. Rozszerzenie kodu to odwzorowanie $\varphi : S^* \to A^*$ dane wzorem

$$\varphi(s_1 s_2 \ldots s_k) := \varphi(s_1)\varphi(s_2)\ldots\varphi(s_k).$$

Kodowanie (kod) jest *jednoznacznie dekodowalne* jeżeli jego rozszerzenie jest nieosobliwe. Innymi słowy, kodowanie jest nieosobliwe, jeżeli mając słowo $w = w_1 w_2 \ldots w_K$ (gdzie w_i to słowa kodowe) możemy jednoznacznie odzyskać jego rozkład na $w_1; w_2; \ldots; w_k$.

Ważnym typem kodów są kody *przedrostkowe - prefiksowe (z ang. prefix)*, to jest takie, że żadne ze słów kodujących nie jest przedrostkiem następnego[1]. Łatwo zauważyć, że kod przedrostkowy jest jednoznacznie dekodowalny.

Pytanie, jeżeli mamy dany alfabet, i chcemy zrealizować kod o zadanej długości - kiedy nam się uda?

Twierdzenie 6.1 (Nierówność Krafta)**.** *Alfabet źródłowy S o m elementach, da się zakodować słowami prefiksowymi z alfabetu kodowego A o d-elementach o długościach $l_1, \ldots, l_m$ wtw. gdy*

$$\sum_{i=1}^{m} d^{-l_i} \leq 1.$$

[1] najwygodniejsze w użyciu, bo nie musimy badać do końca, aby rozpoznać

Dowód. Za [**?**, strona 75]. Mamy dane liczby l_i spełniające nierówność Krafta, i chcemy skonstruować kod przedrostkowy.

Dla $w \in A^*$ i $l \geq \text{length}(w)$ niech

$$A(w,l) := \{v \in A^l : w \text{ jest prefiksem } v\} = \{wu : u \in A^{l-\text{length}(w)}\}.$$

Wtedy $|A(w,l)| = d^{l-\text{length}(w)}$. Oczywiście, jeżeli w_1, w_2 nie są prefiksami jeden drugiego, to $A(w_1,l)$ i $A(w_2,l)$ są rozłączne.

Bez straty ogólności zakładamy, że $1 \leq l_1 \leq \ldots \leq l_m$. Skonstruujemy $w_1, \ldots, w_m \in A^*$ takie, że żaden w_i nie jest prefiksem drugiego. Niech w_1 będzie dowolnie wybranym elementem A^{l_1}. Postępujemy indukcyjnie: zakładamy, że mamy zbudowane $w_1, \ldots, w_k$ takie, że żadne nie jest prefiksem drugiego. Pytamy się, czy istnieje $w_{k+1} \in A^{l_{k+1}}$ takie, że żadne z $w_1, \ldots, w_k$ nie jest jego prefiksem. Oczywiście takie w_{k+1} istnieje wtw gdy

$$A^{l_{k+1}} \setminus \bigcup_{i=1}^{k} A(w_i, l_{k+1}) \neq \emptyset.$$

Z uwagi u góry (i faktu, że są parami rozłączne) dostajemy

$$|\bigcup_{i=1}^{k} A(w_i, l_{k+1})| = \sum_{i=1}^{k} |A(w_i, l_{k+1})|$$

$$= \sum_{i=1}^{k} d^{l_{k+1}-\text{length}(w_i)} = d^{l_{k+1}} \sum_{i=1}^{k} d^{-l_i} < d^{l_{k+1}} \sum_{i=1}^{m} d^{-l_i} \leq d^{l_{k+1}} = |A^{l_{k+1}}|.$$

Czyli ten zbiór jest niepusty, i w konsekwencji znajdziemy szukany kod.

Z drugiej strony, jeżeli mamy kod prefiksowy $w_1, \ldots, w_m$ z długościami $l_1, \ldots, l_m$, to

$$w_m \in A^{l_m} \setminus \bigcup_{j=1}^{m-1} A(w_j, l_m),$$

czyli

$$1 \leq |A^{l_m}| - \sum_{j=1}^{m-1} |A(w_j, l_m)| = d^{l_m} - d^{l_m} \sum_{j=1}^{m-1} d^{-l_j},$$

i w konsekwencji

$$\sum_{j=1}^{m} d^{-l_j} = \frac{1}{d^{l_n}}(d^{l_m} \sum_{j=1}^{m} d^{-l_j})$$

$$= \frac{1}{d^{l_m}}(1 + d^{l_m} \sum_{j=1}^{m-1} d^{-l_j}) \leq \frac{1}{d^{l_m}} d^{l_m} = 1.$$

□

Zadanie 6.1. *Mając dane przykładowe $l_1, \ldots, l_m$ spełniające nierówność Krafta, proszę skonstruować kod prefiksowy o tych długościach.*

Twierdzenie 6.2. *Nierówność Krafta zachodzi dla dowolnych kodów jednoznacznie dekodowalnych.*

W konsekwencji możemy się zredukować to używania kodów przedrostkowych.

Dowód (na wykładzie nie było i nie obowiązuje). Podobnie jak poprzednio słowo w_i ma długość l_i. Rozważamy zdania o coraz większej potencjalnej długości (precyzyjnie mówiąc patrzymy na zdania składające się z k słów kodujących dla coraz większej ilości k).

Z założenia o jednoznacznej dekodowalności dostajemy, że

$$(i_1, \ldots, i_k) \neq (i'_1, \ldots, i'_k) \implies w_{i_1} \ldots w_{i_k} \neq w_{i'_1} \ldots w i'_k.$$

To oznacza, że dla dowolnego naturalnego r funkcja

$$\{(i_j)_{j=1..k} \subset \{1, \ldots, m\}^k \mid k \in \mathbb{N}, \sum_{j=1}^{k} l_j = r\} \ni (i_1, \ldots, i_k) \to w_{i_1} \ldots w_{i_k} \in A^r$$

jest iniektywna (przypominam, że A^r to zdania długości r). Niech $h(r)$ oznacza ile razy r wyraża się w postaci sumy $l_{i_1} + \ldots + l_{i_k}$. Ponieważ wielkość dziedziny jest $h(r)$ i obrazu A^r jest d^r, dostajemy, że

$$h(r) \leq d^r.$$

Teraz dla dowolnej liczby k mamy

$$(\sum_{i=1}^{m} d^{-l_i})^k = (\sum_{i=1}^{m} d^{-l_i}) \cdot \ldots \cdot (\sum_{i=1}^{m} d^{-l_i})$$

$$= \sum_{i_1} \ldots \sum_{i_k} d^{-(l_{i_1} + \ldots + l_{i_k})} = \sum_{r=1}^{kl_{\max}} \frac{h(r)}{d^r},$$

gdzie jak przypominam $h(r)$ oznacza ile razy r wyraża się w postaci sumy $l_{i_1} + \ldots + l_{i_k}$, a $l_{\max}$ to długość najdłuższego słowa. W konsekwencji

$$(\sum_{r=1}^{m} d^{-l_i})^k = \sum_{r=1}^{kl_{\max}} \frac{h(r)}{d^r} \leq \sum_{r=1}^{kl_{\max}} 1 = kl_{\max}.$$

Biorąc pierwiastek k-tego stopnia dostajemy oszacowanie w granicy przez 1. □

6.2 Wartość oczekiwana długości słowa – definicja entropii

Załóżmy, że mamy rozkład prawdopodobieństwa na $S = \{s_1, \ldots, s_m\}$, czyli litera s_i pojawia się z prawdopodobieństwem $p_i = p(s_i)$ (zakładamy dodatkowo, że źródło ma brak pamięci, to znaczy, że to co pojawi się następne nie zależy od tego co pojawiło się poprzednio).

Chcemy kodować zużywając statystycznie/średnio minimalną ilość pamięci. Załóżmy, że mamy dany alfabet kodujący $A = \{0, 1\}$ (czyli $d = 2$) i jednoznacznie dekodowalną funkcję kodującą $\varphi : S \to A^*$ (przyjmujemy $l_i = \text{length}\varphi(s_i)$).

Wartość oczekiwana długości słowa kodującego jest oczywiście dana wzorem

$$L = E(\text{length}(\varphi)) := \sum_{s \in S} p(s) \cdot \text{length}(\varphi(s)) = \sum_i p_i l_i.$$

Pytanie jak dobrać wartości l_i by minimalizować wartość oczekiwaną ilości pamięci. Ponieważ na podstawie nierówności Krafta wiemy jakie długości są dopuszczalne, dostajemy problem minimalizacji

$$L(l_1, \ldots, l_n) := \sum_i p_i l_i$$

przy warunku

$$\sum_i 2^{-l_i} \leq 1.$$

Zapominamy o tym, że są całkowite (dostaniemy przybliżenie), i wtedy możemy zwiększyć L zakładając równość. Otrzymaliśmy więc następujący problem:

Problem 6.1. *Znaleźć minimum*

$$L(r_1, \ldots, r_n) := \sum_i p_i r_i$$

przy warunku $\sum_i 2^{-r_i} = 1$.

Dowód. Rozwiązanie: wykorzystamy metodę mnożników Lagrange'a:

$$J(r_1, \ldots, r_n; \lambda) = \sum_i p_i r_i + \lambda(\sum_i 2^{-r_i} - 1).$$

Różniczkując dostajemy

$$\frac{\partial J}{\partial r_i} = p_i - \lambda 2^{-r_i} \ln 2,$$

i przyrównując do zera dostajemy

$$2^{-r_i} = p_i/(\lambda \ln 2).$$

Podstawiając do warunku na λ, dostajemy $\lambda = 1/\ln 2$, czyli

$$p_i = 2^{-r_i},$$

dając optymalne kody dla $\bar{r}_i = -\log_2 p_i$ i wartość oczekiwaną długości słowa kodującego

$$\sum_i p_i \bar{r}_i = -\sum_i p_i \log_2 p_i.$$

Można pokazać, że to jest minimum globalne □

Oczywiście, dla nas kluczowa jest sytuacja, gdy alfabet kodowy składa się z dwóch liter „0" i „1", i w konsekwencji dostajemy definicję *entropii*:

Definicja 6.3 (Definicja Entropii Shannona). Niech $X = \{x_i\}$ będzie dyskretną przestrzenią probabilistyczną, gdzie prawdopodobieństwo wylosowania punktu x_i wynosi p_i. Wtedy $h(X)$, *entropia* X, wyraża się wzorem

$$h(X) := \sum_i p_i \cdot -\log_2 p_i.$$

Można łatwo pokazać, że jeżeli mamy n elementów możliwych do wylosowania, to entropia szacuje się przez $\log_2 n$.

6.3 Shannon noiseless coding theorem

Kodowanie Shannona polega na zaokrągleniu w górę optymalnych wartości x_i za pomocą wzoru

$$l_i = \lceil x_i \rceil.$$

Korzystając z Tw. Krafta możemy teraz zbudować kod prefiksowy który realizuje te długości (nazwiemy go kodowaniem Shannona). Wtedy mamy

$$h((p_i)_i) = \sum_i -p_i \log_2 p_i = \sum_i p_i x_i \leq \sum_i p_i l_i$$

$$\leq \sum_i p_i \lceil x_i \rceil \leq \sum_i p_i (x_i + 1) = h((p_i)_i) + 1.$$

Widzimy więc, że wartość oczekiwania długości słowa $\sum_i p_i l_i$ przy kodowaniu Shannona nie przekracza wartości entropii plus jeden.

Przyda się nam pojęcie iloczynu kartezjańskiego dwóch przestrzeni probabilistycznych. Zakładamy, że mamy dwie dyskretne przestrzenie probabilistyczne $X = \{x_i\}$, p_i oraz $Y = \{y_j\}$, q_j. Definiujemy rozkład prawdopodobieństwa na iloczynie kartezjańskim $X \times Y$ wzorem

$$p(x_i, y_j) = p_{i,j} = p_i \cdot q_j.$$

W kategorii zmiennych losowych to jest równoważne stwierdzeniu, że X i Y są niezależne.

Przyjmujemy oznaczenie: $\mathrm{sh}(x) = x \cdot -\log_2 x$.

Obserwacja 6.1. *Mamy*

$$h(X \times Y) = h(X) + h(Y).$$

Dowód. Mamy

$$\sum_{ij} \mathrm{sh}(x_i \cdot y_i) = \sum_i x_i \sum_j \mathrm{sh}(y_j) + \sum_i \mathrm{sh}(x_i) \cdot \sum_j y_j.$$

□

Rozumując przez indukcję, otrzymujemy, że

$$h(X_1 \times \ldots \times X_n) = h(X_1) + \ldots + h(X_n).$$

Wniosek 6.1 (Shannon noiseless coding theorem)**.** *Niech będzie dane źródło bez pamięci. Możemy dowolnie blisko zbliżyć się do entropii przy pomocy kodowania.*

Dowód. Zamiast kodować litery z alfabetu S, będziemy kodować słowa n-elementowe, czyli elementy $S \times \ldots \times S$ (naszym nowym alfabetem źródłowym stają się słowa o długości n z alfabetu S). Entropia $h(S^n)$ na podstawie wcześniejszych wzorów wyraża się przez

$$h(S^n) = nh(S).$$

Teraz na podstawie wcześniejszych uwag możemy znaleźć kodowanie, dla którego oczekiwana wartość długości kodu nie przekracza $nh(S) + 1$. W konsekwencji, kodując dłuższe ciągi liter, statystycznie na jeden element z S będziemy potrzebowali $(nh(S) + 1)/n$ (bo kodujemy słowa długości n). Czyli biorąc n odpowiednio duże możemy zbliżyć się do granicy $h(S)$ dowolnie blisko. □

6.4 MLE vs dywergencja Kullbacka-Leiblera

Teraz przejdziemy do jednej z ważniejszych modyfikacji entropii. Idea polega na dokonywaniu kompresji danej zmiennej losowej przy pomocy kodu dopasowanego do drugiej.

Załóżmy, że mamy zmienną losową Y (rozkład q_i), i kod dopasowany do rozkładu X (p_i). Wtedy przez entropię krzyżową rozumiemy

$$H^{\times}(Y\|X) := \sum_i q_i \cdot (-\log_2 p_i).$$

Załóżmy, że chcielibyśmy zobaczyć jaka jest różnica między kodowaniem za pomocą kodu dopasowanego do X, a optymalnym:

$$H^{\times}(Y\|X) - H(Y) = \sum_i q_i \log(q_i/p_i).$$

Tą różnicę oznaczamy $D_{KL}(Y\|X)$ i nazywamy różnicą Kullbacka-Leiblera (spotyka się także inne nazwy, typu entropia relatywna).

Załóżmy teraz, że mamy rodzinę gęstości kodujących $\mathcal{F}$ i chcemy z nich wybrać najlepsze:

$$H^{\times}(Y\|\mathcal{F}) := \inf H^{\times}(Y\|f).$$

Wtedy możemy to zapisać równoważnie:

$$H^{\times}(Y\|\mathcal{F}) = \inf_{f\in\mathcal{F}} \sum_i q_i \cdot (-\log_2 f_i). \tag{6.1}$$

Interpretacja: estymacja Metodą Największej Wiarygodności.

Uwaga 6.1. Załóżmy, że wylosowaliśmy (mamy dane) punkty y_i. Metoda największej wiarygodności polega na szukaniu spośród rodziny rozkładów $\mathcal{F}$ (zdefiniowanych na y_i) tego który najlepiej „przybliża" dane (czyli takiego, że prawdopodobieństwo wylosowania ciągu y jest maksymalne). Przy ustalonym $f \in \mathcal{F}$ prawdopodobieństwo wylosowania ciągu $(y_1, \ldots, y_n)$ wynosi $f_1 \cdot \ldots \cdot f_n$. W konsekwencji szukamy f które realizuje

$$\sup_{f\in\mathcal{F}} f_1 \cdot \ldots \cdot f_n.$$

Rozpatrzmy teraz analogiczną sytuację, gdy y_i wylosowaliśmy k_i razy (ponieważ długość ciągu y wynosi n, wtedy częstość pojawienia się wynosi $q_i = k_i/n$). Wtedy szukamy takiego $f \in \mathcal{F}$ które realizuje

$$\sup_{f\in\mathcal{F}} f_1^{k_1} \cdot \ldots \cdot f_n^{k_n}.$$

To jest równoważne szukaniu $f \in \mathcal{F}$ które realizuje

$$\sup_{f\in\mathcal{F}} k_1 \log_2 f_1 + \ldots + k_n \log_2 f_n = n \sup_{f\in\mathcal{F}} \sum_i q_i \log_2 f_i.$$

Ale to jest dokładnie to samo co wyprowadzone wcześniej we wzorze (6.1). Jak widzimy, szukanie optymalnej gęstości kodującej prowadzi do tych samych wyników co estymacja metodą największej wiarygodności.

6.5 Entropia różniczkowa

Powstaje naturalne pytanie, jak należy postępować w sytuacji gdy rozpatrywane zmienne posiadają ciągły rozkład? To się zdarza przy zapisywaniu dźwięku (ogólnie sygnałów analogowych). Wtedy najpierw dokonujemy zawsze kwantyzacji (dyskretyzacji) z krokiem δ, czyli mianowicie zamiast wektora losowego X rozważamy jego dyskretyzację, to jest

$$X_\delta := \lfloor \delta X \rfloor / \delta.$$

Twierdzenie 6.3. *Niech $\mathbb{X}$ wektor losowy o gęstości f na $\mathbb{R}^N$. Zakładamy dodatkowo, że f jest ciągła i ma zwarty support*[2]*. Wtedy*

$$\lim_{\delta \to 0} \left[h(X_\delta) - N \log_2 \delta - \int \mathrm{sh}(f(x))dx \right] = 0.$$

Dowód. Idea jest podobna do zbieżności całki Riemanna. Dla prostoty zawężam się do sytuacji jednowymiarowej, nie robię także superścisłych oszacowań.

Niech $x_i = i\delta$ (początek i-tego przedziału). Wtedy

$$h(X_\delta) = \sum_i \mathrm{sh}(p_i)$$

gdzie

$$p_i = \mu_X([x_i, x_{i+1}) = \int_{[x_i, x_{i+1})} f(x)dx \approx f(x_i)\delta.$$

Stosując to przybliżenie w powyższym wzorze dostajemy

$$h(X_\delta) \approx \sum_i \mathrm{sh}(f(x_i)\delta) = \sum_i (-f(x_i)\log_2 f(x_i))\delta + \sum_i f(x_i)\delta \cdot (-\log_2 \delta)$$

$$\approx \int \mathrm{sh}(x)dx + \int f(x)dx(-\log_2 \delta) = \int \mathrm{sh}(x)dx - \log_2 \delta.$$

□

Czyli w sposób naturalny prowadzi to do następującej definicji (entropia różniczkowa to asymptotyka):

Definicja 6.4. Przez entropię różniczkową zmiennej ciągłej X o gęstości f rozumiemy

$$h(X) := \int \mathrm{sh}(f(x))dx.$$

Analogicznie jak w przypadku dyskretnym możemy rozważać *nierówność Krafta w wersji ciągłej*:

$$\int 2^{-l(x)}dx \leq 1.$$

Przykład 6.1. Łatwo przeliczyć, że jeżeli X ma rozkład jednostajny na zbiorze W, to

$$h(X) = \log_2(\lambda_N(W)).$$

[2]tak naprawdę twierdzenie idzie przy znacznie słabszych założeniach

Przez analogię do entropii krzyżowej dla zmiennych dyskretnych definiujemy

$$H^{\times}(Y\|X) := \int q(x) \cdot (-\log_2 p(x))dx,$$

gdzie q to gęstość Y, a p gęstość X. Analogicznie także definiuje się dywergencje Kullbacka-Leiblera wzorem

$$D_{KL}(Y\|X) := \int q(x) \cdot \log_2(q(x)/p(x))dx.$$

Rozdział 7

Rozkład normalny

7.1 Dlaczego rozkład normalny?

Jest dużo powodów:

1. jest to odpowiednik funkcji kwadratowej
2. CTW
3. maksymalizacja entropii
4. niezmienniczy na różne operacje

Łatwo sprawdzić, że dla A liniowego mamy

$$E(AX+b) = AEX + Ab,\ \mathrm{cov}(AX+b) = A\mathrm{cov}(X)A^T.$$

Co więcej, jeżeli X ma rozkład gęstości f, to $AX+b$ ma rozkład $|\det(A)|f(A^{-1}(y-b))$.

Powyższa obserwacja może służyć do losowania z rozkładu normalnego wielowymiarowego. Zacznijmy od rozkładu $N(0,1)$ który wiemy jak losować. Następnie zmienna X o rozkładzie $N(0,I_N)$ możemy wylosować biorąc kolejne współrzędne z rozkładu normalnego jednowymiarowego, gęstość jego jest dana wzorem

$$f(x) = \frac{1}{(2\pi)^{N/2}} \exp(-\|x\|^2/2).$$

Gdybyśmy teraz chcieli losować z normalnego o średniej m i macierzy kowariancji Σ, to na podstawie wcześniejszych wystarczy wziąć $\Sigma^{1/2}X + m$, oraz gęstość dana jest wzorem

$$\frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp(-\|x-m\|_\Sigma^2/2),$$

gdzie $\|w\|_\Sigma^2 = (\Sigma^{-1/2}w)^T(\Sigma^{-1/2}w) = w^T\Sigma^{-1}w$ jest kwadratem odległości Mahalanobisa.

Z punktu widzenia statystyki, najważniejszym rozkładem jest rozkład normalny. Wzór na gęstość:

$$\mathcal{N}(m,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(x-m)^2/\sigma^2).$$

Jest ku temu wiele powodów:

1. Centralne twierdzenie graniczne: jeżeli mamy niezależne zmienne losowe o tym samym rozkładzie ze średnią m i wariancją σ^2, to

$$\frac{X_1 + \ldots + X_n}{\sqrt{n}} \to \mathcal{N}(m, \sigma^2).$$

2. wzór wymaga tylko dwóch parametrów, a dobrze opisuje sporą klasę zjawisk
3. logarytm jest funkcją kwadratową (mle za chwilę)

7.2 Wyprowadzenie rozkładu normalnego

Postaram się wyprowadzić wielowymiarowy rozkład normalny korzystając z idei, że chcemy by to był rozkład który możemy efektywnie estymować dla danych, korzystając z zasady maksymalnej wiarygodności (maximal likelihood).

Chciałbym w związku z tym zacząć od przypomnienia zasady maksymalnej wiarygodności. Mając daną rozdzinę rozkładów $(f_\theta)_{\theta\in\Theta}$, chcemy dopasować ją tak by optymalnie pasowało do próbki X. W tym celu rozpatrujemy (asymptotyczne) prawdopodobieństwo wylosowania X:

$$l(X, f_\theta) = \prod_i f_\theta(x_i).$$

Chcemy powyższe zmaksymalizować względem $\theta \in \Theta$, ponieważ jest zazwyczaj łatwiej pracować z sumą niż iloczynem, rozpatrujemy

$$\log l(X, f_\theta) = \sum_i \log f_\theta(x_i).$$

W związku z czym de facto pracujemy na logarytmie z gęstości.

W konsekwencji chcielibyśmy, by ten logarytm był możliwie najprostszą funkcją. Zauważmy, że nie może to być funkcja liniowa, gdyż gdyby logarytm funkcji był liniowy, to funkcja miałaby całkę nieskończoność. W konsekwencji naprostszym wyborem jest przyjąć by logarytm był funkcją kwadratową

$$\log f(x) = -\tfrac{1}{2}x^2 + c.$$

Dobór współczynnika $-1/2$ pełni rolę normalizacyjną, i wyjaśni się później (oczywiście, nie może być dodatni, bo znowu całka nie byłaby skończona). W konsekwencji mamy

$$f(x) = C \cdot \exp(-x^2/2).$$

Musimy jeszcze dobrać C tak, aby całka z f była równa jeden:

$$C \int_{\mathbb{R}} \exp(-x^2/2)dx = 1.$$

Niestety, funcja $\exp(-x^2/2)$ nie da się pocałkować w klasie funkcji elementarnych, więc postaramy się zastosować trik do policzenia $\int_{\mathbb{R}} \exp(-x^2/2)dx$. Otóż spróbujemy policzyć

$$\int_{\mathbb{R}} \exp(-x^2/2)dx \int_{\mathbb{R}} \exp(-y^2/2)dy.$$

Pozornie wydaje się to trudniejsze, ale zauważmy, że powyższą całkę możemy zapisać jako

$$\int_{\mathbb{R}^2} \exp(-(x^2+y^2)/2)dxdy.$$

I teraz zauważamy, że funkcja ta zależy jedynie od odległości od zera! W związku z czym robimy zmianę zmiennych na biegunowe, to znaczy

$$r = \sqrt{x^2 + y^2} \in [0, \infty), \phi : \{\cos\phi = x/r, \sin\phi = y/r\} \in [0, 2\pi).$$

Jakobian wynosi r, co znaczy, że nasza całka sprowadza się do

$$\int_{\mathbb{R}^2} \exp(-(x^2+y^2)/2)dxdy = \int_0^{2\pi}\int_0^{\infty} \exp(-r^2/2)rdrd\phi = 2\pi\int_0^{\infty}\exp(-r^2/2)rdr.$$

Robiąc podstawienie $u = r^2/2$ dostajemy

$$\int_0^{\infty}\exp(-r^2/2)rdr = \int_0^{\infty}\exp(-u)du = 1,$$

co w konsekwencji oznacza, że

$$\int_{\mathbb{R}}\exp(-x^2/2)dx = \sqrt{2\pi} \text{ czyli } C = \frac{1}{\sqrt{2\pi}}.$$

W konsekwencji dostaliśmy gęstość

$$N(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2).$$

Niech $\mathbb{X}$ oznacza zmienną losową o gęstości N. Ponieważ N jest parzyste, oczywiście wartość oczekiwana wynosi zero:

$$E\mathbb{X} = \int xN(x)dx = 0.$$

Łatwo policzyć korzystając z całkowania przez części

$$V\mathbb{X} = \int x^2N(x)dx = 1,$$

co oznacza, że odchylenie standardowe $\mathbb{X}$ wynosi 1.

Ponieważ chcielibyśmy, aby nasza rodzina gęstości była niezmiennicza na dowolne przekształcenia afiniczne. Gęstość po takiej transformacji $x \to \sigma x + m$, to znaczy gęstość zmiennej losowej $\sigma\mathbb{X} + m$ będzie wynosić

$$\mathcal{N}(m, \sigma^2)(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\tfrac{(x-m)^2}{2\sigma^2}),$$

a średnia będzie wynosić m, zaś wariancja σ^2. Powyższy wzór daje definicję rozkładu normalnego jednowymiarowego.

7.3 Generowanie punktów z rozkładu normalnego

Schemat Boxa-Mullera generowania liczb losowych z rozkładu normalnego powiela ideę obliczenia stałej normalizacyjnej. Ponieważ funkcji dystrybuanty nie da się wyliczyć za pomocą funkcji elementarnych, nie można do generowania użyć funkcji odwrotnej do dystrybuanty.

Będziemy starali się wygenerować punkty na płaszczyźnie których obie niezależne współrzędne będą pochodzić z rozkładu normalnego: $\mathbb{X}$ i $\mathbb{Y}$. W tym celu zmienimy układ na biegunowy, a mianowicie będziemy generować zmienne losowe odpowiadające za promień $\mathbb{R}$ i kąt Θ:

$$\mathbb{R} = \sqrt{\mathbb{X}^2 + \mathbb{Y}^2} \text{ oraz } \Theta : \cos\Theta = \mathbb{X}/\mathbb{R}, \sin\Theta = \mathbb{Y}/\mathbb{R}.$$

Oczywiście kąt ma rozkład jednostajny na przedziale $[0, 2\pi)$, zaś podobnie jak w wyliczeniu całki można pokazać, że $\mathbb{S} = \mathbb{R}^2$ ma rozkład wykładniczy o parametrze $\lambda = 2$ [ćwiczenie]. Konkludując losujemy parę punktów z rozkładu normalnego za pomocą wzorów:

$$\mathbb{X} = \mathbb{R}\cos(\Theta) = \sqrt{-2\ln\mathbb{U}}\cos(2\pi\mathbb{V}),$$

oraz

$$\mathbb{Y} = \mathbb{R}\sin(\Theta) = \sqrt{-2\ln\mathbb{U}}\sin(2\pi\mathbb{V}),$$

gdzie $\mathbb{U}$ i $\mathbb{V}$ pochodzą z rozkładu jednostajnego na odcinku $[0, 1)$.

Jeżeli chcemy wylosować punkt z $\mathcal{N}(m, \sigma^2)$, to losujemy zmienną $\mathbb{X}$ z $\mathcal{N}(0, 1)$, i kładziemy $\mathbb{Y} = \sigma\mathbb{X} + m$.

7.4 Estymacja parametrów

Załóżmy, że mamy próbkę $X = (x_i)_{i=1..n} \subset \mathbb{R}$, i chcemy do niej dopasować optymalne parametry rozkładu normalnego m i σ^2, tak aby zgodnie z MLE było optymalnie.

Spróbujemy teraz wyprowadzić w jaki sposób powinien być zdefiniowany rozkład normalny wielowymiarowy. Rozpatrzmy więc $\mathcal{N}(m, \sigma^2)$ i spróbujmy policzyć koszt log-likelihood dla próbki X:

$$\log\mathrm{L}(X, \mathcal{N}(m, \sigma^2)) = \sum_i \log(\mathcal{N}(m, \sigma^2)(x_i)) =$$

$$= \tfrac{n}{2}\ln(2\pi) + \tfrac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i - m)^2.$$

Oczywiście, przy ustalonym σ, wiemy, że powyższe równanie minimalizuje się dla

$$m = \mathrm{mean}X = \frac{1}{n}\sum_i x_i.$$

Przy tak ustalonym m zajmijmy się minimalizacją względem σ, dla prostoty przyjmijmy sobie oznaczenie $S = \sigma^{-2}$. Wtedy mamy funkcję

$$S \to \tfrac{n}{2}\ln(2\pi) - \tfrac{n}{2}\ln(S) - \frac{S}{2}\sum_i (x_i - m)^2,$$

której pochodną przyrównując do zera dostajemy równanie

$$\frac{1}{S} = \tfrac{1}{n}\sum_i (x_i - m)^2,$$

co oznacza, że dostajemy wzór

$$\sigma^2 = \mathrm{Var}X \text{ czyli } \sigma = \sigma_X.$$

W konsekwnecji optymalny wybór dla m i σ to średnia i odchylenie standardowe z próbki. Analogicznie to podejścia z kompresji, można to traktować jako wyprowadzenie wartości średniej i odchylenia standardowego.

7.5 Rozkład normalny wielowymiarowy

Postaramy się teraz zdefiniować rozkład normalny wielowymiarowy. Zacznijmy od najprostszej definicji gęstości dla rozkładów wielowymiarowych, a mianowicie, jeżeli mamy gęstości na prostej $f_1, \ldots, f_D$, to definiujemy gęstość $F = f_1 \otimes \ldots \otimes f_d$ na $\mathbb{R}^D$ za pomocą iloczynu

$$F(x_1, \ldots, x_D) = f_1(x_1) \cdot \ldots \cdot f_D(x_D).$$

Zgodnie z definicją, mamy, że współrzędne są od siebie niezależne, i w konsekwencji macierz kowariancji jest diagonalna, łatwo sprawdzić, że

$$\mathrm{mean}F = [\mathrm{mean}f_1, \ldots, \mathrm{mean}f_D]^T \text{ oraz } \mathrm{cov}F = \mathrm{diag}(\mathrm{Var}f_1, \ldots, \mathrm{Var}f_D).$$

W sensie zmiennych losowych, powyższa gęstość to gęstość zmiennej losowej $[\mathbb{X}_1, \ldots, \mathbb{X}_D]^T$, gdzie $\mathbb{X}_i$ to są niezależne zmienne losowe o gęstościach f_i.

Stosując powyższą procedurę dla rozkładu normalnego $\mathcal{N}(0,1)$ dostajemy rozkład

$$N(x_1, \ldots, x_D) = \mathcal{N}(0,1)(x_1) \cdot \ldots \cdot \mathcal{N}(0,1)(x_D) = \frac{1}{\sqrt{(2\pi)^D}} \exp(-\tfrac{1}{2}\|x\|^2).$$

Korzystając z powyżeszego, dostajemy, że jeżeli $\mathbb{X}$ ma gęstość N

$$\mathrm{mean}\mathbb{X} = 0,\ \mathrm{cov}\mathbb{X} = I.$$

Aby zdefiniować ogólny rozkład wielowymiarowy, zastosujemy analogiczną metodę do sytuacji jednowymiarowej, a mianowicie będziemy chcieli rozszerzyć o transformacje afiniczne $x \to Ax + b$.

Niech więc $\mathbb{X}$ ma gęstość $N(x)$, i policzmy gęstość $g(y)$ zmiennej $\mathbb{Y} = A\mathbb{X} + b$. Najpierw jedynie zauważmy, że

$$E\mathbb{Y} = AE\mathbb{X} + b \text{ oraz } \mathrm{cov}\mathbb{Y} = A\mathrm{cov}XA^T,$$

co oznacza, że

$$E\mathbb{Y} = b \text{ oraz } \mathrm{cov}\mathbb{Y} = AA^T.$$

Oznaczenie $b = m$.

Korzystając z wzoru ? wyprowadzonego wcześniej, mamy dla odwracalnych Φ:

$$g_{\Phi(\mathbb{X})}(y) = \frac{1}{|d\Phi(g^{-1}(y))|} f_{\mathbb{X}}(g^{-1}(y)).$$

gdzie g oznacza gęstość $\Phi(\mathbb{X})$. Stosując powyższe do naszego g, dostajemy

$$g(y) = \frac{1}{|A|} N(A^{-1}(y-m)) = \frac{1}{\sqrt{2\pi}|A|} \exp(-\tfrac{1}{2}\|A^{-1}(y-m)\|^2).$$

Korzystając z $\|w\|^2 = w^T w$, oraz wprowadzając oznaczenia $\Sigma = AA^T$, $\|w\|_\Sigma^2 = w^T\Sigma^{-1}w$ (norma Mahalanobisa), powyższy wzór upraszczamy do końcowego wzoru na rozkład normalny o średniej m i kowariancji Σ:

$$\mathcal{N}(m,\Sigma)(y) = g(y) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} \exp(-\tfrac{1}{2}\|y-m\|_\Sigma^2).$$

RYSUNKI: poziomice, etc.

7.6 log-likelihood

Niech $X = (x_i)_{i=1..n}$ będzie dane. Dostajemy wzór

$$\mathrm{logL}(X, \mathcal{N}(m, \Sigma)) = \sum_i \ln\left((2\pi)^D |\Sigma|\right)^{-1/2} \exp(-\tfrac{1}{2}\|x_i - m\|^2_\Sigma)\big)$$

$$= -\frac{Dn}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \tfrac{1}{2}\sum_i \|x_i - m\|^2_\Sigma.$$

Aby znaleźć maksimum powyższej funkcji, rozpatrzymy jej gradient ze względu na m i Σ.

Oznaczmy (przy ustalonym Σ)

$$\phi(m) = -\frac{Dn}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \tfrac{1}{2}\sum_i \|x_i - m\|^2_\Sigma.$$

Przypominam, że gradient ∇ to operacja odpowiadająca pochodnej, ale jest zdefiniowana tylko dla funkcji skalarnych ψ. Jeżeli $\psi : E \to \mathbb{R}$ jest funkcją skalarną (gdzie E to przestrzeń z iloczynem skalarnym), to gradient $\psi(x)$ to jedyny elementem przestrzeni $w \in E$, taki, że

$$\psi(x + h) = \psi(x) + \langle w, h\rangle + o(h).$$

Wtedy oczywiście

$$\nabla\phi(m) = \nabla(\sum_i \|x_i - m\|^2_\Sigma) = \sum_i \nabla\|x_i - m\|^2_\Sigma.$$

Na podstawie wcześniejszych faktów (Przykład ...) korzystając z symetryczności Σ, wiemy, że gradient funkcji (ze względu na zmienną m)

$$m \to \|x_i - m\|^2_\Sigma = \|m - x_i\|^2_\Sigma = \langle (m - x_i), \Sigma^{-1}(m - x_i)\rangle$$

wynosi

$$\Sigma^{-1}(m - x_i) + (\Sigma^{-1})^T(m - x_i) = 2\Sigma^{-1}(m - x_i).$$

Konkludując

$$\nabla\phi(m) = \sum_i 2\Sigma^{-1}(x_i - m).$$

Przyrównując gradient do zera, dostajemy w oczywisty sposób, że

$$m = \frac{1}{n}\sum_i x_i = \mathrm{mean}X.$$

Otrzymujemy w ten sposób, że aby optymalnie dopasować rozkład $\mathcal{N}(m, \Sigma)$ do danych, przy zmiennym m i zafiksowanym Σ, powinniśmy wycentrować rozkład normalny w środku zbioru danych X. Można w ten sposób powiedzieć, że optymalizacja rozkładu normalnego prowadzi do średniej.

Teraz zajmiemy się szukaniem optymalnego Σ. W tym celu pokażemy parę faktów dotyczących gradientu funkcji macierzowych. Przypominam, że iloczyn skalarny dwóch macierzy $A, B \in \mathbb{R}^{n\times k}$ wyraża się wzorem

$$\langle A, B\rangle = \mathrm{tr}(A^T B) = \mathrm{tr}(AB^T) = \sum_{i,j} a_{ij}b_{ij}.$$

Aby przejść dalej będziemy potrzebować następujące obserwacje.

Lemat 7.1. *Mamy*

$$\det(I+H) = 1 + \operatorname{tr}(H) + o(H).$$

Dowód. Zgodnie z definicją wyznacznika mamy (przez ϵ oznaczam permutację identycznościową):

$$\det(I+H) - 1 = \sum_{\sigma \text{ permutacja}} \prod_{i=1}^{D} (I+H)_{i\sigma(i)} - 1 = \sum_{\sigma=\epsilon} \prod_i (I+H)_{i\sigma(i)} - 1 + \sum_{\sigma\neq\epsilon} \prod_i (I+H)_{i\sigma(i)}$$

$$= (1+H_{11}) \cdot \ldots \cdot (1+H_{DD}) - 1 + \sum_{\sigma\neq\epsilon} \prod_i (I+H)_{i\sigma(i)}.$$

Zajmiemy się teraz analizą obu składników w powyższym wzorze. Oczywiście

$$(1+H_{11}) \cdot \ldots \cdot (1+H_{DD}) - 1 = H_{11} + \ldots + H_{DD} + o(H) = \operatorname{tr}(H) + o(H).$$

Rozpatrzmy drugi czynnik. Weźmy więc dowolną permutację σ która nie jest permutacją identycznościową. Oznacza to, że istnieje takie j, że

$$\sigma(j) = k \neq j.$$

W konsekwencji oczywiście $\sigma(k) \neq k$ (w przeciwnym razie nie byłaby to permutacja). Oznacza to, że indeksy jk i $k\sigma(k)$ są poza przekątną, co oznacza, że

$$(I+H)_{jk} = h_{jk} \text{ oraz } (I+H)_{k\sigma(k)} = h_{k\sigma(k)}$$

i w konsekwencji

$$\prod_{i=1}^{D} (I+H)_{i\sigma(i)} = h_{jk}h_{k,\sigma(k)} \cdot \prod_{i:i\neq j,k} (I+H)_{i\sigma(i)} \in o(H).$$

□

Stosujemy oznaczenie

$$A^{-T} = (A^{-1})^T.$$

Stwierdzenie 7.1. *Mamy*

$$\nabla \det(A) = \det A \cdot A^{-T}.$$

Dowód. Mamy

$$\det(A+H) = \det A \cdot (\det(I + A^{-1}H)) = \det A \cdot (1 + \operatorname{tr}(A^{-1}H) + o(A^{-1}H))$$

$$= \det A \cdot (1 + \langle A^{-T}, H\rangle + o(H)) = \det(A) + \langle \det A \cdot A^{-T}, H\rangle + o(H).$$

□

Wniosek 7.1. *Mamy*

$$\nabla \ln(\det A) = A^{-T}.$$

Dowód. Korzystamy z tego, że

$$\nabla f(g(x)) = f'(g(x)) \cdot \nabla g(x).$$

Wtedy

$$\nabla \ln(\det A) = \frac{1}{\det A} \cdot \det A A^{-T} = A^{-T}.$$

□

Lemat 7.2. *Mamy*

$$(I+H)^{-1} = I - H + o(H).$$

Dowód. Jest to bezpośredni wniosek z wzoru na sumę szeregu geometrycznego

$$(I-Q)^{-1} = I + Q + Q^2 + \ldots + Q^n + \ldots \text{ dla } Q : \|Q\| < 1.$$

□

Stwierdzenie 7.2. *Mamy*

$$(\Sigma+H)^{-1} = \Sigma^{-1} - \Sigma^{-1}H\Sigma^{-1} + o(H).$$

Dowód. Ponieważ

$$(AB)^{-1} = B^{-1}A^{-1},$$

mamy

$$(\Sigma+H)^{-1} = \Sigma^{-1}(I+H\Sigma^{-1})^{-1} = \Sigma^{-1} - \Sigma^{-1}H\Sigma^{-1} + o(H).$$

□

Zacznijmy od oznaczenia: jeżeli $F(x_1, \ldots, x_n)$ jest funkcją n zmiennych, to przez $\nabla_{x_i} F$ oznaczam gradient względem i-tej zmiennej.

Stwierdzenie 7.3. *Mamy (u, v ustalone)*

$$\nabla_\Sigma u^T\Sigma^{-1}v = -\Sigma^{-T}uv^T\Sigma^{-T}.$$

Dowód. Mamy (stosujemy „trace trick": $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ i własność $(AB)^T = B^TA^T$):

$$u^T(\Sigma+H)^{-1}v - u^T\Sigma^{-1}v = -u^T\Sigma^{-1}H\Sigma^{-1}v + o(H)$$
$$= -\mathrm{tr}(u^T\Sigma^{-1}H\Sigma^{-1}v) + o(H) = -\mathrm{tr}(\Sigma^{-1}vu^T\Sigma^{-1}H) + o(H)$$
$$= -\langle \Sigma^{-T}uv^T\Sigma^{-T}, H\rangle + o(H).$$

□

Możemy teraz przejść do głównego wyniku.

Twierdzenie 7.1. *Niech*

$$\Phi(\Sigma) = -\frac{Dn}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \tfrac{1}{2}\sum_i \|x_i - m\|^2_\Sigma.$$

Wtedy

$$\nabla\Phi(\Sigma) = -\frac{n}{2}\Sigma^{-T} + \frac{1}{2}\sum_i \Sigma^{-T}(x_i-m)(x_i-m)^T\Sigma^{-T}.$$

Dowód. Korzystamy z poprzednich wyników. □

Na podstawie powyższego twierdzenia, dostajemy, że gradient się zeruje gdy

$$-\frac{n}{2}\Sigma^{-T} + \frac{1}{2}\sum_i \Sigma^{-T}(x_i-m)(x_i-m)^T\Sigma^{-T} = 0,$$

czyli po przemnożeniu przez Σ^{-T} z prawej i lewej strony mamy

$$\Sigma^T = \frac{1}{n}\sum_i (x_i-m)(x_i-m)^T \text{ czyli } \Sigma = \Big(\frac{1}{n}\sum_i (x_i-m)(x_i-m)^T\Big)^T = \frac{1}{n}\sum_i (x_i-m)(x_i-m)^T = \mathrm{cov}X.$$

Uwaga 7.1. Rozkład osobliwy – regularyzacja.

ε - maszynowe

Ledoit-Wolff - wzór i motywacja.

Jako druga opcja - cross-validacja

Rozdział 8

Estymacja gęstości

8.1 Motywacja

Chcemy wnioskować o rozkładzie gęstości na podstawie próbki. Jedyna w miarę wiarygodna weryfikacja przez walidację krzyżową. Wygrywa ten, który ma lepsze MLE na zbiorze testującym.

Zazwyczaj plus polega na tym, że mamy możliwość generowania z danych (ale to też się robi – na następnym wykładzie będzie o modelach generatywnych).

Jeżeli mamy taką estymację, to możemy zrobić klastrowanie bazujące na gęstości.

Uwaga 8.1. Ostrzeżenie – klątwa wymiarowości, kula, kwadrat, etc

W wysoko-wymiarowej kostce w zasadzie wszystko jest na brzegu.

Kula stanowi mały procent w kostce.

Rozkład normalny koncenttruje się na otoczeniu sfery.

8.2 Histogram i DBScan

histogram – dzielimy na pudełeczka, i zliczamy ilość w każdym. Czyli

$$f \sim \frac{1}{N} \sum_i p_{x_i}$$

gdzie p_x to funkcja stała na kostce $h[)^D$ o całce równej jeden.

Rozsądniej byłoby te pudełeczka ustalać tak, by były zcentrowane w punktach zbioru danych Wtedy mamy

$$f(x) \sim \frac{1}{N} \sum_i K(x - x_i)$$

gdzie K to byłaby przeskalowana funkcja charakterystyczna kuli o promieniu ε (normalizacja).

jaka jest wartość w danym pudełeczku? To byłby estymator gęstości? bez normalizacji

$$f(x) \sim \mathrm{card}\{i : x \in B(x_i, \varepsilon)\} = \mathrm{card}\{i : d(x, x_i) \leq \varepsilon\}.$$

To podejście służy jako idea DBScan.

Opiszmy w tym celu jeszcze klastrowanie gęstościowe. Gdybyśmy mieli gęstość f, to rozbijamy na klastry korzystając z poziomu $C > 0$. Traktujemy gęstość jako góry, i poruszamy się tylko po wysokości $\geq C$. Składowe spójne stanowią nasze klastry (rysunek był na wykładzie).

Rzeczy o gęstości mniejszej niż C traktujemy jako outliersy (rzeczy rzadko występujące). W praktyce ustalamy C tak, aby ilość outliersów była między 10-30 procent.

DBScan (wersja uproszczona). Ustalamy jak w powyższej estymacji ε i kładziemy dla każdego punktu $x_k \in X$

$$f_k = \text{card}\{i \neq k : d(x_i, x_k) \leq \varepsilon\}.$$

Tworzymy graf w którym łączymy krawędzią wszystkie punkty nie dalsze niż ε. Dla C definiujemy

$$X_C = \{x_k \in X : f_k \geq C\}.$$

Składowe spójne stanowią nasze klastry.

8.3 Metody jądrowe

Mamy $X \subset \mathbb{R}^D$. W poprzednim trochę głupie było branie ostrych.

Więc ustalamy jakąś funkcję nieujemną K całkowalną do jedynki, którą kładziemy w każdym punkcie:

$$f(x) \sim \frac{1}{N} \sum_i K(x - x_i).$$

W praktyce jeszcze skalujemy jądro:

$$f(x) \sim \frac{1}{N} \sum_i \frac{1}{h^D} K(\frac{x - x_i}{h}).$$

Stosuje się najczęściej jądro gaussowskie $K(x) = N(0, I)$.

Powstaje oczywiście pytanie jak ustalić szerokość jądra, jak za mała to mamy overfitting, a jak za duża, to jest znowy za bardzo rozmyte.

Podejście Silvermana - optymalny dla znormalizowanych gaussowskich.

lepsza cross-validacja

różne modele gaussowskie najpierw - przeskalowana identyczność i diagonalne.

8.4 GMM

EM - jako idea, do czego służy

celem jest sytuacja, gdy mamy pewne zmienne ukryte (hidden) które chcemy oszacować do wyjaśnienia rzeczywistości.

Zastosowanie do GMM (hidden). Celem jest dokonanie estymacji gęstości w klasie rozkładów

$$p_1 N(m_1, \Sigma_1) + \ldots + p_n N(m_n, \Sigma_n).$$

postarać się wyjaśnić ideę

Wzór rekurencyjny – algorytm podobny w sensie działania do k-means:

1. startujemy wielokrotnie od gaussów (równe p_i), środki w losowych punktach zbioru danych, kowariancje równe kowariancji zbioru

2. dla każdego z punktów x ze zbioru X zbieramy wagi w jakim stopniu każdy gauss tłumaczy:

$$w_i(x) = p_i N_i(x) / (\sum_j p_j N_j(x))$$

3. obliczamy ważone średnie m_i i kowariancje Σ_i, natomiast waga to sumaryczna waga tłumacząca

$$p_i = \frac{1}{n} \sum_j w_i(x_j) = \frac{\sum_j w_i(x_j)}{\sum \sum}$$