Podstawy klasycznej (nienadzorowanej) analizy danych

J. Tabor

18 października 2017

Spis treści

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

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

I wyklad

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

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

- współczynnik korelacji - macierz kowariancji

- histogram - motywacja

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

PLANY: dokończyć histogram, gęstość normalizacja więcej-wymiarowa

Rozdział 1

Podstawowe charakterystyki danych

1.1 Dane skalarne

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

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

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

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

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

Średnia z próbki $X = (x_i)$ to

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

Średnia ma pewne minusy:

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

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

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

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

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

[1] Szpinak - cytowanie

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

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

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

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

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

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

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

Pokażemy najprostsze własności wariancji.

Obserwacja 1.1. *Mamy*

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

Dowód. Mamy

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

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

□

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

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

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

Wtedy

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

oraz

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

Dowód. Mamy z poprzedniej obserwacji oraz (1.1)

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

□

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

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

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

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

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

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

1.2 Funkcja kosztu

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

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

Dwa pytania:

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

Błąd można rozumieć na wiele sposobów. Dla przykładu możemy rozpatrzeć

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

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

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

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

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

Łatwo widać, że

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

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

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

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

Błąd dany przez sumę modułów. Rozważmy błąd dany przez:

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

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

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

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

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

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

co daje tezę. □

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

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

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

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

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

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

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

1.3 Histogram

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

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

1.4 Współczynnik korelacji

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

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

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

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

Czyli czy

$$y = a_1 x + b \text{ lub } x = a_2 y + b \text{ gdzie } x = (x_1, \ldots, x_N), y = (y_1, \ldots, y_N) \in \mathbb{R}^N. \tag{1.2}$$

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

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

Łatwo wtedy sprawdzić, że (1.2) jest równoważne stwierdzeniu, że wektory v_1, v_2 są współliniowe:

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

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

Dygresja 1.2. *Załóżmy, że mamy dwa wektory $x, y \in \mathbb{R}^N$. Chcemy umieć sprawdzić, czy są one współliniowe, czyli czy istnieje α_1 bądź (równoważnie) α_2 takie*

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

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

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

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

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

Tak więc sprawdzenie zależności liniowej sprowadza się do wyliczenia

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

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

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

Zadanie 1.8. *Zakładamy, że x, y spełniają* (1.2)*. Kładziemy*

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

Pokaż, że

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

Zadanie 1.9. *Proszę wyliczyć korelacje między pierwwszą i drugą współrzędną dla a) $X = \{(l, 2l+1)\}$, $l = 1..10$, b) $X = \{(\cos(2\pi k/6), \sin(2\pi k/6)) : k = 0..5\}$.*

1.5 Dane wektorowe

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

Najprostszym i najczęściej stosowanym jest oczywiście średnia:

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

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

$$\mathrm{mean}X = (\mathrm{mean}X_1, \ldots, \mathrm{mean}X_D). \tag{1.3}$$

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

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

Jest to bezpośredni wniosek z następującej oberwacji.

Obserwacja 1.3. *Własność:*

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

wynika z bezpośredniego rozpisania wzorów:

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

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

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

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

Po odjęciu otrzymujemy to co chcieliśmy.

Dygresja 1.3. *Gdybyśmy chcieli zdefiniować analog wielowymiarowy mediany, należałoby rozważyć minima funkcji*

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

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

Macierz kowariancji definuje się biorąc kowariancje każdych współrzędnych:

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

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

Obserwacja 1.4. *Mamy*

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

Dowód. Niech

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

Wtedy

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

Oczywiście

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

czyli

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

Na podstawie (1.3) mamy $(\mathrm{mean}X)^l = \mathrm{mean}(X^l)$, co daje tezę obserwacji. □

Zadanie 1.10. *Proszą pokazać, że dla liniowego A mamy*

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

Zadanie 1.11. *Proszą pokazać, że dla liniowego A mamy*

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

Zadanie 1.12. *Proszę pokazać, że*

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

Wsk.: proszę powtórzyć rozumowanie dla skalarów.

Zadanie 1.13. *Niech $X = (x_1, \ldots, x_N)$, $Y = (y_1, \ldots, y_K)$. Znamy średnie i macierze kowariancji X oraz Y. Proszę wyliczyć średnią i macierz kowariancji dla $X \cup Y = (x_1, \ldots, x_N, y_1, \ldots, y_K)$.*

Wsk.: proszę powtórzyć rozumowanie dla skalarów.

1.6 Macierze symetryczne, nieujemnie i dodatnio określone

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

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

$$Av = \lambda v.$$

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

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

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

Mówimy, że macierz A jest symetryczna, jeżeli

$$A = A^T.$$

FAKT. Niech A będzie macierzą symetryczną. Wtedy

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

Przypominam, że wektory (v_i) są ortonormalne, jeżeli

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

co jest równoważne temu, że

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

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

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

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

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

Ważną klasę macierzy symetrycznych tworzą macierze nieujemnie i dodatnio określone:

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

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

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

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

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

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

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

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

Algorytm 1.1. Algorytm liczenia pierwiastka z dodatnio określonej macierzy symetrycznej A:

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

2, Końcowy wzór:

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

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

Stwierdzenie 1.1. *Niech $\Sigma = \mathrm{cov}X$ dla $X \subset \mathbb{R}^D$. Wtedy:*

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

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

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

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

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

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

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

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

Także

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

gdyż $\langle x, v\rangle = x^T v$.

Punkty 2 i 3 pozostawiam bez dowodu. □

Zadanie 1.14. *Proszę policzyć pierwiastek z macierzy*

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

Zadanie 1.15. *Niech A będzie macierzą symetryczną, a V bazą ortonormalną składającą się z wektorów własnych A.*

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

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

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

1.7 Whitening, odległość Mahalanobisa

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

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

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

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

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

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

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

Problem 1.2. *Mamy dane $X \subset \mathbb{R}^D$. Czy (i ew. kiedy) można znaleźć taką macierz odwracalną A, że*

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

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

Problem 1.3. *Niech $\Sigma = \mathrm{cov} X$ dla pewnego zbioru danych X. Czy (i ew. kiedy) można znaleźć taką macierz odwracalną A, że*

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

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

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

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

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

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

Konkludując, whitening dany jest przez

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

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

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

Wprowadźmy teraz oznaczenie na normę Mahalanobisa:

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

Wtedy mamy

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

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

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

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

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

Zadanie 1.16. *Niech*

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

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

Zadanie 1.17. *Mamy*

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

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

Zadanie 1.18. *Metryka Mahalanobisa jest niezmiennicza na transformacje afiniczne w następującym sensie*

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

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

Rozdział 2

k-means

k-means jest tak naprawdę metodą kompresji/dyskretyzacji. Ale jest używany do klastrowania.

2.1 Zafiksowane centra $v_1, \ldots, v_k$

Problem 2.1. *Postawienie problemu: Mamy dany zestaw możliwych punktów których używamy do dyskretyzacji (kompresji)* $V = \{v_1, \ldots, v_k\} \subset \mathbb{R}^N$.

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

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

tak by zminimalizować całkowity (kwadratowy) błąd popełniony przy dyskretyzacji

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

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

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

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

Obserwacja 2.1. *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`

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

Problem 2.2. *Postawienie problemu: Mamy daną funkcję indeksującą j. Chcemy znaleźć, dla zestawu danych $X \subset \mathbb{R}^N$, zestaw możliwych punktów których używamy do dyskretyzacji (kompresji) $V = \{v_1, \ldots, v_k\} \subset \mathbb{R}^N$ tak by zminimalizować całkowity (kwadratowy) błąd popełniony przy dyskretyzacji*

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

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

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

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

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

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

Ale my już wiemy jakie jest rozwiązanie! Po prostu

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

2.3 Ogólny problem

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

Problem 2.3. *Chcemy znaleźć, dla zestawu danych $X \subset \mathbb{R}^N$, zestaw możliwych punktów których używamy do dyskretyzacji (kompresji) $V = \{v_1, \ldots, v_k\} \subset \mathbb{R}^N$ oraz funkcję indeksującą j tak by zminimalizować całkowity (kwadratowy) błąd popełniony przy dyskretyzacji*

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

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

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

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

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

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

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

Inicjalizacja początkowych punktów:

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

k-means++ algorytm:

1. Choose one center uniformly at random from among the data points. For each data point x, compute $D(x)$, the distance between x and the nearest center that has already been chosen.

2. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D^2(x)$.

3. Repeat Steps 2 and 3 until k centers have been chosen.

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

Zadanie 2.1. *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

2.4 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 2.2.

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

Dowód. Dowód wynika bezpośrednio z Obserwacji 1.3 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

$$\begin{aligned}
\mathrm{SE}(X_1 \cup X_2, m_{X_1 \cup X_2}) &= \mathrm{SE}(X_1, m_{X_1 \cup X_2}) + \mathrm{SE}(X_2, m_{X_1 \cup X_2}) \\
&= \mathrm{SE}(X_1) + |X_1| \cdot \|m_{X_1 \cup X_2} - m_{X_1}\|^2 + \mathrm{SE}(X_2) + |X_2| \cdot \|m_{X_1 \cup X_2} - m_{X_2}\|^2 \\
&= \mathrm{SE}(X_1) + \mathrm{SE}(X_2) + |X_1| \cdot (\tfrac{|X_2|}{|X_1|+|X_2|})^2 \|m_{X_1} - m_{X_2}\|^2 + |X_2| \cdot (\tfrac{|X_1|}{|X_1|+|X_2|})^2 \|m_{X_1} - m_{X_2}\|^2 \\
&= \mathrm{SE}(X_1) + \mathrm{SE}(X_2) + \tfrac{|X_1||X_2|}{|X_1|+|X_2|} \|m_{X_1} - m_{X_2}\|^2.
\end{aligned}$$

□

Bezpośrednio z powyższej obserwacji dostajemy:

Obserwacja 2.3. *1. Mamy*

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

2. Oraz

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

Wniosek 2.1. *Jeżeli mamy klastry X_i i X_j, oraz punkt $x \in X_i$, to będzie się nam opłacało go przełączyć do j wtw gdy:*

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

czyli na podstawie powyższego gdy:

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

czyli gdy

$$\frac{|X_j|}{|X_j|+1}\|x - m_j\|^2 < \frac{|X_i|}{|X_i|-1}\|x - m_i\|^2. \tag{2.1}$$

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 $s_{\sigma(l)} \mathrel{+}= x_l$, $N_{\sigma(l)}{}^{++}$. 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 (2.1), i wtedy updatujemy odpowiednio średnie i liczności klastrów.