Analiza Danych

J. Tabor

6 czerwca 2017

Spis treści

ZASADY: Ocena końcowa będzi obliczana jako 40% wyniku z ćwiczeń + 60 % wyniku z pisemnego egzaminu. Egzamin pisemny będzie na podstawie wykładu.

Osoba prowadząca ćwiczenia może zwolnić maksymalnie do 4 najlepszych studentów z egzaminu pisemnego (ocena z ćwiczeń jest wtedy przepisywana jako ocena z przedmiotu).

Rozdział 1

Plany

- dane - typy danych: tekst, dźwięk (muzyka,mowa), grafika
- dane vs zmienna losowa vs gęstość
- rozkłady, generowanie danych
- metoda największej wiarygodności
- rozkład normalny, estymacja parametrów
- kompresja
- estymacja kowariancji: ledoit-wolff
- fourier, spektrum, filtry liniowe, n-gramy
- metody minimalizacji, metody gradientowe
- entropia, kompresja Huffman, DUDA?
- pojęcie błędu, kwadratowy vs inne
- estymacja jądrowa, em - działanie
- modele gaussowskie - CEC
- walidacja krzyżowa, leave one out, bootstrap, k-fold
- hdbscan?
- PCA+ICA

Rozdział 2

Wstęp

2.1 Podział danych ze względu na pochodzenie

2.1.1 Człowiek

Obraz
- klasyczne RGB
- inny zakres, więcej typów fotoreceptorów (ptaki)
- hiperspektralne (przykłady zastosowań: szukanie surowców, rak, etc)
- filmy (obraz w czasie)

Dźwięk
- człowiek słyszy 20Hz-20kHz
- poza zakresem: nietoperz, wieloryb, trzęsienie ziemi

Tekst
- bardzo ważne z punktu widzenia zastosowań

2.1.2 Repozytorium UCI, MNIST, etc

Dane opisujące wyniki badań, eksperymentów, etc.

UCI jako baseline eksperymentów `http://mlr.cs.umass.edu/ml/`, MNIST: `http://yann.lecun.com/exdb/mnist/`.

2.2 Podział danych ze względu na strukturę

2.2.1 Macierze

Dźwięk, obraz.

2.2.2 Ciagi symboli

Tekst, DNA

2.2.3 Grafy

- związki chemiczne
- sieci społecznościowe

2.2.4 Co to znaczy, że rozumiemy dane?

ETAP 0. Znać i umieć zinterpretować wstępne charakterystyki danych: średnia, mediana, kowariancja, ilość grup, etc.

ETAP 1. Sklejać bardziej inteligentnie wykorzystując proste konstrukty. Nie wywołuje totalnego dysonansu.

ETAP 2. Przewidywać, uzupełniąć brakujące/zasłonięte fragmenty. Kompresować.

ETAP 3. Rozumieć kontekst. Przerzucać między sobą reprezentacje.

Obrazy - potrafimy robić okulary, człowieka

mężczyzna + korona = król

obraz Wojtka w stylu van Gogha

chemia - znamy aktywne, mamy jakas klasę zwiazakow chemicznych i chcemy znaleźć aktywnyw danej klasie (albo o danych własnościach chemicznych).

WYKŁAD będzie dotyczył głównie etapów 1-2, etap 3 to głównie Deep Learning.

2.3 Kluczowa trudność w analizie danych

2.3.1 Wektor losowy

Dane są próbką X stworzoną przez czarną skrzynkę $\mathbb{X}$ która je generuje. Precyzyjnie mówiąc ta czarna skrzynka nazywa się *wektor losowy* - jeżeli dane są w $\mathbb{R}^D$ i zmienna losowa - jeżeli dane są w $\mathbb{R}$. Mając dane $\mathbb{X}$ i zbiór[1] $A \subset \mathbb{R}^D$ mamy dobrze określone prawdopodobieństwo $P(\mathbb{X} \in A)$.

Nasza czarna skrzynka $\mathbb{X}$ ma na środku czerwony przycisk, który jeżeli go naciśniemy wypluwa nam zestaw danych o interesującej nas liczebności (tak naprawdę zakładamy jeszcze, że kolejne dane są od siebie niezależnes).

- załóżmy, że zdajemy egzamin z matematyki dyskretnej prowadzony przez prof. Dyskretnego od 10 lat. Wtedy jako próbkę z danych mamy 20 egzaminów (10 w I terminie i 10 w drugim). Natomiast czarną skrzynką która te dane (zadania egzaminacyjne) generuje jest mózg profesora Dyskretnego.

- mogą nas interesować dane z badania pacjentów (wiek, waga, morfologia, ect) podejrzanych na jakąś chorobę. Wtedy czarną skrzynką może być konsorcjum lekarskie, które jeżeli mu zapłacimy 100 pln, łapie jakiegoś pacjenta w szpitalu, przeprowadza żadane badania i wysyła nam ich wyniki.

Problem 2.1 (Główny problem)**.** *Nasza trudność polega na tym, że mamy do dyspozycji próbkę X, a interesuje nas działanie czarnej skrzynki (wektora losowego) $\mathbb{X}$. Dodatkowo często zdarza się, że nasza czarna skrzynka utworzyła pewien zestaw danych, i nie może już utworzyć więcej (bo dla przykładu wszyscy pacjenci w szpitali już zostali przebadani).*

Jak państwo widzą, nawet jeżeli dane są proste, jak tekst stanowiący zawartość egzaminu z matematyki dyskretnej, czarna skrzynka która je generuje może być bardzo skomplikowana (mózg profesora Dyskretnego).

[1] Formalnie rzecz biorąc musimy dodatkowo założyć, że A jest mierzalny.

2.3.2 Przeuczanie i walidacja krzyżowa (cross-validation)

Konsekwencją powyższej dychotomii – mamy dane X a interesuje nas metoda która je generuje $\mathbb{X}$ – jest powstanie efektu przeuczania (nadmiernego dofitowania do danych, overfitting) oraz walidacji krzyżowej (jako metody radzenia sobie z problemem).

Rozpatrzmy następujący przykład.

Przykład 2.1. Załóżmy, że chcemy się nauczyć matematyki dyskretnej do egzaminu. Bierzemy więc wszystkie egzaminy z ostatnich 10 lat, uczymy się ich rozwiązań. Załóżmy, że umiemy już wszystkie rozwiązać, więc wnioskujemy, że dostaniemy pięć. Przychodzi egzamin, i zaskoczeni dostajemy trzy. Pytanie dlaczego, skoro przecież przerobiliśmy wszystkie zadania z naszego zestawu danych?

Otóż kluczowe jest to, że naszym prawdziwym celem było nie przygotowanie do przeszłych egzaminów (czyli do próbki X), ale do następnego egzaminu który zostanie wyprodukowany przez czarną skrzynkę $\mathbb{X}$ (mózg profesora dyskretnego).

W związku z tym powstają dwa pytania:

1. jak sprawdzić jaki jest nasz prawdziwy stopień przygotowania, czyli oszacować jaką dostaniemy ocenę na egzaminie?

2. jak się przygotować do następnego egzaminu?

Na pytanie 2 będziemy się starali odpowiedzieć w trakcie wykładu, natomiast pytanie 1 ma bardzo proste rozwiązanie. Otóż przygotowując się do egzaminu powinniśmy odłożyć jeden z egzaminów (nazwijmy go testowy) i z niego nie przerabiać zadań. Dopiero po przygotowaniu się na podstawie pozostałych (tak zwanych treningowych) zadań, sprawdzamy swoją wiedzę rozwiązując egzamin testowy które wcześniej nie widzieliśmy.

To jak ten egzamin testowy rozwiązaliśmy powinno być dobrym oszacowaniem tego co dostaniemy przy prawdziwym egzaminie.

Wrócimy do tego dokładniej jeszcze w przyszłości, ale to podejście sformalizowane prowadzi do tak zwanej *walidacji krzyżowej* (cross-validation).

Rozdział 3

Opis danych

3.1 Histogram i gęstość

Powstaje teraz naturalne pytanie w jaki sposób charakteryzować nasze czarne skrzynki do produkcji danych. Załóżmy bowiem, że mamy dwie skrzynki $\mathbb{X}_1, \mathbb{X}_2$ które mogą nam produkować dane. Wtedy utożsamiamy je ze sobą pisząc $\mathbb{X}_1 = \mathbb{X}_2$ gdy z równym prawdopodobieństwem wylosują każdy podzbiór:

$$P(\mathbb{X}_1 \in A) = P(\mathbb{X}_2 \in A) \text{ dla dowolnego podzbioru } A.$$

Pomimo tego, że powyższa definicja wygląda ładnie, jest mało praktyczna w sprawdzaniu, gdyż wymaga sprawdzania warunku po podzbiorach. W związku z tym potrzebne są praktyczne warunki które pozwolą nam to sprawdzić.

Zacznijmy analizę tej sytuacji od danych dyskretnych w $\mathbb{R}$, to znaczy zakładamy, że możemy wylosować tylko N możliwych wyników $x_0, \ldots, x_{N-1} \in \mathbb{R}$. Losując teraz za pomocą naszej skrzynki coraz większą ilość danych, możemy zliczyć jak często wylosujemy każdy z elementów x_i. Asymptotycznie 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). Teraz łatwo zauważyć, że utożsamiamy ze sobą te zmienne losowe, które te rozkłady mają jednakowe.

Teraz zajmiemy się przypadkiem ciągłym, gdzie możemy potencjalnie wylosować dowolne liczby rzeczywiste z pewnego zakresu $[a, b)$. Zacznijmy najpierw od wprowadzania 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 zakres $[a, b)$ jaki przyjmuje nasza zmienna losowa na K rozłączne „pudełka" równej długości:

$$P_i = [a + \tfrac{i}{K}(b-a), a + \tfrac{i+1}{K}(b-a)) \text{ dla } i = 0, \ldots, K-1.$$

Mając zestaw danych $X = (x_i)$ możemy zliczyć ile procent danych wpadło danych do każdego pudełka, i w konsekwencji dostajemy histogram[1]:

$$h_i = \tfrac{1}{\mathrm{card} X} \mathrm{card}\{i : x_i \in P_i\}.$$

W przypadku gdy mamy zmienną losową $\mathbb{X}$, powyższy wzór to będzie po prostu

$$h_i = P(\mathbb{X} \in P_i)$$

[1] Często się zaznacza także ilość elementów które wpadły do danego pudełka, bądź gęstość prawdopodobieństwa – wtedy pole pod wykresem jest równe jeden.

gdzie $P(\mathbb{X} \in Q)$ oznacza prawdopodobieństwo, że nasza czarna skrzynka $\mathbb{X}$ wylosowała nam element należący do przedziału Q.

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 danego prostokąta (odpowiadającemu graficznie naszemu pudełku) jest taka, by jego pole było równe częstości 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[2] 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:

$$P(\mathbb{X} \in A) = \int_A f_{\mathbb{X}}(x)dx. \tag{3.1}$$

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 (3.1), że dwa wektory losowe są utożsamiane jeżeli mają równą gęstość.

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

3.2 Charakterystyki opisowe

3.2.1 Zmienne losowe

Zaczniemy od przypomnienia podstawowych charakterystyk dla danych. 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

[2]UWAGA: nie zawsze ta granica istnieje!

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

$$\text{mean}X = \frac{1}{|X|} \sum_i x_i.$$

Zadanie 3.2. *Proszę wyliczyć wzory na* $\text{mean}((x_i + y_i)_i)$ *oraz* $\text{mean}(CX)$.

Średnia z próbki to estymator[3] wartości oczekiwanej $E\mathbb{X}$, która w przypadku gdy $\mathbb{X}$ ma gęstość $f_{\mathbb{X}}$ wyraża się wzorem

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

Wprost z powyższej definicji mamy

$$E(\mathbb{X}_1 + \mathbb{X}_2) = E\mathbb{X}_1 + E\mathbb{X}_2.$$

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[4]. 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 wynik dzielący próbkę na dwie „połówki”[5]:

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

gdzie $X_{\leq m} = \{x \in X : x \leq m\}$. Proszę zauważyć, że względem powyższej definicji mediana jest przedziałem (w praktyce jedyność jest wtedy gdy zbiór danych ma nieparzystą ilość elementów, zaś w przypadku gdy zbiór ma parzystą liczbę elementów za medianę przyjmuje się dowolnego reprezentanta tego przedziału). Jak łatwo widać, mediana jest reprezentowana przez realna wartość ze zbioru danych, a co więcej jest relatywnie nieczuła na outliersy. Pokażę potem państwu, że oba te pojęcia są konsekwencją wyboru funkcji kosztu jaki rozpatrujemy przy zastępowaniu danych.

Zadanie 3.3. *Proszę policzyć medianę* $X = \{1, 2, 5, 2, 3\}$.

[3]To znaczy, jeżeli wielkość próbki rośnie do nieskończoności, to zmierza do szukanej wartości.

[4]Szpinak - cytowanie

[5]Proszę zauważyć, że połówki te nie muszą być równe.

W przypadku rozkładów które mają gęstość szukamy takiego m by:

$$\int_{(-\infty,m]} f_{\mathbb{X}}(x)dx = \int_{[m,\infty)} f_{\mathbb{X}}(x)dx.$$

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

$$\sigma(X)^2 = \mathrm{Var}X = \frac{1}{|X|}\sum_i (x_i - \mathrm{mean}X)^2.$$

Zauważmy, że

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

$$= \frac{1}{|X|}\sum_i x_i^2 - 2\mathrm{mean}X \cdot \mathrm{mean}X + \frac{1}{|X|}\sum_i (\mathrm{mean}X)^2 \frac{1}{|X|}\sum_i x_i^2 - (\mathrm{mean}X)^2.$$

Oznacza to, że

$$\mathrm{Var}X = \frac{1}{|X|}\sum_i x_i^2 - (\mathrm{mean}X)^2.$$

W przypadku zmiennych losowych mamy wzór:

$$\mathrm{Var}\mathbb{X} = E(\mathbb{X} - E\mathbb{X})^2 = E\mathbb{X}^2 - (E\mathbb{X}).$$

I znowu w przypadku gęstości wzór który dostajemy to

$$\mathrm{Var}\mathbb{X} = \int (x - E\mathbb{X})^2 f_{\mathbb{X}}(x)dx.$$

Przykład 3.1. 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:

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

Analogicznie do normalizacji względem rozkładu używa się normalizacji względem współczynników rozkładu:

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

Zadanie 3.4. *Proszę sprawdzić, że po dokonaniu tej procedury dostajemy rozkład o średniej zero i odchyleniu jeden.*

Zadanie 3.5. *Proszę dokonać normalizacji (średnia=0,odch=1) próbki* $X = \{1, 3, 7, 11\}$.

Sposób wizualizacji - barplot, zamieszcza oprócz zakresu $(\min, \max)$ średnią m oraz $m \pm \sigma$ (czasami się także znaczy kwantyle, czyli poziomy odpowiadające 25% i 75% danych).

Oprócz powyższych mierników rozważa się jeszcze momenty wyższych rzędów:

$E(\mathbb{X}^k) = \int x^k f_{\mathbb{X}}(x)dx$ dla rozkładów ciągłych, lub $E(\mathbb{X}^k) = \frac{1}{|X|}\sum_i x_i^k$ dla danych dyskretnych.

które jak zobaczymy potem mogą być zastosowane do wykrywania na ile dany rozkład różni się od rozkładu normalnego. Nie każdy rozkład ma momenty, czy nawet średnią[6], ale jeżeli dwa rozkłady mają wszystkie momenty i są one równe momenty, to są identyczne (to wynika z tego, że wielomiany są gęste w przestrzeni funkcji).

Zadanie 3.6. *Policz momenty*

3.2.2 Wektory losowe

Analogicznie jak w przypadku zmiennych losowych definiuje się wartość oczekiwaną dla wektora losowego w $\mathbb{R}^D$:

$$E\mathbb{X} = E(\mathbb{X}_1, \ldots, \mathbb{X}_D) = (E\mathbb{X}_1, \ldots, E\mathbb{X}_D).$$

W przypadku gdy rozkład ma gęstość lub jest dyskretny powyższy wzór redukuje się do

$$E\mathbb{X} = \sum p_i x_i \text{ lub } \int x f_{\mathbb{X}}(x) dx.$$

W przypadku próbki $X = (x_i)$ estymatorem wartości oczekiwanej jest po prostu wartość średnia

$$\text{mean}X = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Łatwo sprawdzić, że dla A liniowego mamy

$$E(A\mathbb{X} + b) = AE\mathbb{X} + Ab.$$

Zadanie 3.7. *Proszę sprawdzić powyższy wzór.*

Jak widzimy, wartość oczekiwana jest liczona osobno dla każdej współrzędnej. Powstaje więc oczywiste pytanie, w jaki sposób te zmienne są ze soba związane, na ile wartość jednej wpływa na inne.

Zajmiemy się najpierw najprostszym przypadkiem, gdy $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)$ jest wektorem losowym na płaszczyźnie. 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):

$$\mathbb{X}_2 \approx a_1 \mathbb{X}_1 + b_1 \text{ lub } \mathbb{X}_1 = a_2 \mathbb{X}_2 + b_2.$$

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

Dygresja 3.1. *Załóżmy, że mamy dwa wektory $v_1, v_2 \in \mathbb{R}^N$. Chcemy umieć sprawdzić, czy są one współliniowe, czyli czy istnieje α_1 bądź (równoważnie) α_2 takie*

$$v_1 = \alpha_2 v_2 \textit{ lub } v_2 = \alpha_1 v_1.$$

[6]Przykładem jest rozkład Cauchy'ego którego gęstość wyraża się wzorem $f(x) = \frac{1}{\pi(1+x^2)}$.

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:*

$$\cos(\angle v_1, v_2) = \frac{\langle v_1, v_2 \rangle}{\|v_1\| \cdot \|v_2\|}.$$

Tak zdefiniowany współczynnik jest powszechnie stosowanym miernikiem współliniowości. Używa go się między innymi w NLP do zdefiniowania podobieństwa między tekstami (jest nieczuły na długość tekstu).

Dla prostoty, współczynnik korelacji wyprowadzimy dla próbki $(X_1, X_2) = (x_1^i, x_2^i)_i$ wielkości N wygenerowanej przez wektor losowy $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)$. Bierzemy

$$x_1 = (x_1^1, \ldots, x_1^N) \text{ i } x_2 = (x_2^1, \ldots, x_2^N).$$

Chcemy sprawdzić, czy powyższe wektory są w zależności liniowej

$$x_2^i = a_1 x_1^i + b_1 \text{ bądź } x_1^i = a_2 x_2^i + b_2. \tag{3.2}$$

Trochę przeszkadza b, więc przesuwamy do zera (odejmując od obu średnią) rozpatrując wektory $v_1 = (v_1^i)$, $v_2 = (v_2^i)$ zdefiniowane przez

$$v_1^i = x_1^i - \mathrm{mean} X_1 \text{ oraz } v_2^i = x_2^i - \mathrm{mean} X_2,$$

łatwo wtedy sprawdzić, że (3.2) jest równoważne stwierdzeniu, że wektory v_1, v_2 są współliniowe:

$$v_2 = a_1 v_1 \text{ bądź } v_1 = a_2 v_2,$$

co na podstawie wcześniejszej dygresji sprowadza się do wyliczenia

$$\rho = \frac{\sum_i (x_1^i - \mathrm{mean} X_1)(x_2^i - \mathrm{mean} X_2)}{\sqrt{\sum_i (x_1^i - \mathrm{mean} X_1)^2} \sqrt{\sum_i (x_2^i - \mathrm{mean} X_2)^2}} = \frac{\mathrm{cov}(X_1, X_2)}{\sigma(X_1)\sigma(X_2)},$$

gdzie przez $\mathrm{cov}(X_1, X_2)$ oznaczamy uśredniony iloczyn skalarny pomiędzy $(x_2^i - \mathrm{mean} X_1)$ i $(x_2^i - \mathrm{mean} X_2)$:

$$\mathrm{cov}(X_1, X_2) = \frac{1}{N} \sum_i (x_1^i - \mathrm{mean} X_1)(x_2^i - \mathrm{mean} X_2).$$

W przypadku wektorów losowych powyższe wzory stają się

$$\rho = \frac{\mathrm{cov}(\mathbb{X}_1, \mathbb{X}_2)}{\sigma(\mathbb{X}_1) \cdot \sigma(\mathbb{X}_2)},$$

gdzie

$$\mathrm{cov}(\mathbb{X}_1, \mathbb{X}_2) = E(\mathbb{X}_1 - E\mathbb{X}_1)(\mathbb{X}_2 - E\mathbb{X}_2) = E(\mathbb{X}_1 \mathbb{X}_2) - E\mathbb{X}_1 E\mathbb{X}_2.$$

Przyjmuje się, że jeżeli współczynnik korelacji na moduł jest większy of 0.5, to jest korelacja liniowa.

W przypadku wektorów losowych częściej niż współczynnik korelacji liczy się macierz kowariancji, która jest zdefiniowana przez

$$\mathrm{cov}\mathbb{X} = [\mathrm{cov}(\mathbb{X}_i, \mathbb{X}_j)]_{i,j} = E((\mathbb{X} - E\mathbb{X}) \cdot (\mathbb{X} - E\mathbb{X})^T) = E(\mathbb{X}\mathbb{X}^T) - (E\mathbb{X})(E\mathbb{X})^T.$$

Proszę zauważyć, że mając macierz kowariancji można wyliczyć współczynniki korelacji. Co więcej, różne współrzędne są liniowo niezależne gdy macierz ta jest diagonalna. Gdy $\mathbb{X}$ ma gęstość $f_{\mathbb{X}}$, mamy

$$\mathrm{cov}\mathbb{X} = \int (x - E\mathbb{X})(x - E\mathbb{X})^T f_{\mathbb{X}}(x) dx.$$

Dla próbki X wzory to

$$\mathrm{cov}X = \frac{1}{N} \sum_i (x_i - \mathrm{mean}X)(x_i - \mathrm{mean}X)^T. \tag{3.3}$$

Zadanie 3.8. *Nasz zestaw danych składa się z punktów $\{(1,2),(2,3),(5,5)\}$ na płaszczyźnie. Proszę wyliczyć współczynnik korelacji między współrzędnymi X i Y. Proszę policzyć macierz kowariancji.*

Ponieważ Wprost z definicji widać, że macierz kowariancji jest symetryczna.

3.3 Whitening, odległość Mahalanobisa

Z punktu widzenia wielu metod nauczania maszynowego, najlepiej jeżeli dane są znormalizowane, czyli w naszym przypadku jak nie ma między nimi zależności liniowej, średnia jest zero a macierz kowariancji jest macierzą identycznościową.

Uzyskanie tego, by wartość oczekiwana była zero, jest łatwe - po prostu przesuwamy

$$Y = X - \mathrm{mean}X.$$

Natomiast powstaje oczywiście pytanie, w jaki sposób zmodyfikować próbkę, by współrzędne były liniowo niezależne. Interesuje nas w jaki sposób liniowo przekształcić dane, musimy więc wiedzieć jak się przekształca macierz kowariancji. Dla prostoty korzystając z (3.3) mamy

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

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 istotną klasę stanowią odwzorowania symetryczne, aby zagwarantować jednoznaczność możemy zawężamy się do szukania A w klasie odwzorowań symetrycznych[7]. I wtedy jak wiemy z algebry liniowej rozwiązanie jedyne rozwiązanie jest dane przez pierwiastek[8]

[7] Rozważamy ogólną sytuację w późniejszym rozdziale dotyczącym ICA

[8] Przypominam algorytm liczenia pierwiastka z dodatnio określonej macierzy symetrycznej A: a) Policz wartości własne $(\lambda_1, \ldots, \lambda_D)$ i wektory własne $V = [v_1, \ldots, v_D]$. Czyli w postaci macierzowej dla $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ dostajemy $AV = V\Lambda$ (wybieramy wektory własne ortonormalne, czyli $V^T V = I$). b) Końcowy wzór:

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

Konkludując, końcowe rozwiązanie dane jest

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

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:

$$\|x\|_\Sigma^2 = x^T \Sigma^{-1} x.$$

Wtedy mamy

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

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).

Faktem który pokazuje wagę metryki mahalanobisa, jest to, że jest ona niezależna na transformacje afiniczne zbioru.

Obserwacja 3.1. *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 $\phi(x) = Ax + b$.

Dowód. Weźmy dowolne $x_i, x_j \in X$ i rozpatrzmy odwzorowanie afiniczne $\phi : x \to Ax + b$ (gdzie A jest odwracalne, a b translacja). Wtedy mamy

$$\|x_i - x_j\|_{\mathrm{cov}X}^2 = (x_i - x_j)^T \mathrm{cov}X^{-1}(x_i - x_j),$$

a

$$\|\phi(x_i) - \phi(x_j)\|_{\mathrm{cov}\phi(X)}^2 = \|Ax_i - Ax_j\|_{A\mathrm{cov}XA^T} = (A(x_i - x_j))^T (A\mathrm{cov}XA^T)^{-1}(A(x_i - x_j))$$

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

□

3.4 Dane z wagami

W praktyce w wielu przypadkach spotkamy się z sytuacją, gdy będziemy chcieli mieć dane z dodatkową wartością (częstość, waga). Czyli jeżeli $X = (x_i)$, a $w = (w_i) \subset \mathbb{R}_+$, to zapisujemy wtedy X z wagami jako $X_w = (x_i, w_i)$.

Zazwyczaj zakładamy, że wagi są (wspólnie) ograniczone. Wszystkie wzory dotyczące średnich, etc się przekładają, dla przykładu

$$EX_w = \frac{1}{\sum_i w_i} \sum_i w_i x_i \text{ czy } \mathrm{cov}X = \frac{1}{\sum_i w_i} \sum_i w_i (x_i - EX_w)(x_i - EX_w)^T.$$

Analogicznie możemy rozpatrywać sytuację dla wektorów losowych, z tym, że tutaj formalna definicja jest bardziej skomplikowana, w związku z czym ją pomijam. Jedynie warto wspomnieć, że jeżeli wektor losowy $\mathbb{X}$ w $\mathbb{R}^D$ ma gęstość f, i $w : \mathbb{R}^D \to \mathbb{R}_+$ to ograniczona z góry funkcja wagująca, to zmienna losowa $\mathbb{X}_w$ powstała przez zwagowanie $\mathbb{X}$ ma gęstość

$$f_w(x) = \frac{1}{\int w(z) f(z) dz} w(x) f(x).$$

Rozdział 4

Rozkłady danych

4.1 Rozkłady bazowe

4.1.1 Rozkład dyskretny na $\{0, \ldots, M-1\}$

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 4.1. 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 4.2. Zły dobór parametrów może mieć tragiczne skutki – niesławny tu jest RANDU zaprojektowany w latach 60tych przez IBM-a:

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

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`.

4.1.2 $\mathrm{unif}_{[a,b]}$: rozkład równomierny na odcinku $[a, b]$

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 $[0, 1]$.

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 $\mathbb{X}$.

4.1.3 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 4.1. *Rozkład wykładniczy ma gęstość daną wzorem:*

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

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 4.1. *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*

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

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

4.1.4 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 4.1. 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.

4.1.5 Wagowanie

Cel: rownoważenie klas. Jaka gęstość.

4.1.6 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).

Rozdział 5

Kompresja stratna

5.1 Wektoryzacja – kompresja grupy danych

Kompresja stratna i bezstratna. Dane zajmowałyby za dużo pamięci.

Problem 5.1. *Zastąpić grupę punktów/danych $X = \{x_i\}_{i=1..k} \subset \mathbb{R}^d$ za pomocą jednego, tak by był minimalny błąd.*

Dwa pytania:

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

Błąd oczywiście musi być funkcją nieujemną.

5.1.1 Błąd (średnio-)kwadratowy w $\mathbb{R}$

Dygresja 5.1. *Funkcja ax^2+bx+c, gdzie $a > 0$, osiąga minimum w punkcie $-b/(2a)$. Wartość minimalna wynosi $-\Delta/(4a)$.*

Jednym z najczęściej stosowanych sposobów pomiaru błędu jest 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

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

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

$$v = E(X) = \tfrac{1}{k}\sum_i x_i,$$

zaś wartość tego minimalnego błędu jest równa

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

gdzie pierwszą równość jedno dostajemy bezpośrednio wstawiając, a drugie z wzoru, że minimum funkcji kwadratowej wynosi $-\Delta/4a$. Jeżeli weźmiemy pod uwagę wartość tego błędu uśrednionego (t.j. podzielonego przez ilość elementów k zbioru X) to dostajemy definicję tak zwanej *wariancji*:

$$\mathrm{Var}(X) = \frac{1}{k}\sum_i |x_i - E(X)|^2 = E(X^2) - E(X)^2 = \tfrac{1}{k}\sum_i x_i^2 - (\tfrac{1}{k}\sum_i x_i)^2. \qquad (5.1)$$

Często piszemy EX zamiast $E(X)$ o ile nie powoduje to nieporozumień. Przez σ oznaczamy *odchylenie standardowe* dane wzorem $\sigma^2 = \mathrm{Var}(X)$ (pierwiastek z wariancji). Z (5.1) mamy

$$\sigma_X^2 = \mathrm{Var}(X) = E((X - EX)^2) = EX^2 - (EX)^2.$$

Uwaga 5.1. Dobre: szybko się liczy (liniowo względem ilości danych), naturalna motywacja.

Złe: otrzymujemy często wynik który nie istnieje, duża czułość na wartości oddalone (outliers=zaburzenia/b
w danych).

5.1.2 Błąd dany przez moduł

Zobaczymy teraz co by się stało gdybyśmy rozważali błąd innego typu.

Załóżmy, że interesuje nas błąd dany przez:

$$\sum_i |x_i - \bar{x}|.$$

Lemat 5.1. *Załóżmy dodatkowo, że X jest posortowany, to znaczy $x_1 \le x_2 \le \cdots \le x_{k-1} \le x_k$. Rozpatrzmy funkcję*

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

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

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

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

co daje tezę. □

Wniosek 5.1. *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 $[x_{(k+1)/2}, \infty)$.*

Definicja 5.1. W ten sposób otrzymujemy definicję mediany – dowolny punkt taki, że ilość punktów ze zbioru silnie mniejszych jest równa ilości silnie większych. Jak nieparzysta ilość danych, to jednoznacznie zdefiniowana, jeżeli parzysta, to przedział. Jak widzimy ważne jest by dane wstępnie posortować.

Uwaga 5.2. Dobre: w miarę szybko się liczy ($n \log n$ względem ilości danych), naturalna motywacja, zawsze otrzymujemy realną daną, mniejszy wpływ outliersów.

Złe: nie ma wzoru dla sytuacji wyżej wymiarowej (nawet na płaszczyźnie)!

Zadanie 5.1. *Policzyć średnią i medianę dla liczb, zaburzyć jedną – zobaczyć jaki jest wpływ wartości oddalonych (outliersów).*

5.1.3 Sytuacja wyżej wymiarowa

Pokażemy, że jest pełna analogia z przypadkiem jednowymiarowym. Zastąpienie grupy danych $X = \{x_i\} \subset \mathbb{R}^N$ przy pomocy jednej v - kompresja minimalizowanie wartości:

$$v \to \sum_i \|x_i - v\|^2 \tag{5.2}$$

Twierdzenie 5.1. *Rozpatrzmy funkcję*

$$g : x \to a\langle x, x\rangle + \langle b, x\rangle + c.$$

Wtedy

$$g(x) = a\langle x + b/(2a), x + b/(2a)\rangle - (\langle b, b\rangle - 4ac)/(4a) = a\|x + b/(2a)\|^2 - \Delta/(4a).$$

Oznacza, że jeżeli $a > 0$ to minimum jest przyjmowane dla $x = -b/(2a)$ i wynosi $-\Delta/(4a)$ (bo $\|x + b/(2a)\|^2$ jest minimalizowane dla $x = -b/(2a)$).

Dowód. Taki sam jak dla jednowymiarowego przypadku (sprowadzanie do postaci kanonicznej). □

Wzór który wylicza (5.2) do jednej postaci – kluczowa jest obserwacja, że suma funkcji kwadratowych jest funkcją kwadratową!

Stwierdzenie 5.1. *Mamy wzór:*

$$\frac{1}{k}\sum_i \|x_i - v\|^2 = \|v\|^2 - 2\langle E(X), v\rangle + \frac{1}{k}\sum_i \|x_i\|^2 = \|v - E(X)\|^2 + (\tfrac{1}{k}\sum_i \|x_i - E(X)\|^2).$$

Widzimy kiedy się ten obiekt minimalizuje! W konsekwencji, podobnie jak w przypadku jednowymiarowym, dla zbioru danych $X = \{x_i\}_{i=1..k}$ w sposób naturalny zdefiniowaliśmy średnią:

$$E(X) = \tfrac{1}{k}\sum_i x_i,$$

oraz uogólniony odpowiednik wariancji (dla wygody używam tego samego oznaczenia):

$$\mathrm{Var}(X) = \tfrac{1}{k}\sum_i \|x_i - E(X)\|^2.$$

Ćwiczenie 5.1. Proszę pokazać, że

$$\mathrm{Var}(X) = \tfrac{1}{2k^2}\sum_{i,j} \|x_i - x_j\|^2.$$

Ćwiczenie 5.2 (do metody Hartigana). Korzystając z poprzedniego ćwiczenia, proszę pokazać, że przy rozbiciu X na dwa rozłączne podzbiory X_1, X_2 mamy

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \tfrac{k_1 k_2}{k^2}(EX_1 - EX_2)^2.$$

5.2 Podstawy minimalizacji funkcji

W przypadku błędy danego przez sumę modułów (norm) w przypadku wielowymiarowym, nie istnieje już jawny wzór na rozwiązanie. W związku z czym stosuje się metody minimalizacji.

Problem: szukanie minimum funkcji $f : \mathbb{R}^D \to \mathbb{R}$. Najczęściej spotykane są dwa podejścia:

- metody gradientowe, bądź bazujące na gradientowych,
- metody zastępujące funkcję inną, której minimum znamy, czy potrafimy oszacować.

Metody gradientowe są metodami ciągłymi, ich idea polega na tym, że chcemy „spadać w dół" po wykresie funkcji (można sobie wyobrazić, że wykres funkcji to są góry, z których chcemy jak najszybciej zejść). Startujemy zazwyczaj od losowo wybranego punktu x_0 (w związku z tym zazwyczaj losujemy ten punkt wielokrotnie, chyba, że mamy gwarancję, że minimum jest globalne). Stosuje się je między innymi do optymalizacji sieci neuronowych.

Metody polegające na zastępowaniu, polegają na tym, że zastępujemy rozważaną funkcją inną, której minimum potrafimy znaleźć, a która lokalnie w otoczeniu punktu x_0 zachowuje się jak rozważana przez nas (albo przynajmniej ją ogranicza od góry). Pokażę państwu przykład tej metody na podstawie IRLS.

5.2.1 Metody gradientowe

Poniżej naszkicuję podstawową metodę gradientową (drugie podejście będzie dokładnie opisane w następnej sekcji). Przez gradient funkcji $f : \mathbb{R}^D \to \mathbb{R}$ nazywamy kierunek największego wzrostu, to jest

$$\nabla f = [\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_D}]^T \in \mathbb{R}^D.$$

Gradient zasadniczo zachowuje się tak samo jak pochodna, ale w przeciwieństwie do pochodnej jest elementem przestrzeni (natomiast do określenia gradientu musi być w tej przestrzeni ustalony iloczyn skalarany). Jeżeli funkcja w danym punkcie x_0 ma ekstremum, to $\nabla f(x_0) = 0$ (w szczególności często szukanie ekstremów sprowadza się do liczenia gradientu).

Przytoczę dla wygody podstawowe własności gradientu:

- jeżeli $f(x) = a + \langle w, x\rangle$, to oczywiście $\nabla f(x) = w$,
- jeżeli $f(x_0 + h) = f(x_0) + \langle w, h\rangle + o(h)$ to $\nabla f(x_0) = w$,
- gradient sumy/różnicy, to suma/różnica gradientów,
- dla funkcji skalarnych f, g: $\nabla(f(x)g(x)) = \nabla f(x) \cdot g(x) + f(x)\nabla g(x)$ (analogicznie dla ilorazu),
- $\nabla F(g(x)) = F'(g(x)) \cdot \nabla g(x)$.

Przykład 5.1. Policzmy dla przykładu gradient funkcji $\phi : x \to x^T Ax = \langle x, Ax\rangle$:

$$\begin{aligned}\phi(x + h) &= \langle x + h, Ax + Ah\rangle = \langle x, Ax\rangle + \langle x, Ah\rangle + \langle h, Ax\rangle + o(h)\\ &= \phi(x) + \langle (A + A^T)x, h\rangle + o(h),\end{aligned}$$

co oznacza, że

$$\nabla\phi(x) = Ax + A^T x.$$

Jako wniosek (korzystamy z ostatniego punktu dla $g(x) = x - w$) dostajemy

$$\nabla\langle x - w, A(x - w)\rangle = A(x - w) + A^T(x - w).$$

Uwaga 5.3. Jeżeli chcemy teraz znaleźć minimum danej funkcji kosztu, to musimy schodzić w „dół", czyli poruszać się przeciwnie do gradientu:

$$x_{k+1} = x_k - h\nabla f(x_k),$$

gdzie h jest małe. Przerywamy, albo gdy wzrost, albo gdy spadek jest bardzo mały.

Docelowo znajdziemy punkty x takie, że $\nabla f(x) = 0$.

5.2.2 Metoda IRLS=(iterated reweighted least squares)

Wracamy do szukania najlepszego przybliżenia, czyli szukania wielowymiarowej mediany.

Dany zbiór $(x_i)_{i=1..N} \subset \mathbb{R}^d$ który chcemy zareprezentować przez jeden m, aby zminimalizować sumaryczny błąd. W naszym przypadku chcemy zminimalizować

$$L : m \to \sum_i f(\|x_i - m\|^2),$$

gdzie f ustalona funkcja.

Proszę zauważyć, że jeżeli $w \to f(\|w\|^2)$ jest wypukłe (tak będzie w naszym przypadku), mamy globalne minimum.

Przypadek $f(r) = r$ już omawialiśmy, prowadzi do klasycznej metody najmniejszych kwadratów, gdzie

$$m = \frac{1}{N}\sum_i x_i.$$

Analogicznie z wagami w_i. Wtedy minimum funkcji

$$m \to \sum_i w_i \|x_i - m\|^2$$

jest przyjmowane w **ważonej średniej**

$$\frac{1}{\sum_i w_i} w_i x_i.$$

Rozważaliśmy jednowymiarowy przypadek: mediana: $f(r) = \sqrt{r}$. Pytanie: jak znaleźć jednowymiarową medianę w przypadku wielowymiarowym, czyli minimum $m \to \sum_i \|x_i - m\|$, a w ogólności

$$L : m \to \sum_i f(\|x_i - m\|^2).$$

IRLS=**Iterated reweighted least squares**

Pracuje gdy f jest funkcją wklęsłą (co zachodzi w przypadku mediany, gdzie $f(r) = \sqrt{r}$). Załóżmy, że mamy znalezione w k-tym kroku przybliżenie m_k minimum. Chcemy poprawić m_{k+1} zamiast m_k, by zachodziło

$$L(m_{k+1}) \leq L(m_k).$$

Ponieważ f jest wklęsłe, w każdym punkcie

$$r_i := \|x_i - m_k\|^2$$

dziedziny możemy oszacować f z góry przez funkcję podpierającą w r_i daną wzorem

$$f(r) \leq f(r_i) + f'(r_i)(r - r_i) = f'(r_i) \cdot r + (f(r_i) - r_i f'(r_i)) = w_i^k r + (f(r_i) - r_i f'(r_i)),$$

gdzie waga

$$w_i^k := f'(r_i) = f'(\|x_i - m_k\|^2).$$

Zauważmy, że RHS jest funkcją liniową która zgadza się z f w r_i, i ogranicza z góry, co oznacza, że

$$\sum_i f(\|x_i - m\|^2) \leq \sum_i \left[f'(r_i) \cdot \|x_i - m\|^2 + (f(r_i) - r_i f'(r_i))\right]$$

$$= \sum_i w_i^k \cdot \|x_i - m\|^2 + \text{const.}$$

W konsekwencji w następnej iteracji dostajemy

$$m_{k+1} := \frac{1}{\sum_i w_i^k} \sum_i w_i^k x_i.$$

Uwaga 5.4. Stosujemy metodę najmniejszych kwadratów, ale ze zmieniającymi się wagami, które zależą od odległości od poprzedniego.

Wróćmy do przypadku mediany w $\mathbb{R}^d$. Jako m_0 bierzemy średnią, albo losowo wybrany element z danych. Mając dane m_k, powtarzamy procedurę

- wylicz nowe wagi punktów:
$$w_i^k = \frac{1}{\|x_i - m_k\|}$$
- wylicz nową średnią x_i ze względu na wagi w_i^k:

$$m_{k+1} = \frac{1}{\sum_i w_i^k} \sum_i w_i^k x_i = m_k + \frac{1}{n} \sum_{i=1}^n \frac{x_i - m_k}{\|x_i - m_k\|},$$

dopóty ciąg m_k się nie ustabilizuje, to znaczy $m_{k+1} \approx m_k$.

Uwaga 5.5. Zobaczmy, że

$$m_{k+1} = m_k + \frac{1}{\sum_i w_i^k} \sum_i \frac{x_i - m_k}{\|x_i - m_k\|}.$$

To oznacza, że nie ma znaczenia jak daleko jest dany punkt od m_k, ważny jest tylko kierunek w którym on leży – jak widzimy, w powyższym wzorze uśredniamy kierunki do wszystkich punktów ze zbioru, a wielkość kroku zależy od tego jak daleko m_k leży od tych elementów zbioru.

5.3 Klastrowanie k-means

k-means jest tak naprawdę metodą kompresji/wektoryzacji.

5.3.1 Zafiksowane centra $v_1, \ldots, v_k$

Problem 5.2. *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:

$$j(x) = \operatorname*{argmin}_{l=1..k} \|x - v_i\|.$$

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

Problem 5.3. *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

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

5.3.3 Ogólny problem

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

Problem 5.4. *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 5.2. *Napisać k-means korzystający z metody Lloyda.*

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

5.3.4 Diagram Voronoi

Chcemy patrzeć, gdzie wpadnie nowy punkt – podział przestrzeni.

Oczywiście tym które minimalizuje odległość od x, czyli kładziemy

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

Czyli inaczej mówiąc, przybliżamy x najbliższym elementem ze zbioru V. 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:

Następne ważne twierdzenie, będzie dotyczyło obserwacji które punkty są bliżej wybranym punktom.

Obserwacja 5.1. *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`

5.4 Podejście Hartigana

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 5.2. *Własność:*

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

wynika z bezpośredniego rozpisania wzorów:

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

$$= \sum_i \langle x_i, x_i\rangle - 2|X|\langle m_x, v\rangle + |X|\langle v, v\rangle,$$

a podstawiając w powyższym $v = m_X$ dostajemy

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

Po odjęciu otrzymujemy to co chcieliśmy.

Obserwacja 5.3.

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

Dowód. Dowód wynika bezpośrednio z poprzedniej obserwacji 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 5.4. *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 5.2. *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

$$\frac{|X_j|}{|X_j|+1}\|x - m_j\|^2 < \frac{|X_i|}{|X_i|-1}\|x - m_i\|^2. \tag{5.3}$$

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

Algorytm Hartigana dla k-means.
Na wejściu:

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

Wyznaczamy średnie tych klastrów i ich liczności: $s_i = 0 \in \mathbb{R}^N$, $N_i = 0 \in \mathbb{N}$ dla $i = 1..k$. For $l = 1..n$ do $s_{\sigma(l)} \mathrel{+}= x_l$, $N_{\sigma(l)} \mathrel{++}$. Po przejściu kładziemy

$$m_i = s_i / N_i \text{ dla } i = 1..l.$$

Z każdym klastrem wiążemy jego średnią oraz liczność, chodzimy kolejno (wielokrotnie) po wszystkich punktach zbioru aż do uzyskania stabilizacji, zmieniamy przynależność gdy zajdzie (5.3), i wtedy updatujemy odpowiednio średnie i liczności klastrów.

5.5 PCA

Niech $X = (x_i)$ będzie podzbiorem $\mathbb{R}^D$. Chcemy znaleźć podprzestrzeń (afiniczną bądź liniową) danego wymiaru $d < D$ która najlepiej przybliża nam nasze dane.

Idea problemu pochodzi od kompresji – tak jak w k-means zastępowaliśmy dane przez k-punktów, tak tutaj zmniejszamy ilość parametrów potrzebnych do opisu danych – stopień kompresji r (compression rate) będzie wynosił D/d.

Dygresja 5.2. *Załózmy, że ktoś nam podał tą przestrzeń V, i chcemy dokonać kompresji punktu x. W praktyce to co robimy, to zastępujemy punkt x przez punkt $p_V x$, który oznacza punkt najbliższy do x leżący w V. Wtedy kwadratowy błąd to*

$$d^2(x, V) = \|x - p_V x\|^2 = \inf\{\|x - y\|^2 : y \in V\}.$$

Na szczęście jawny wzór na $p_V x$ możemy uzyskać przez rzutowanie ortogonalne. Jak mamy podprzestrzeń liniową V o bazie ortonormalnej $[v_1, \ldots, v_d]$, to rzut ortogonalny jest dany przez

$$p_V x = \sum_i \langle v_i, x\rangle v_i.$$

Jeżeli afiniczna, i $\bar{x} \in V$ i $[v_1, \ldots, v_d]$ to baza ortonormalna $V - \bar{x}$, to

$$x \to \bar{x} + \sum_i \langle v, x_i - \bar{x}\rangle v_i.$$

Błąd tak zdefiniowanej kompresji to $\|x - p_V x\|$.

Aby określić co rozumiemy przez „najlepiej przybliża" potrzebujemy definicji: przez *kwadratowy błąd* (squarred error) popełniony przy zastąpieniu X przez optymalne elementy z popdrzestrzeni $V \subset \mathbb{R}^D$ rozumiem

$$d^2(X; V) := \sum_i \|x_i - p_V x_i\|^2.$$

Załóżmy, że mamy ustaloną podprzestrzeń liniową V która nas interesuje do tworzenia kompresji. Pierwsze pytanie którym się zajmiemy, to jakiej translacji v należałoby dokonać na V, aby zminimalizować błąd kompresji przy przybliżeniu X za pomocą $v + V$ (wtedy pytamy się o podprzestrzeń afiniczną). Jak zobaczymy, optymalny wynik jest wtedy gdy przesuniemy naszą podprzestrzeń V tak by przechodziła przez środek X:

$$v = \mathrm{mean} X.$$

Obserwacja 5.5. *Niech będzie dana podprzestrzeń wektorowa V przestrzeni $\mathbb{R}^D$. Wtedy*

$$d^2(X, v + V) = d^2(p_{V^\perp}(X), p_{V^\perp}(v)) = \sum_i \|p_{V^\perp}(x_i) - p_{V^\perp}(v)\|^2. \tag{5.4}$$

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

Dowód. Oczywiście wystarczy pokazać to tezę dla pojedynczego $x \in \mathbb{R}^N$. Czyli, że

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

ale to wynika z informacji dotyczących przestrzeni Hilberta.

Skoro tak, to chcemy znaleźć punkt w przestrzeni $V^\perp$ który najlepiej przybliża $p_{V^\perp}(\mathrm{x})$, ale na podstawie obserwacji zrobionych dla k-means, to jest środek ciężkości, czyli $Ep_{V^\perp}(\mathrm{x}) = p_{V^\perp}(E\mathrm{x})$, z liniowości. □

Zadanie 5.3. *Proszę znaleźć* $d^2(\mathrm{x}; V)$ *dla* $\mathrm{x} = [1, 2, 3; 3, 2, 1]$, $V = \{[x_1, x_2]^T \subset \mathbb{R}^2 : x_1 + x_2 = 0\}$.

Zazwyczaj w związku z powyższym przesuwamy (ustawiamy) środek układu współrzędnych w środku ciężkości (czyli po przesunięciu mamy $E\mathrm{x} = 0$). I teraz pytamy się o to w jaki sposób dobrać podprzestrzeń liniową, by była najlepsza dokładność aproksymacji?

5.6 Dowód minimalizacji

Aby to rozstrzygnąć zaczniemy od policzenia odległości od podprzestrzeni: mamy daną bazę $v^1, \ldots, v^k$ przestrzeni V. Interesuje nas wartość

$$d^2(\mathrm{x}; V) := \sum_{i=1}^{K} \|\mathrm{x}^i - P_V \mathrm{x}^i\|^2 = \sum_{i=1}^{K} \|\mathrm{x}^i\|^2 - \|P_V \mathrm{x}^i\|^2.$$

Czyli aby zminimalizować $d^2(\mathrm{x}; V)$ wystarczy zmaksymalizować $\sum_{i=1}^{K} \|P_V \mathrm{x}^i\|^2$. W tym celu przyda się nam następujący wzór:

Lemat 5.2. *Niech* V *podprzestrzeń o bazie ortonormalnej* $v = [v^1, \ldots, v^k]$. *Wtedy*

$$\sum_{i=1}^{K} \|P_V \mathrm{x}_i\|^2 = \mathrm{tr}(v^T \Sigma_{\mathrm{x}} v),$$

gdzie $\Sigma_{\mathrm{x}} = \mathrm{x}\mathrm{x}^T$.

Dowód. Ponieważ dla $x \in X$ mamy

$$\|P_V x\|^2 = \sum_{i=1}^{k} \langle x, v_i \rangle^2 = \sum_i (v_i' x) \cdot (x' v_i) =$$

$$\sum_i v_i'(xx')v_i = \mathrm{tr}(v' x x' v).$$

Ponieważ $\mathrm{x}\mathrm{x}' = \sum_i \mathrm{x}^i \mathrm{x}^{iT}$ dostajemy tezę lematu. □

Uwaga 5.6. Zauważmy, że jeżeli dokonaliśmy przesunięcia środka układu do środka ciężkości danych (dane traktujemy jako zbiór punktów, a nie jako ciąg), to wtedy $E\mathrm{x} = 0$, czyli

$$\frac{1}{K}\Sigma_{\mathrm{x}} = E(\mathrm{x}\mathrm{x}') - E(\mathrm{x})E(\mathrm{x}') = \mathrm{cov}(\mathrm{x}),$$

gdzie $\mathrm{cov}(\mathrm{x})$ to macierz kowariancji danych x o wyrazach $\mathrm{cov}(\mathrm{x}_i, \mathrm{x}_j)$ (gdzie przypominam, że indeks u dołu to branie odpowiedniej współrzędnej).

Uwaga 5.7 (interpretacja geometryczna Σ_{x}). Załóżmy, że chcemy wyznaczyć kierunek najbardziej reprezentatywny dla zestawu danych.

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.

Ponieważ jak łatwo zauważyć macierz Σ_{x} jest macierzą symetryczną, przydadzą nam się podstawowe informacje na temat macierzy symetrycznych:

- macierz symetryczna ma rzeczywiste wartości własne;
- można znaleźć bazę ortonormalną składającą się z wektorów własnych;
- jeżeli macierz symetryczna A jest nieujemnie określona, to znaczy $y \to y'Ay$ jest nieujemna, to wartości własne A są nieujemne.

Teraz jesteśmy już w stanie sformułować główne twierdzenie obecnej sekcji.

Twierdzenie 5.2. *Rozpatrzmy wszystkie K-wymiarowe podprzestrzenie V o bazie ortonormalnej v. Wtedy wartość*

$$\mathrm{tr}(v' \Sigma_{\mathrm{x}} v)$$

jest maksymalna, gdy v to pierwsze K-elementów bazy ortonormalnej składającej się z wektorów własnych macierzy Σ_{x} ustawionych malejąco po wartościach własnych.

Wnioski:

- dostajemy transformatę PCA (Karhunen-Loeve): najpierw zmiana środka ciężkości, następnie wybór bazy ortonormalnej dla macierzy Σ_{x} (w scilabie patrz komenda `spec`, ustawia o ile pamiętam po wartościach własnych rosnąco, czyli jeżeli chcemy wziąć Karhunen-Loeve Transform, to musimy brać ostatnie K wektorów z bazy);
- proszę zauważyć, że ponieważ nasza baza diagonalizuje Σ_{x}, to po transformacie współrzędne przestają być skorelowane.

Dowód za PCA Jolliffe:
A macierz składająca się z orotonormalnych wektorów własnych, $A'\Sigma A = \Lambda$.
wspolrzedne w tej bazie ortonormalnej to A'

Lemat 5.3. *Dla każdej liczby całkowitej $q, 1 \leq q \leq p$, rozważmy ortonormalną transformację liniową*

$$y = B'x$$

where y is a q-element vector, and B' is a $(q \times p)$ matrix, and let $\Sigma_y = B'\Sigma B$ be the variance-covariance matrix for y. Then the trace of Σ_y, denoted by $\mathrm{tr}(\Sigma_y)$, is maximized by taking $B = A_q$, where A_q consistis of the first q columns of A.

Dowód. Let β_k be the k-th column of B, as the columns of A form a basis for p-dimensional space, we have

$$\beta_k = \sum_{j=1}^{P} c_{jk}\alpha_j, k = 1, 2, \ldots, q,$$

where $c_{jk}, j = 1, 2, \ldots, p\ k = 1, 2, \ldots, q$ are appropriately defined constants. Thus $B = AC$, where C is the $(p \times q)$ matrix with $(j, k)$the element c_{jk} and

$$B'\Sigma B = C'A'\Sigma AC = C'\Lambda C = \sum_{j=1}^{p} \lambda_j c_j c_j',$$

where c_j' is the jth row of C. Therefore

$$\mathrm{tr}(B'\Sigma B) = \sum_j \lambda_j \mathrm{tr} c_j c_j' = \sum_j \lambda_j \mathrm{tr} c_j' c_j =$$

$$\sum_{j,k} \lambda_j c_{jk}^2. \tag{5.5}$$

Now

$$C = A'B \text{ so } C'C = B'AA'B = B'B = I_q,$$

because A is orthogonal and the columns of B are orthonormal. Hence

$$\sum_{j,k} c_{jk}^2 = q \tag{5.6}$$

and the columns of C are also orthonormal. The matrix C can be thought of the first a columns of a $(p \times p)$ orthogonal matrix D, say. But the rows of D are orthonormal and so satisfy $d_j'd_j = 1$, $j = 1, \ldots, p$. As the rows of C consist of the first q elements of the rows of D, it follows that $c_j'c_j \leq 1\ j = 1, \ldots, p$, that is

$$\sum_{k=1}^{q} c_{jk}^2 \leq 1. \tag{5.7}$$

Now $\sum_{k=1}^{q} c_{jk}^2$ is the coefficient of λ_j in (5.5), the sum of these coefficients is q from (5.6) and none of the coefficients can exceed 1, from (5.7). Because $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, it is fairly clear that

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

will be maximized if we can find a set of c_{jk} for which

$$\sum_{k=1}^{q} c_{jk}^2 = \begin{cases} 1, j = 1..q \\ 0, j = q+1..p \end{cases} \tag{5.8}$$

But if $B' = A_q'$ then

$$c_{jk} = \begin{cases} 1, j = 1 \leq j = k \leq q \\ 0, \text{ elsewhere.} \end{cases}$$

which satisfies (5.8). Thus $\mathrm{tr}(\Sigma_y)$ achievies ist maximum value when $B' = A_q'$. □

Rozdział 6

Kompresja bezstratna

6.1 Entropia

6.1.1 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. □

6.1.2 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 $\log_2 n$.

6.1.3 Kodowanie, entropia i tw. McMillana

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. □

6.1.4 Entropia krzyżowa i 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.

6.2 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 zmiennej losowej X rozważamy jej dyskretyzację, to jest

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

Twierdzenie 6.3. *Niech X zmienna losowa 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

7.6.1 Wzór na log-likelihood

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

$$\log \mathrm{L}(X, \mathcal{N}(m, \Sigma)) = \sum_i \ln \frac{1}{\sqrt{2\pi|\Sigma|}} \exp(-\tfrac{1}{2}\|x_i - m\|^2_\Sigma)$$

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

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

Oznaczmy (przy ustalonym Σ)

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

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.

7.6.2 Gradient względem macierzy

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*

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

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

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

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

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

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

□

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)*

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

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

$$u^T(\Sigma + H)^{-1} v - u^T \Sigma^{-1} v = -u^T \Sigma^{-1} H \Sigma^{-1} v + o(H)$$

$$= -\mathrm{tr}(u^T \Sigma^{-1} H \Sigma^{-1} v) + o(H) = -\mathrm{tr}(\Sigma^{-1} v u^T \Sigma^{-1} H) + o(H)$$

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

□

Możemy teraz przejść do głównego wyniku.

Twierdzenie 7.1. *Niech*

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

Wtedy

$$\nabla \Phi(\Sigma) =$$

7.6.3 Rozkład osobliwy – regularyzacja

ε - maszynowe

Ledoit-Wolff - wzór i motywacja.

Jako druga opcja - cross-validacja

Rozdział 8

Estymacja gęstości

8.1 Parametryczna

kNN?

8.2 Nieparametryczna

8.3 GMM

różne modele gaussowskie najpierw

Rozdział 9

Klastrowanie

9.1 Typy - hierarchiczne, soft, probabilistyczne, fuzzy

9.2 Gęstościowe - dbscan

9.3 k-means i wariacje: CEC

9.4 Spektralne

9.5 Semi-supervised

Rozdział 10

ICA

10.1 Motywacja

10.2 Optymalizacja