Wprowadzenie do modeli generatywnych

J. Tabor

27 marca 2018

Spis treści

Rozdział 1

Zabawa modelami generatywnymi

1.1 Horse to zebra and toster to cat

Jak ktoś chce się pobawić sieciami neuronowymi polecam `http://playground.tensorflow.org/`.

Załóżmy, że mamy zbiór danych $X \subset \mathbb{R}^N$ Większość modeli generatywnych jako składnik zawiera

- latent (hidden) space $Z = \mathbb{R}^D$ [przestrzeń ukryta?], używa się D od 50 do 4000 (takie widziałem, najczęściej jest chyba $D = 300$ bo jeszcze się szybko uczy, a pozwala sensownie opisywać większość obiektów)
- funkcję $\Phi : Z \to \mathbb{R}^N$ która domyślnie ma umożliwiać generowanie danych
- teraz nowy punkt dostajemy biorąc punkt z Z wylosowany zgodnie z $N(0, I_D)$, i obkładając go przez Φ.

Z powyższego widzimy, że warto rozumieć co to jest rozkład normalny, przy czym też należy skupić uwagę na sytuacji wysoko-wymiarowej.

Przykłady zastosowań modeli generatywnych i głębokich sieci neuronowych:

- horse to zebra `https://youtu.be/9reHvktowLY` and `https://steemit.com/machinelearning/@teemujvi/how-an-ai-turns-a-horse-into-a-zebra-using`
- Film „Twój Vincent”: `http://www.filmweb.pl/video/zwiastun/nr+1+polski-43872`
- różne przykłady: `https://www.kdnuggets.com/2017/04/unpaired-image-translati` `html` Praca: `https://arxiv.org/pdf/1703.10593.pdf`
- everything to cat `https://affinelayer.com/pixsrv/` oraz `https://affinelayer.com/pix2pix/`

Uwaga – pomimo dużego sukcesu, to co te metody umieją, to głównie modyfikować typ/rodzaj szeroko rozumianej tekstury. Nie widzą/rozumieją głębszych zależności (typu malarz wydłużał twarze, to nasza metoda też wydłuży), mają problemy z zamianami typu pies -> kot (przypominam, że zebra i koń mają dokładnie ten sam szkielet).

Rozdział 2

Rozkład normalny - przypomnienie podstawowych faktów

W tym rozdziale przypominam (bez dowodów) podstawowe fakty dotyczące rozkładu normalnego. Są one zasadniczo niezbędne do dalszej pracy.

2.1 Rozkład normalny jednowymiarowy

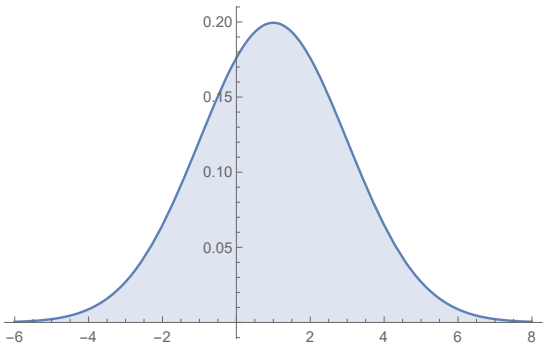
Zamieszczam jedynie podstawowe informacje, po więcej informacji odsyłam do `https://en.wikipedia.org/wiki/Normal_distribution`

ROZKŁAD NORMALNY JEDNOWYMIAROWY o średniej w m i odchyleniu standardowym σ ma gęstość daną wzorem:

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

WŁASNOŚCI:

- rozkład symetryczny, środek w m, punkty przegięcia w $m \pm \sigma$,
- prawo 3σ: poza przedziałem $[m - 3\sigma, m + 3\sigma]$ znajduje się $< 0.27\%$ danych
- ogony zmierzeją bardzo szybko do zera - rząd typu e^{-x^2}



Rysunek 2.1: Normalny o średniej $m = 1$ i odchyleniu standardowym $\sigma = 1$.

Zadanie 2.1. *Proszę pokazać, że* $C = \int N(0,1)(x)dx = 1$. *Wsk.: proszę rozważyć całkę*

$$\int_{\mathbb{R}^2} N(0,1)(x)N(0,1)(y)dxdy,$$

i zauważyć, że powyższa całka wynosi C^2. *Z drugiej strony proszę zmienić współrzędne na biegunowe i wyliczyć tę całkę.*

Zadanie 2.2. *Mamy zmienną losową* X *o standardowym rozkładzie normalny* $N(0,1)$. *Proszę dla dużych* a *oszacować wielkość ogona, czyli*

$$P(X \geqslant a) = \int_a^\infty N(0,1)(x)dx.$$

Wsk.: aby wyliczyć asymptotykę $\int_a^\infty \exp(-\frac{1}{2}x^2)dx$ *proszę zastosować całkowanie przez części:*

$$\int_a^\infty \exp(-\frac{1}{2}x^2)dx = \int_a^\infty x\exp(-\frac{1}{2}x^2)\cdot\frac{1}{x}dx =$$

$$\begin{cases} u' = x\exp(-\frac{1}{2}x^2) \\ v = 1/x \end{cases} \quad = \int_a^\infty u'v = [uv]_a^\infty - \int_a^\infty uv' = \dots$$

Następnie trzeba pokazać, że drugi czynnik jest mały w porównaniu do pierwszego.

2.2 Rozkład normalny wielowymiarowy

Aby podać definicję rozkładu wielowymiarowego będę potrzebował *metryki Mahalanobisa*. Załóżmy, że mamy zbiór danych $X = (x_i)_{i=1..N} \subset \mathbb{R}^D$ i chcemy do niego dopasować odległość euklidesową (to znaczy zadaną przez iloczyn skalarny). Wtedy

- wyliczamy średnią ze zbioru: $m_X = \mathrm{mean}(X) = \frac{1}{N}\sum_{i=1}^N x_i$,
- wyliczamy macierz kowariancji X:

$$\Sigma = \mathrm{cov}(X) = \frac{1}{N}\sum_{i=1}^N (x_i - m_X)\cdot(x_i - m_X)^T \in \mathbb{R}^{D\times D},$$

- definiujemy odległość i iloczyn skalarny Mahalanobisa

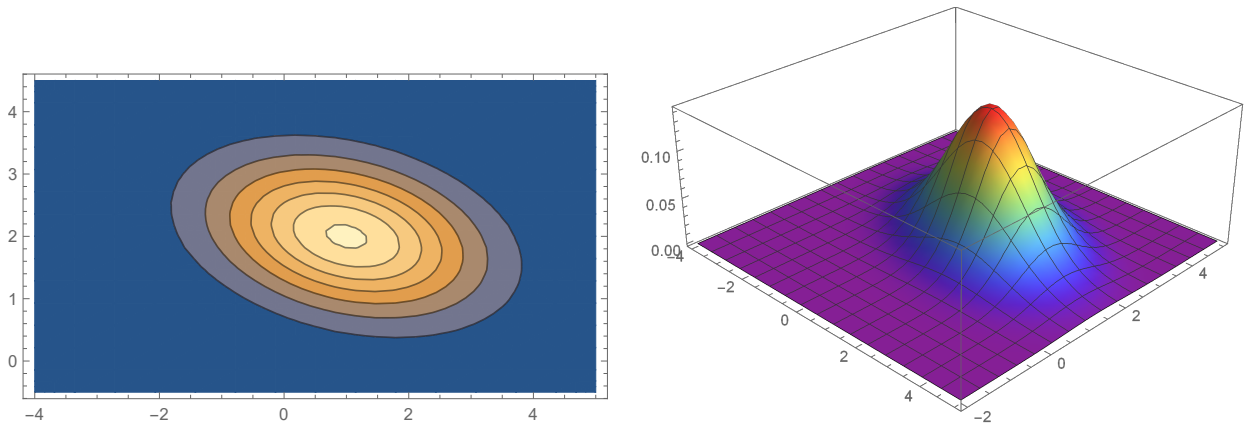
$$\|x\|_\Sigma^2 = x^T\Sigma^{-1}x,\ \langle x, y\rangle_\Sigma = x^T\Sigma^{-1}y.$$

Okazuje się, że tak zdefiniowana odległość dużo lepiej oddaje wewnętrzną strukturę danych niż kanoniczna odległość euklidesowa.

Zadanie 2.3 (python)**.** *Proszę wziąć parę przykładowych zbiorów z repozytorium UCI, zawęźić się do dwóch wymiarów, i narysować na zbiorze danych koła jednostkowe o środku w średniej danych – domyślne oraz Mahalanobisa.*

I teraz już jesteśmy w stanie zdefiniować rozkład normalny wielowymiarowy $N(m, \Sigma)$ w $\mathbb{R}^D$, gdzie m będzie średnią, a Σ macierzą kowariancji:

$$N(m,\Sigma)(x) = \frac{1}{(2\pi)^{D/2}\det^{1/2}\Sigma}\exp(-\tfrac{1}{2}\|x-m\|_\Sigma^2) \text{ dla } x \in \mathbb{R}^D.$$



Rysunek 2.2: Rozkład normalny wielowymiarowy (poziomice i widok 3D)

Ważna własność rozkładów gaussowskich, to ich niezmienniczość na transformacje afiniczne:

Stwierdzenie 2.1. *Jeżeli N wymiarowy wektor losowy X pochodzi z rozkładu normalnego $N(m, \Sigma)$ (co zapisuję skrótowo jako $X \sim N(m, \Sigma)$), i $A \in \mathbb{R}^{K \times N}$, $b \in \mathbb{R}^K$, to*[1]

$$Y = AX + b \sim N(Am + b, A\Sigma A^T).$$

2.3 Rodziny gaussowskie

Ogólna klasa rozkładów normalnych $N(m, \Sigma)$ jest czasem niewygodna w użyciu – zobaczmy, że jeżeli wymiar danych D jest duży (na przykład $D = 1000$) to macierz kowariancji jest opisywana przez około $D^2/2 = 500000$ współczynników. To prowadzi do problemów – szacowanie takiej ilości współczynników może być trudne, i może mieć dużą złożoność numeryczną.

W konsekwencji rozpatruje się klasy gaussów które wymagają mniejszej ilości parametrów, i w związku z tym są łatwiejsze w implementacji i użyciu.

GAUSSY RADIALNE / SFERYCZNE. W tym przypadku rozważamy kowariancję która jest proporcjonalna do identyczności:

$$\Sigma = \alpha I.$$

Jak widzimy dostajemy wtedy jeden parametr do estymacji. Gęstość jest dana wzorem

$$N(m, \alpha I)(x) = \frac{1}{(2\pi\alpha)^{D/2}} \exp(-\tfrac{1}{2\alpha}\|x - m\|^2).$$

Poziomice to sfery.

KOWARIANCJA DIAGONALNA. Jest to uogólnienie poprzedniego, zakładamy, że macierz kowariancji jest diagonalna:

$$\Sigma = \mathrm{diag}[\sigma_1^2, \ldots, \sigma_D^2].$$

Wtedy możemy pracować jakby na każdej współrzędnej z osobna (współrzędne są niezależne). Wzór na gęstość się faktoryzuje po współrzędnych:

$$N(m, \mathrm{diag}[\sigma_1^2, \ldots, \sigma_D^2])(x) = N(m_1, \sigma_1^2)(x_1) \cdot \ldots \cdot N(m_D, \sigma_D^2)(x_D).$$

Poziomice to sfery o osiach równoległych do osi układu współrzędnych. Wygodne w użyciu, ilość parametrów do oszacowania taka sama jak wymiar danych.

2.4 Generowanie rozkładów za pomocą mieszanek gaussowskich

Nasze zadanie brzmi następująco – mamy zbiór $X = (x_i)_{i=1..n} \subset \mathbb{R}^D$ wygenerowany przez rozkład o nieznanej gęstości f, i chcemy dokonać estymacji (oszacowania) gęstości f. Dodatkowo tutaj zawężamy się do sytuacji, gdy f jest dany za pomocą kombinacji gaussów.

[1] Poniższe stwierdzenie zachodzi także w przypadku gdy A nie ma pełnego rzędu, ale wtedy Y ma rozkład normalny osobliwy – pomijam milczeniem definicję rozkładów normalnych osobliwych (singular normal density), czyli takich dla których rozkład mieści się na podprzestrzeni afinicznej, co jest równoważne temu, że macierz kowariancji jest nieodwracalna.

Są dwa najczęściej stosowane podejścia, jedno to GMM (gaussian mixture models = mieszanki gaussowskie, bazuje na EM), drugie na KDE (kernel density estimation). UWAGA: opisane tu metody działają wiarygodnie dla małych wymiarów ($D < 5$), dla większych wymiarów trzeba podchodzić do wyników nieufnie.

GMM. W podejściu GMM zakładamy, że nasza gęstość f jest przybliżana przez mieszankę niewielkiej ilości k dowolnych gaussów (zazwyczaj $k < 20$):

$$f \approx \sum_{l=1}^{K} p_l N(m_l, \Sigma_l).$$

Podejście to ma bardzo łatwą iteracyjną procedurę (EM = expectation maximization, zbliżona do k-means), która taką aproksymację optymalizuje. Jeżeli mamy taką mieszankę, to z niej losuje się już prosto: z prawdopodobieństwem p_i wybieramy rozkład i-ty $N(m_l, \Sigma_l)$, i z niego wtedy generuję losowo punkt.

Podejście jest ryzykowne w dużym wymiarze, bo może być zbyt dużo parametrów przy kowariancjach do wiarygodnego oszacowania (w związku z tym czasami się używa modeli gaussowskich opisanych w poprzedniej sekcji zamiast dowolnych gaussów aby zmniejszyć liczbę parametrów). Dodatkowo metoda może nie zadziałać dobrze, gdy dane mają jakąś bogatszą geometrię, i nie składają się z niewielkiej ilości elipsoidalnych grup (na przykład spirala).

KDE. UWAGA: domyślnie w tym podejściu zakładamy, że zbiór jest po whiteningu, czyli kowariancja jest równa identyczność.

W podejściu KDE zakładamy, że nasza szukana gęstość f powstaje przez rozmycie gaussowskie zbioru danych X:

$$f \approx \frac{1}{n} \sum_{i=1}^{n} N(x_i, h^2 I).$$

Inaczej mówiąc, rozmywamy każdy punkt. Powyższe ma interpretację fizyczną: a mianowicie możemy to interpretować w ten sposób, że w każdym punkcie zbioru danych mamy skupioną energię cieplną, puszczamy czas, i patrzymy w jaki sposób się ciepło rozchodzi. Losowanie z takiego rozkładu jest łatwe, tak samo jak poprzedni po prostu z prawdopodobieństwem $1/n$ losujemy punkt z rozkładu $N(x_i, h^2 I)$.

Zauważmy, że dla małych h mamy overfitting, bo nasza aproksymacja będzie zasadniczo skupiona w zbiorze danych, zaś dla bardzo dużych h wszystko nam się rozmyje. W związku z tym powstaje bardzo ważne pytanie, jaki jest optymalny dobór h.

WZÓR SILVERMANA. Istnieje wzór Silvermana na optymalne h:

$$h = h_{opt} = \left(\tfrac{4}{n(d+2)}\right)^{1/(d+4)}. \tag{2.1}$$

Proszę zauważyć, że wraz ze wzrostem wymiaru (przy ustalonym n) powyższa funkcja zmierza do 1. Wzór Silvermana jest wyliczony przy założeniu, że dane są gaussowskie – czyli praktycznie rzecz biorąc stanowią jedną spójną grupę. Jeżeli dane nie są gaussowskie, to wzór Silvermana nie działa za dobrze, i ma tendencje do nadmiernego rozmywania.

WALIDACJA KRZYŻOWA. Efekty są zazwyczaj dużo lepsze, gdy h dobiera się przy pomocy walidacji krzyżowej, tzn. dzielimy zbiór na testowy X_T i walidujący X_W (zazwyczaj jest to 5-krotna lub 10-krotna walidacja). Dobieramy h tak by gęstość wyliczona na bazie X_T dawała optymalną wartość log-likelihood na X_W:

$$h_{opt} = \operatorname{argmax}\{ \sum_{x \in X_W} \log[\tfrac{1}{|X_T|} \sum_{z \in X_T} N(z, h^2 I)(x)]\}.$$

Potencjalnie jest oczywiście dosyć wolne.

W przypadku walidacji krzyżowej można się spytać, czy nie należy szukać kernel-a (czyli w naszym przypadku $h^2 I$) w klasie szerszej niż przeskalowanie identyczności. Ogólnie jest to trudny problem, ale łatwo jest w przypadku kowariancji diagonalnych, bo nam się problem rozbija na D zadań jednowymiarowych, i w każdym z osobna robimy przeszukiwanie.

Zadanie 2.4. *Załóżmy, że chcemy by $h \approx 1/2$. Proszę wyliczyć z* (2.1) *w zależności od wymiaru d, ile powinniśmy mieć punktów n_d w zbiorze danych. Ile wynosi n_{10}, n_{100}?*

Zadanie 2.5 (python)**.** *Proszę zaimplementować w sytuacji jednowymiarowej szukanie h przez jednokrotną walidację krzyżową.*

Zadanie 2.6 (python)**.** *Proszę wylosować $n = 100$ danych z rozkładu*

$$\tfrac{1}{2}N(-2,1) + \tfrac{1}{2}N(2,1).$$

Proszę porównać na rysunku prawdziwą gęstość, z KDE danym przez wzór Silvermana oraz walidację krzyżową.

Rozdział 3

Porównywanie dwóch rozkładów

Często spotkamy się z problemem porównania ze sobą dwóch gęstości (czasami w sytuacji gdy nie mamy gęstości, a mamy tylko próbkę). Przypominam, że f jest gęstością, o ile

$$f \geqslant 0, \int f(x)dx = 1.$$

W naszym przypadku będziemy szczególnie zainteresowani sytuacją, gdy te gęstości są dane za pomocą rozkładów normalnych lub ich mieszanek. Zazwyczaj do porównywania używamy pojęcia odległości (metryki).

3.1 Metryka Hellingera

Metryk na rozkładach jest dosyć dużo, jedna z często spotykanych to *metryka Hellingera:*

$$H^2(f,g) = \tfrac{1}{2}\int(\sqrt{f(x)} - \sqrt{g(x)})^2 dx = 1 - \int \sqrt{f(x)g(x)}dx.$$

Z powyższych wzorów łatwo widać, że

$$H(f,g) \in [0,1].$$

Potencjalnie powyższa metryka jest fajna, bo jest zadana przez iloczyn skalarny:

$$\langle f,g\rangle_H = \tfrac{1}{2}\int \sqrt{f(x)}\sqrt{g(x)}dx$$

Ma dodatkowo bonus, że jest dobrze zdefiniowana dla wszystkich gęstości. W praktycznych implementacjach rzadko się ją używa, bo nie ma jawnych wzorów dla mieszanek gaussowskich.

Okazuje się, że można łatwo policzyć odległość Hellingera dla dwóch gaussów. Będziemy potrzebować następującego wzoru:

$$\int N(m_1,\Sigma_1)(x)(m_2,\Sigma_2)(x)dx = N(m_1 - m_2, \Sigma_1 + \Sigma_2)(0). \tag{3.1}$$

3.2 Norma L^2

Drugą często spotykaną metryką na rozkładach jest norma L^2 zadana przez iloczyn skalarny:

$$\|f-g\|_{L^2}^2 = \int |f(x)-g(x)|^2 dx \text{ oraz } \langle f,g\rangle_{L^2} = \int f(x)g(x)dx.$$

Oczywiście mamy

$$\|f-g\|_{L^2}^2 = \langle f,f\rangle_{L^2} - 2\langle f,g\rangle_{L^2} + \langle g,g\rangle_{L^2}.$$

Pozornie wydaje się zbliżone do metryki Hellingera, ale ma przewagę[1] polegającą na tym, że jest dana jawnym wzorem dla mieszanek gaussowskich.

Oczywiście mamy zawsze

$$\langle \sum_i f_i, \sum_j g_j\rangle = \sum_{i,j}\langle f_i, g_j\rangle.$$

W konsekwencji, dla f, g danych przez mieszanki gaussowskie

$$f = \sum_i p_i N(m_i, \Sigma_i), g = \sum_j \tilde{p}_j N(\tilde{m}_j, \tilde{\Sigma}_j)$$

dostajemy korzystając z (3.1)

$$\langle f, g\rangle_{L^2} = \sum_{i,j} p_i \tilde{p}_j N(m_i - \tilde{m}_j, \Sigma_i + \tilde{\Sigma}_j)(0).$$

Zadanie 3.1. *Korzystając z* (3.1)*, proszę wyliczyć jawny wzór na*

$$H^2(f,g)$$

dla f, g *gaussowskich. Wsk: proszę pokazać, że* $\sqrt{N(m,\Sigma)} = C_{m,\Sigma} N(m, 2\Sigma)$ *dla pewnej stałej* $C_{m,\Sigma}$.

Zadanie 3.2 (z *)**.** *Proszę udowodnić* (3.1).

Zadanie 3.3. *Proszę policzyć*

$$\|f-g\|_{L^2} \text{ dla } f = N(-1,1), g = N(1,1)$$

3.3 Test normalności Bowmana-Fostera

[1] Ma też pewne minusy, a mianowicie nie jest do końca zgodna ze strutkurą gęstości – a mianowicie nie jest zupełna.

Rozdział 4

Entropia oraz dywergencja Kullbacka-Leiblera

W poniższym rozdziale zrobimy wprowadzenie do entropii oraz dywergencji Kullbacka-Leiblera, najczęściej stosowanej miary porównywania rozkładów w nauczaniu maszynowym.

4.1 Entropia: przypadek dyskretny i ciągły

Do wprowadzenia dywergencji Kullbacka-Leiblera będziemy potrzebowali pojęcia entropii. Postaram się tu pokazać główną ideę, natomiast nie przedstawiam wielu dowodów.

Entropia. Zakładamy, że mamy źródło S które wysyła nam sygnały s_i z prawdopodobieństwem p_i. Chcemy teraz kodować te sygnały za pomocą kodów binarnych o odpowiednio długościach l_i. Zakładamy dodatkowo, że kody te mają być jednoznacznie dekowalne. Okazuje się wtedy, na podstawie nierówności Krafta, że warunek na istnienie kodu jednoznacznie dekodowalnego o długościach l_i to

$$\sum_i 2^{-l_i} \leqslant 1. \tag{4.1}$$

Oczekiwana długość kodu sygnału z S to oczywiście

$$\sum_i p_i l_i. \tag{4.2}$$

Używając wykładników Lagrange'a można łatwo zminimalizować powyższą wartość przy warunku (4.1), dostajemy łatwo, że globalne minimum jest realizowane przy $l_i = -\log_2 p_i$ (i jest jedyne), co oznacza, że minimalna wartość oczekiwana dana jest wzorem

$$h(p) = \sum_i p_i \cdot -\log_2 p_i \text{ gdzie } p = (p_i).$$

Powyższą wartość nazywamy *entropią* źródła (dyskretnego) S.

Wprost z definicji widzimy, że entropia (jako minimalna oczekiwana długość kodu) jest funkcją nieujemną.

Entropię uogólnia się na przypadek ciągły – wtedy zastępuje się $\log_2$ przez logarytm naturalny, i mówi się wtedy często o *entropii różniczkowej*. Czyli dla gęstości f na $\mathbb{R}^D$ kładziemy

$$h(f) = \int_{\mathbb{R}^D} f(x) \cdot -\ln f(x) dx.$$

Jeżeli X jest wektorem losowym o gęstości f_X, to używam oznaczenia

$$h(X) = h(f_X).$$

Pokażę wzór na entropię rozkładu normalnego. W tym celu przydadzą się nam następujące obserwacja.

Obserwacja 4.1. *Mamy*

$$h(N(0,I)) = \frac{D}{2}\ln(2\pi e).$$

Dowód. Mamy

$$\begin{aligned} &-\int_{\mathbb{R}^D} N(0,I)(x)\cdot[-\frac{D}{2}\ln(2\pi) + \|x\|^2/2]dx \\ &= \frac{D}{2}\ln(2\pi) - \frac{1}{2}\int_{\mathbb{R}^D} N(0,I)(x)\|x\|^2 dx. \end{aligned}$$

Teraz mamy, korzystając z faktu $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ (tak zwany „trace trick") oraz tego, że macierz kowariancji $N(m,\Sigma)$ to Σ:

$$\int_{\mathbb{R}^D} N(0,I)(x)\|x\|^2 dx = \int_{\mathbb{R}^D} N(0,I)(x)\mathrm{tr}(x^T x)dx$$

$$= \int_{\mathbb{R}^D} N(0,I)(x)\mathrm{tr}(xx^T)dx = \mathrm{tr}\big(\int_{\mathbb{R}^D} N(0,I)(x)xx^T dx\big) = \mathrm{tr}(I) = D,$$

co kończy dowód. □

Obserwacja 4.2. *Niech X będzie wektorem losowym i niech $\phi : x \to Ax + b$ będzie odwracalnym odwzorowaniem afinicznym. Wtedy*

$$h(AX+b) = \ln|\det A| + h(X).$$

Dowód. Jak wiemy gęstość $Y = AX + b$ dana jest wzorem

$$f_Y(y) = |\det A|^{-1} f(\phi^{-1}y).$$

Wtedy

$$\begin{aligned} h(f_Y) &= -\int_{\mathbb{R}^D} |\det A|^{-1} f(\phi^{-1}y)\ln[|\det A|^{-1} f(\phi^{-1}y)]dy \\ &= \int_{\mathbb{R}^D} |\det A| f(\phi^{-1}y)\ln[|\det A| f(\phi^{-1}y)]dy \\ &= \ln|\det A| - \int_{\mathbb{R}^D} |\det A|^{-1} f(\phi^{-1}y)\ln f(\phi^{-1}y)dy. \end{aligned}$$

Oczywiście

$$\int_{\mathbb{R}^D} |\det A|^{-1} f(\phi^{-1}y)\ln f(\phi^{-1}y)dy = \left[\begin{smallmatrix} y = \phi(x), \\ dy = |\det A| dx \end{smallmatrix}\right] = h(f).$$

□

Teraz jesteśmy już gotowi do wyliczenia entropii rozkładu normalnego.

Twierdzenie 4.1. *Mamy*

$$h(N(m,\Sigma)) = \tfrac{D}{2}\ln(2\pi e) + \tfrac{1}{2}\det\Sigma.$$

Dowód. Niech X zmienna losowa o rozkładzie $N(0,I)$. Wyliczymy najpierw takie A i b by $AX+b \sim N(m,\Sigma)$. Ponieważ

$$AX + b \sim N(A0 + b, AIA^T)$$

wystarczy położyć

$$A = \Sigma^{1/2} \text{ oraz } b = m.$$

Korzystając z poprzedniej obserwacji dostajemy

$$h(N(m,\Sigma)) = h(AX+b) = \ln|\Sigma^{1/2}| + h(N(0,I)) = \frac{1}{2}\log|\Sigma| + h(N(0,I)).$$

Obserwacja 4.1 kończy dowód. □

Zadanie 4.1. *Korzystając z mnożników Lagrange'a proszę zminimalizować* (4.2) *przy warunku* (4.1).

4.2 Dywergencja Kullbacka-Leiblera: przypadek dyskretny

Dywergencja Kullbacka-Leiblera $D_{KL}(f,g)$ jest także miarą podobieństwa dwóch rozkładów f i g, z tym, że w przeciwieństwie do miar opisanych w poprzedniej sekcji nie jest symetryczna, co skutkuje tym, że nie jest metryką.

Dywergencja Kullbacka-Leiblera. Załóżmy teraz, że dostaliśmy od kogoś kod zoptymalizowany dla częstości q_i (czyli o długościach $-\log_2 q_i$), i chcemy się zorientować ile średnio stracimy bitów informacji kodując za pomocą kodów zoptymalizowanych do q_i zamiast do p_i. Wtedy interesuje nas

$$D_{KL}(p,q) = \sum_i p_i \cdot (-\log_2 q_i) - \sum_i p_i \cdot (-\log_2 p_i) = \sum_i p_i \log_2(p_i/q_i).$$

Powyższą wartość nazywamy *dywergencją Kullbacka-Leiblera*. Oczywiście, z faktu, że globalne minimum entropii jest jednoznacznie wyznaczone przez długości $-\log_2 p_i$, dostajemy, że

$$D_{KL}(p,q) \geqslant 0 \text{ oraz } D_{KL}(p,q) = 0 \text{ wtw. gdy } p = q.$$

Aby pokazać interpretację przez log-likelihood (likelihood=wiarygodność), przypomnę najpierw tę metodę.

Załóżmy, że mamy rozkład dyskretny w którym możemy wylosować punkty z $X = \{x_1,\ldots,x_k\}$. Zakładamy, że mamy próbkę n-elementową wylosowaną z naszego rozkładu i punkt x_i wylosowaliśmy n_i razy – czyli możemy oszacować prawdopodobieństwo p_i wylosowania x_i jako

$$p_i = n_i/n.$$

Zakładamy teraz, że mamy sparametryzowaną rodzinę $(p^\theta)_{\theta\in\Theta}$ rozkładów prawdopodobieństwa na X. Wtedy przez p_i^θ oznaczamy prawdopodobieństwo wylosowania punktu x_i.

Problem na jaki stara się odpowiedzieć metoda największej wiarygodności (MLE=maximium likelihood estimation) jest następujący:

Problem 4.1. *Jak dobrać $\theta \in \Theta$ by rozkład p^θ najbardziej „przypominał" nasze dane?*

Odpowiedź jest następująca: wybieramy ten parametr θ dla którego prawdopodobieństwo wylosowania próbki $\{(x_i, n_i)_i\}$ jest maksymalne:

$$p_\theta((x_i, k_i)) = \prod_i (p_i^\theta)_i^n.$$

Ponieważ zamiast mnożyć łatwiej dodawać, w MLE maksymalizujemy zazwyczaj log-likelihood naszej próbki:

$$\operatorname*{argmax}_\theta \sum_i n_i \ln p_i^\theta.$$

Dzielac przez n widzimy, że możemy równoważnie maksymalizować

$$\operatorname*{argmax}_\theta \sum_i p_i \ln p_i^\theta.$$

Teraz rozpatrzymy odpowiedź teorio-informatyczną na Problem 4.1. Otóż wybierzmy ten parametr θ, dla którego najmniej tracimy na kodowaniu naszej próbki za pomocą kodu dostowanego przez p^θ:

$$\operatorname{argmin} D_{KL}(p\|p_\theta).$$

Rozpisując powyższe, widzimy, że dostajemy

$$\operatorname*{argmin}_\theta D_{KL}(p\|p_\theta) = \operatorname*{argmin}_\theta \sum_i p_i \log_2(p_i/p_i^\theta)$$

$$= \operatorname*{argmin}_\theta [h(p) - \sum_i p_i \log p_i^\theta] = \operatorname*{argmax}_\theta \sum_i p_i \log_2 p_i^\theta.$$

Widzimy więc, że dostajemy dokładnie ten sam wzór co dla MLE!

4.3 Wzór dla rozkładów normalnych

Dywerencja Kullbacka-Leiblera uogólnia się także na przypadek ciągłych rozkładów $p(x), q(x)$, z tym, że wtedy dla wygody rachunkowej zastępuje się zwykle $\log_2$ przez $\ln$:

$$D_{KL}(p, q) = \int p(x) \ln(p(x)/q(x))dx.$$

Wyliczymy jawny wzór na wartość dywergencji pomiędzy dwoma rozkładami normalnymi. Pokażemy najpierw ważną własność dywergencji KL, a mianowicie niezależność na transformacje afiniczne. Wprowadzam jeszcze oznaczenie – dla wektorów losowych X, Y o gęstościach odpowiednio p, q kładę

$$D_{KL}(X\|Y) = D_{KL}(p\|q).$$

Przypominam wzór: jeśli wektor losowy X ma gęstość f, a Φ jest odwracalną funkcją różniczkowalną (formalnie dyffeomorfizm), to $\Phi(X)$ ma gęstość

$$f_\phi(z) = \frac{1}{|d_{\Phi^{-1}(z)}\Phi|} f(\Phi^{-1}(z)), \tag{4.3}$$

gdzie $|A|$ oznacza wyznacznik macierzy A.

Stwierdzenie 4.1. *Niech X, Y będą zmiennymi losowymi o gęstościach f, g i niech $x \to Ax+b$ będzie odwracalnym odwzorowaniem afinicznym. Wtedy*

$$D_{KL}(X\|Y) = D_{KL}(AX + b\|AY + b).$$

Dowód. Kluczowe jest podanie wzoru na gęstości f_W, g_Z zmiennych $W = AX+b$ i $Z = AY+b$ (następnie wystarczy tylko zmienić zmienne w całkowaniu) – wprost z (4.3) zastosowanego dla funkcji $\Phi(x) = Ax + b$, mamy

$$f_W(z) = \frac{1}{|A|} f(\Phi^{-1}z) \text{ oraz } g_Z = \frac{1}{|A|} g(\Phi^{-1}z).$$

I teraz stosując odpowiednią zmianę zmiennych dostajemy

$$D_{KL}(AX + b\|AY + b) = D_{KL}(f_W\|g_Z) = \int f_W(z) \ln(f_W(z)/g_W(z))dz$$

$$= \int \frac{1}{|A|} f(\Phi^{-1}z) \ln(f(\Phi^{-1}z)/g(\Phi^{-1}z))dz =$$

$$= \left[\begin{array}{c} x = \Phi^{-1}z \\ dx = \frac{1}{|A|}dz \end{array} \right] = \int f(x) \ln(f(x)/g(x))dx = D_{KL}(X\|Y).$$

□

Na podstawie powyższego wzoru postaram się pokazać w punktach jak można wyliczyć wzór na D_{KL} pomiędzy rozkładami normalnymi.

Twierdzenie 4.2. *Dla rozkładów normalnych w $\mathbb{R}^D$ mamy*

$$D_{KL}(N(m_1, \Sigma_1)\|N(m_2, \Sigma_2)) = \tfrac{1}{2} \left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) - D + \|m_2 - m_1\|^2_{\Sigma_2} - \ln\det(\Sigma_2^{-1}\Sigma_1) \right).$$

Dowód. Pokażę dwa główne kroki w rozumowaniu.

ETAP 1. Pokażę, że możemy się zredukować do rozważenia przypadku

$$D_{KL}(N(m, \Sigma)\|N(0, I)).$$

Niech X_i oznacza rozkład normalny o gęstości $N(m_i, \Sigma_i)$. Korzystając z wcześniejszych wzorów na rozkład normalny po transformacji $x \to Ax + b$ mamy

$$AX_i + b \sim N(Am_i + b, A\Sigma_i A^T).$$

Teraz wystarczy tak dobrać A i b, aby $AX_2 + b \sim N(0, I)$, czyli by

$$Am_2 + b = 0,\ A\Sigma_2 A^T = I.$$

Podobnie jak wcześniej rozwiązaniem (jedynym w klasie macierzy symetrycznych) jest

$$A = \Sigma_2^{-1/2} \text{ i w konsekwencji } b = -A^{-1}m_2 = -\Sigma_2^{-1/2}m_2.$$

Konkludując dostajemy

$$D_{KL}(N(m_1, \Sigma_1)\|N(m_2, \Sigma_2)) = D_{KL}(N(m, \Sigma)\|N(0, I))$$

dla

$$m = Am_1 + b = \Sigma_2^{-1/2} m_1 - \Sigma_2^{-1/2} m_2,$$
$$\Sigma = \Sigma_2^{-1/2} \Sigma_1 \Sigma_2^{-1/2}.$$

ETAP 2. Używając „trace trick” wyliczymy $D_{KL}(N(m,\Sigma)\|N(0,I))$.
Mamy

$$D_{KL}(N(m,\Sigma)\|N(0,I)) = h(N(m,\Sigma)) - \int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot \ln N(0,I)(x) dx.$$

Oczywiście wzór na entropię rozkładu normalnego znamy. Pozostaje nam zatem wyliczenie drugiej części powyższego wzoru:

$$\int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot \ln N(0,I)(x) dx = \int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot [-\tfrac{D}{2}\ln(2\pi) - \tfrac{1}{2}\|x\|^2] dx.$$

Oczywiście

$$\int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot \tfrac{D}{2}\ln(2\pi) dx = \tfrac{D}{2}\ln(2\pi).$$

Korzystając podobnie jak w przypadku wyliczenia entropii, z „trace trick”, dostajemy

$$\int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot \|x\|^2 dx = \int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot \mathrm{tr}(x^T x) dx$$

$$= \int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot \mathrm{tr}(xx^T) dx = \mathrm{tr}\Big(\int_{\mathbb{R}^D} N(m,\Sigma)(x) \cdot xx^T dx\Big) = \mathrm{tr}(\Sigma).$$

Łącząc wszystkie powyższe fakty dostajemy tezę twierdzenia. □

Korzystając z powyższych etapów, można łatwo pokazać, że końcowy wzór przyjmuje postać

Zadanie 4.2. *Czy teza Stwierdzenia 4.1 zachodzi dla dowolnych odwracalnych różniczkowalnych Φ (niekoniecznie afinincznych)? Proszę udowodnić, albo podać kontrprzykład.*

Zadanie 4.3. *Proszę uzupełnić dowód Twierdzenia 4.2.*

Rozdział 5

Obiekty wysokowymiarowe

5.1 Objętość figur D-wymiarowych

Przez kostkę D-wymiarową P w $\mathbb{R}^D$ rozumiem dowolny zbiór postaci

$$P = \{x = (x_1, \ldots, x_D) : x_i \in [p_i, q_i]\}.$$

Powyższe zapisujemy też jako

$$P = \bigtimes_i [p_i, q_i].$$

Teraz brzeg kostki definiuje się jako

$$\partial P = \{x \in P \mid \exists i : x_i \in \{p_i, q_i\}\}.$$

Przypominam definicję objętości dla kostki wielowymiarowej:

$$\lambda(P) = \prod_i |q_i - p_i|$$

Jeżeli teraz dana figura W składa się sumy parami rozłącznych kostek P_l (gdzie dopuszczam przecięcie brzegów), czyli

$$W = \bigcup_l P_l,$$

to kładziemy

$$\lambda_D(W) = \sum_l \lambda_D(P_l).$$

Jeżeli mamy dowolny „porządny” zbiór $W \subset \mathbb{R}^D$, to możemy go dowolnie dokładnie przybliżyć od góry przez zbiór będący sumą malutkich kostek (analogicznie z dołu). Wtedy, jeżeli różnica między objętością z góry i z dołu jest asymptotycznie zero, to definiujemy objętość jako granicę dowolnej z tych wartości.

Można pokazać, że ogólnie dla zbioru $W \subset \mathbb{R}^D$ D-wymiarowa objętość W jest dana przez

$$\lambda_D(W) = \int_W 1 dx.$$

Korzystając z powyższego dostajemy stosując zmianę zmiennych:

Wniosek 5.1. *Dla odwracalnego odzworowania liniowego $\phi : x \to Ax + b$ mamy*

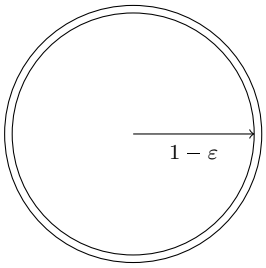
$$\lambda_D(\phi(W)) = |\det A| \cdot \lambda_D(W).$$

Własność zbiorów wysokowymiarowych - wszystko przy brzegu:

$$\lambda((1-\varepsilon)A) = (1-\varepsilon)^D \lambda(A).$$

Czyli

$$\frac{\lambda((1-\varepsilon)A)}{\lambda(A)} = (1-\varepsilon)^D \leqslant e^{-\varepsilon D}.$$



Rysunek 5.1: Kula.

Załóżmy więc, że dajemy ε małe, ale wymiar duży.
Wtedy z powyższego wzoru widzimy, że dla odpowiedniego wymiaru i tak wszystko jest przy brzegu.

Proszę zauważyć, że powyższe twierdzenie jest prawdziwe dla zbiorów wypukłych, w szczególności także dla kostki wielowymiarowej.

Zadanie 5.1 (z *)**.** *Proszę pokazać, że pole prostokąta nie zależy od tego w jaki sposób podzielimy go na parami rozłączne prostokąty.*

5.2 Obiekty niżejwymiarowe

Bardzo ważnym pojęciem są obiekty niżej wymiarowe – najważniejszą klasę takich obiektów tworzą rozmaitości. Jest to uogólnienie krzywych i powierzchni (sfery, płaszczyzny) na przypadek wysokowymiarowy.

Ogólnie będą nas interesowały nas powierzchnie i rozmaitości, czyli funkcje dane (przynajmniej) lokalnie przez obrazy funkcji różniczkowalnych w sposób ciągły (i o rzędzie maksymalnym) $\phi : \mathbb{R}^D \to \mathbb{R}^N$.

Niech M będzie zbiorem. Mówimy, że różniczkowalna funkcja

$$\phi : U \to V$$

jest D-wymiarową mapą w otoczeniu punktu $m \in M$, jeżeli U jest otwartym otoczeniem m, a

Definicja 5.1. Mówimy, że $M \subset \mathbb{R}^N$ jest rozmaitością (różniczkową) D-wymiarową, jeżeli jest lokalnie dana jako obraz

$$\phi(U).$$

[doprecyzowac ? atlas]

jest globalnie dany jako obraz funkcji ϕ

5.3 Objętość kuli i sfery

Przyda się nam nam informacja jak można całkować funkcje radialne [informacyjnie].

Otóż jak mamy funkcję $f : \mathbb{R}^D \to \mathbb{R}$ to możemy jej całkę policzyć całkując po promieniach i sferach:

$$\int_{\mathbb{R}^D} f(x)dx = \int_0^\infty \int_{S_r} f(x)dr.$$

Teraz stosując zmianę zmiennych łatwo dostajemy

$$\int_{S_r} f(x)dx = r^{D-1} \int_{S_1} f(rx)dx,$$

gdyż sfera S jest obiektem $D-1$ wymiarowym. Czyli

$$\int_{\mathbb{R}^D} f(x)dx = \int_0^\infty r^{D-1} \int_{S_1} f(rx)dxdr \tag{5.1}$$

Przykład 5.1. Policzmy najpierw (znamy) wartość $I = \int e^{-x^2}dx$. Mamy oczywiście

$$I \cdot I = \int e^{-x^2}dx \cdot \int e^{-y^2}dy = \int e^{-x^2-y^2}dxdy.$$

Skoro jest to funkcja radialna (zależy tylko od odległości od zera), zmienia się łatwo na współrzędne biegunowe, i tu korzystamy z poprzedniego wzoru

$$I^2 = \int_0^\infty \int_{S_r} e^{-r^2}d\omega dr = \int_0^\infty 2\pi r e^{-r^2}dr = \pi \int_0^\infty e^{-u}du = \pi.$$

Czyli $I = \sqrt{\pi}$.

Korzystając z tego policzymy łatwo objętość kuli i sfery. Zaczniemy najpierw od sfery jednostkowej. Wprowadzam oznaczenie

$$A(d) = \lambda_{D-1}(S_1).$$

Oczywiście wtedy mamy

$$\lambda_{D-1}(S_r) = A(d)r^{D-1}.$$

Aby to wyliczyć, zrobimy całkę

$$I(d) = \int e^{-x_1^2+\ldots+x_d^2}dx = \Big[\int e^{-x^2}dx\Big]^d = \pi^{d/2}.$$

Ale z drugiej strony na podstawie (5.1)

$$I(d) = \int_0^\infty r^{d-1} \int_{S_1} e^{-r^2}dxdr = A(d) \int_0^\infty e^{-r^2}r^{d-1}dr.$$

Ale

$$\int_0^\infty e^{-r^2}r^{d-1}dr = [r^2 = t] = \frac{1}{2}\int_0^\infty e^{-t}t^{\frac{d}{2}-1}dt = \frac{1}{2}\Gamma(\frac{d}{2}),$$

gdzie Γ oznacza funkcję Gamma Eulera (patrz Zadanie 5.2). Czyli dostajemy

$$A(d) = \frac{\pi^{d/2}}{\frac{1}{2}\Gamma(d/2)}.$$

Teraz już łatwo policzymy objetosc kuli D-wymiarowej, oznaczenie $B_r = B(0,r)$. Jak poprzednio przechodzimy na współrzędne sferyczne:

$$\lambda_D(B_r) = \int_{B_r} 1dx = \int_0^r \int_{S_r} 1dxdr = \int_0^r A(d)r^{D-1}dr = \frac{A(D)}{D}r^D.$$

Powyższe wzory mają istotne konsekwencje jeżeli chodzi o losowanie punkty z kuli D-wymiarowych. Otóż dla małego D, możemy losować punkt z kostki $[-1,1]^D$ i odrzucać, jeżeli wypadł poza kulą. W przypadku wysokich wymiarów nie ma to sensu, bo w zasadzie nie da się trafić!

W związku z tym, przechodzi się na współrzędne sferyczne

$$B_1 \ni x \to [\|x\|, \frac{x}{\|x\|}] \in [0,1] \times S_1$$

i losuje się niezależnie promień $r \in [0,1]$ (odległość od zera) oraz punkt v na sferze S_1. Mając je już wylosowane kładziemy

$$x = rv.$$

Oczywiście, powstaje pytanie jak wylosować promień i punkt na sferze. Wbrew pozorom punkt v na sferze losujemy prosto:

$$v = \frac{x}{\|x\|} \text{ gdzie } x \text{ wylosowany z rozkładu } N(0, I_D).$$

Pozostaje jedynie pytanie w jaki sposób wylosować promień.

Stwierdzenie 5.1. *Promień (odległość od zera) $R = \|X\|$ losowo wybranego punktu X z rozkładu jednostajnego na B_1 ma rozkładu potęgowy o dystrybuancie*

$$P(R \leqslant r) = \begin{cases} 0 \textit{ dla } r < 0, \\ r^D \textit{ dla } r \in [0,1], \\ 1 \textit{ dla } r > 1, \end{cases}$$

i gęstości

$$f(r) = \tfrac{1}{D} r^D \mathbb{1}_{[0,1]}.$$

Dowód. Niech X będzie wektorem losowym mającym rozkład jednostajny na kuli jednostkowej. Aby wyznaczyć rozkład R odległości X od zera, wystarczy policzyć dystrybuantę (dla $r \in [0,1]$):

$$P(R \leqslant r) = \lambda_D(B_r)/\lambda_D(B_1) = r^D.$$

□

W przypadku gdy mamy ciągłą dystrybuantę Φ danego rozkładu R, to losuje się bardzo prosto (tak zwana metoda odwracania dystrybuanty): losujemy punkt U z rozkładu jednostajnego na $[0,1]$, i R uzyskujemy biorąc

$$\Phi^{-1}(U).$$

Waracając do naszego przykładu, aby wylosować promień R losujemy punkt U z rozkładu jednostajnego na $[0,1]$ i bierzemy

$$R = \sqrt[D]{U}.$$

Zadanie 5.2. *Funkcja Γ Eulera (uogólnienie silni) dana jest wzorem*

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt.$$

Proszę sprawdzić indukcyjnie, że dla x naturalnych $\Gamma(x) = (x-1)!$. Wsk.: całkowanie przez części.

Zadanie 5.3. *Proszę sprawdzić w zależności od wymiaru jaki procent kostki $[-1,1]^D$ stanowi kula B_1 w niego wpisana.*

Zadanie 5.4. *Ile należy wylosować punktów z kostki $[-1,1]^D$, dla $D = 100$, by mieć szansę 90 procent na trafienie do kuli B_1.*

Zadanie 5.5 (python)**.** *Proszę zaimplementować losowanie punktu z rozkładu jednostajnego na kuli jednostkowej.*

5.4 Rozkład χ^2_D

Przyda się nam rozkład χ^2, czyli rozkład kwadratu odległości od zera w wielowymiarowym rozkładzie normalnym. Jest to rozkład zmiennej losowej

$$\|X\|^2 \text{ gdzie } X \sim N(0, I_D).$$

Wprost z definicji widzimy, że

$$X_1^2 + \ldots + X_D^2 \sim \chi^2_D \text{ dla niezależnych zmiennych} X_i \in N(0,1).$$

Ale powyższe oznacza, że dla $Z \sim \chi^2_D$ mamy

$$E(Z) = D \text{ oraz } V(Z) = D,$$

gdyż wartość oczekiwana sumy to suma wartości oczekiwanych, a wariancja sumy niezależnych zmiennych losowych to suma wariancji. Korzystając teraz z centralnego tw. granicznego dostajemy, że

$$\frac{1}{\sqrt{D}}(Z - D) \approx N(0,1),$$

czyli

$$Z \approx N(D, D).$$

W praktyce można pokazać, że powyższe przybliżenie jest bardzo dokładne dla $D \geqslant 30$.

Zobaczmy jakie powyższe ma konsekwencje. Pokażemy, że w dużym wymiarze w zasadzie cała masa rozkładu normalnego koncentruje się w pasie o ustalonej szerokości wokół sfery o promieniu $\sqrt{D}$.

Niech [[przeliczyc rachunki bo cos nie tak]

$$C = \sqrt{D - a^2} \approx \sqrt{D} - \frac{a^2}{2\sqrt{D}}.$$

Policzmy więc (zakładamy, że a jest ustaloną stałą rzędu $O(1)$):

$$P(X : \|X\| \in [C - a, C + a]) = P(X : \|X\|^2 \in D - 2a\sqrt{D}, D + 2a\sqrt{D}])$$
$$= P(W \in [D - 2a\sqrt{D}, D + 2a\sqrt{D}]) \text{ dla } W \sim \chi^2_D.$$

Korzystając z przybliżenia rozkładem normlanym dostajemy

$$\approx P(Z \in [D - 2a\sqrt{D}, D + 2a\sqrt{D}]) \text{ dla } Z \sim N(D, D),$$

czyli

$$= P(Z \in [-2a, 2a]) \text{ dla } Z \sim N(0,1).$$

Korzystając z wcześniej wyliczonego faktu, że dla $r \geqslant 2$ mamy bardzo dokładne przybliżenie

$$P(Z \geqslant r) = \int_r^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx \frac{1}{\sqrt{2\pi} r} \exp(-r^2/2),$$

dostajemy dla $Z \sim N(0,1)$

$$P(Z \in [-2a, 2a]) = 1 - 2P(Z \geqslant 2a) \approx 1 - \frac{2}{\sqrt{2\pi} 2a} \exp(-(2a)^2/2) = 1 - \frac{1}{\sqrt{2\pi} a} \exp(-2a^2).$$

Konkludując dostajemy, że poza zbiorem (otoczenie sfery)

$$\{x : \|x\| \in [D - \frac{a^2}{2D} - a, D + \frac{a^2}{2D} + a]\}$$

znajduje się około

$$\frac{1}{\sqrt{2\pi}a} \exp(-2a^2)$$

danych.

A teraz pokażemy, że losowo wybrane punkty są do siebie prostopadłe.

5.5 Twierdzenie Cramera-Wolda

Rozdział 6

Rzutowania ortogonalne i PCA

6.1 Bazy i bazy ortonormalne

Przypominam, że ciąg wektorów $v = [v_1, \ldots, v_N]$ nazywamy bazą przestrzeni $\mathbb{R}^N$ jeżeli każdy punkt $x \in \mathbb{R}^N$ ma jednoznacznie wyznaczone współrzędne α_i w bazie v, czyli

$$x = \alpha_1 v_1 + \ldots + \alpha_N v_N.$$

Można pokazać, że v_i jest bazą, wtedy i tylko wtedy gdy v jest macierzą odwracalną. Stosując zapis macierzowy

$$\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}$$

dostajemy równanie

$$x = v\alpha,$$

czyli współrzędne x w bazie v są dane przez

$$\alpha = v^{-1}x.$$

Mówimy, że ciąg wektorów v_i jest ortogonalny, jeżeli każde dwa różne elementy są prostopadłe, czyli gdy

$$\langle v_i, v_j \rangle = v_i^T v_j = 0 \text{ dla } i \neq j.$$

Ciąg jest ortonormalny, jeżeli dodatkowo jej wektory są normalne, czyli mają długość 1. Baza ortogonalna, to baza która jest ortogonalna.

Stwierdzenie 6.1. *Ciąg $v = [v_1, \ldots, v_D]$ jest ortogonalny wtw. gdy*

$$v^T v = I_D.$$

Ogólnie każdą bazę możemy zortogonalizować (ortogonalizacja Grama-Schmidta), i każdy ciąg ortonormalny możemy rozszerzyć do bazy ortonormalnej.

Stwierdzenie 6.2. *Długość wektora wyraża się standardowym wzorem na współrzędnych w dowolnej bazie ortonormalnej (w konsekwencji też odległość między dwoma punktami można tak liczyć).*

6.2 Rzutowania ortogonalne

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

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

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

$$p_V(x) = \langle x, v_1 \rangle v_1 + \ldots + \langle x, v_k \rangle v_k. \tag{6.1}$$

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

$$p_V = VV^T.$$

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

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

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

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

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

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

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

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

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

Mamy

$$x = p_V(x) + p_{V^\perp}(x) \text{ dla } x \in \mathbb{R}^D. \tag{6.2}$$

W konsekwencji

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

Zadanie 6.1. *Korzystając z* (6.1) *proszę wyliczyć wzór na rzutowanie ortogonalne na*

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

Proszę podać macierz tego rzutowania.

Zadanie 6.2. *Korzystając z* (6.2) *proszę wyliczyć wzór na rzutowanie ortogonalne na*

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

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

6.3 Optymalne położenie środka

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

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

Obserwacja 6.1. *Niech będzie dana podprzestrzeń wektorowa V przestrzeni $\mathbb{R}^D$. Wtedy*

$$d^2(X, v + V) = \mathrm{SE}(p_{V^\perp}(X)) + |X| \cdot \|p_{V^\perp}(\mathrm{mean}X - v)\|^2 \tag{6.3}$$

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

Dowód. Oczywiście

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

Czyli na podstawie Obserwacji **??**

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

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

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

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

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

6.4 Sytuacja jednowymiarowa

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

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

Możemy oczywiście założyć, że v ma normę jeden, wtedy

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

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

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

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

Teraz minimalizacja

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

jest oczywiście równoważna maksymalizacji

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

Konkludując mamy problem znalezienia

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

Ponieważ

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

to

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

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

W konsekwencji sprowadziliśmy do problemu

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

Problem 6.1. *Mamy daną macierz symetryczną nieujemnie określoną Σ. Należy znaleźć*

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

Stwierdzenie 6.3. *Maksimum jest realizowane przez wektor własny odpowiadający największej wartości własnej Σ.*

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

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

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

Wtedy dla

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

mamy

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

Interesuje nas w takim razie szukanie maksimum

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

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

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

Czyli w konsekwencji maksimum jest osiągane dla

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

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

Uwaga 6.1 (interpretacja geometryczna). Załóżmy, że chcemy wyznaczyć kierunek najbardziej reprezentatywny dla zestawu danych X (zakładamy, że średnia jest zero).

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

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

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

i po przeliczeniu dostajemy, że powyższe odwzorowanie dane jest wzorem

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

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

6.5 Sytuacja wyżej-wymiarowa

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

Stwierdzenie 6.4. *Niech $X = (x_i)_{i=1..n}$ zbiór danych, $v = [v_1, \ldots, v_k]$ baza ortonormalna pewnej podprzestrzeni $V \subset \mathbb{R}^D$. Wtedy*

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

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

Interesuje nas wartość

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

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

Oczywiście

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

Z drugiej strony, mamy

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

czyli

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

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

Proszę zauważyć, że

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

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

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

Twierdzenie 6.1. *Rozpatrzmy wszystkie q-wymiarowe podprzestrzenie V o bazie ortonormalnej v w przestrzeni p-wymiarowej. Wtedy wartość*

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

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

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

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

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

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

Łatwo można sprawdzić, że

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

gdzie

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

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

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

czyli

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

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

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

W konsekwencji wylądowaliśmy na problemie maksymalizacji

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

Widać, że rozwiązanie jest maksymalne gdy

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

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

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

Stwierdzenie 6.5. *Mamy*

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

I wtedy ustalamy jaki procent wariancji chcemy mieć wyjaśniony.

Zadanie 6.3. *Mamy przestrzeń w* $\mathbb{R}^4$ *zadaną przez* $(1,1,0,0), (0,0,1,1)$ *i zbiór o kowariancji* I *i liczności* 100. *Proszę policzyć błąd popełniony przy rzutowaniu.*

Rozdział 7

Autoenkodery

7.1 Hipoteza rozmaitości

7.2 Sytuacja liniowa: autoenkodery=PCA

7.3 Sieci neuronowe

7.4 Metody gradientowe

7.5 Autoenkodery

7.6 Generatywne autoenkodery

Rozdział 8

Modele wariacyjne