Ozon Masters ML1 Арехин Станислав Олегович

Теор. задание 1

Задание №1

$\boxed{\text{стр 1}}$

<u>Дано:</u> n объектов равномерно распред. по d-мерному единич. шару.

<u>Найти:</u> выражение для медианы от начала координат до ближайшего соседа.

<u>Решение:</u>

1) Объём d-мерного шара $V_d(\rho) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}\rho^d \sim \rho^d$.
По условию задачи шар единичный $\Rightarrow$
$V_d(1) = 1^d = 1$. Обозначим шар как B_1

2) Впишем в шар B_1 новый шар радиусом $r<1$ с центром в ~~то~~ начале координат. Объем этого шара $V_{B_r} = r^d$. Обозначим B_r.
Пусть событие A – объект не попал в шар B_r,
$\bar{A}$ – объект попал в шар B_r. $P(\bar{A}) = \frac{V_{B_r}}{V_{B_1}} = \frac{r^d}{1} = r^d$

$\Rightarrow P(A) = 1 - P(\bar{A}) = 1 - r^d$

3) Попадание объектов в область – события независимые, кроме того, они равновозможные.
Пусть ~~В~~ событие B – все n объектов не попали в шар B_r. Тогда $P(B) = P(\prod_{i=1}^{n} A_i) =$
$= \prod_{i=1}^{n} P(A_i) = P(A)^n = (1-r^d)^n$

4) Пусть шар B_r, выражающий медианну расстояния до ближайшего объекта от начала координат. Следовательно, вер-ть того, что все точки не попали в шар B_r должна быть равна 0,5, то есть

$P(B) = 0{,}5$

см. след. лист

№1 продолжение

$P(B)=0{,}5 \Leftrightarrow (1-r^d)^n = 0{,}5 \Leftrightarrow 1-r^d = 0{,}5^{1/n} \Leftrightarrow$
$\Leftrightarrow r^d = 1-0{,}5^{1/n} \Leftrightarrow r = (1-0{,}5^{1/n})^{1/d}$, где

r — медиана расстояния до ближайшего ~~...~~ объекта от ~~...~~ начала координат, n — объём выборки, d — размерность пр-ва.

5) Проинтерпретируем в терминах применимости метода ближайшего соседа.

1. Пусть $d \to \infty$. Тогда $\lim\limits_{d\to\infty} (1-0{,}5^{1/n})^{1/d} = 1$.

То есть медиана расстояния от начала координат до ближайшего соседа $\to 1$, что означает, что все объекты будут равноудалены от начала координат (расположены на границе шара), проще говоря все объекты ближайшие. ~~Это~~ Из этого можно сделать вывод о неприменимости метода ближайшего соседа для пр-в большой размерности.

d фиксировано, например 3.

2. Пусть $n \to \infty$, d фиксировано, например 3.
Тогда $\lim\limits_{n\to\infty} (1-0{,}5^{1/n})^{1/3} = 0$. В этом случае медиана расстояния от начала координат до ближайшего объекта $\to 0$ для $\forall d$. Это значит, что для выборок очень большого объема невозможно применить метод ближайшего соседа, т.к. ближайшего объекта нет.

d фиксировано, например 3

3. Пусть $n \to 0$, d фиксировано, например 3
Тогда $\lim\limits_{n\to 0} (1-0{,}5^{1/n})^{1/3} = 1$. То есть см. следующий из

№1 продолжение.

То есть, если объектов выборки очень мало, медиана → 1, то есть все объекты будут ближайшими и нельзя применить метод ближайшего соседа.

(4.) Итого, для применения метода ближайшего соседа необходимо иметь выборку оптимального объема и размерности. Если это не учитывать, то метод работает плохо.

Ozon Masters ML1 — Теор. задание 1

Арешин Станислав Олегович

стр1

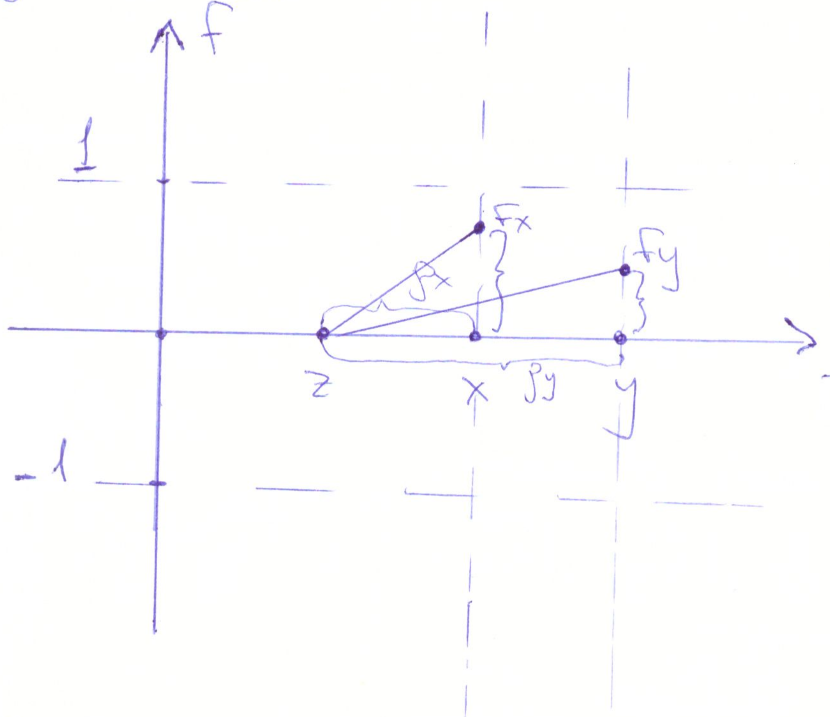
Задача №2

Дано: объекты x, y, z; $\rho(z,x)=\rho_x$; $\rho(z,y)=\rho_y$; $\rho_x < \rho_y$. Добавляется новый признак, который $\sim R(-1;1)$

Найти: Вер-ть того, что ближайшим объектом к z станет не x а y.

Решение

1) Пусть объекты x, y, z изначально находятся в пр-ве $\mathbb{R}^1$. При добавлении нового признака пр-во становится $\mathbb{R}^2$. Изобразим на графике: Новый признак обозначим f. Соответственно для x будет f_x, для y будет f_y:



Тогда в новом пр-ве:

$$\rho(z,x) = \rho_{zx} = \sqrt{\rho_x^2 + f_x^2}$$

И $$\rho(z,y) = \rho_{zy} = \sqrt{\rho_y^2 + f_y^2}$$

То есть:

$$\rho_{zx}^2 = \rho_x^2 + f_x^2$$

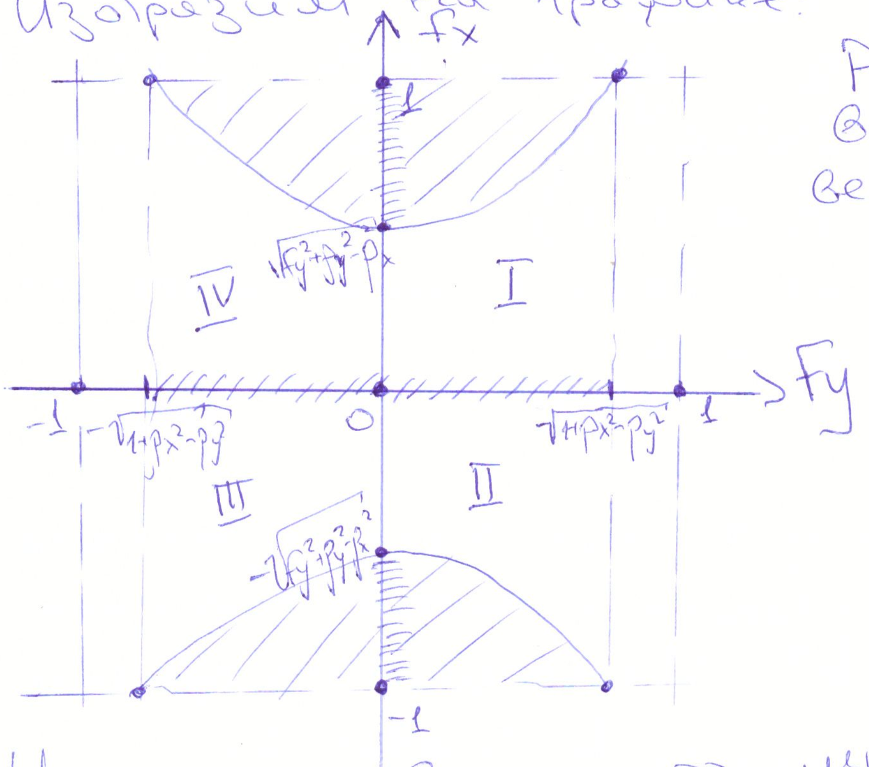
$$\rho_{zy}^2 = \rho_y^2 + f_y^2$$

Вероятность того, что ближайшим к z станет y, а не x $P(\rho_{zx}^2 - \rho_{zy}^2 > 0) = P(\rho_{zy}^2 - \rho_{zx}^2 < 0) = P(\rho_y^2 + f_y^2 < \rho_x^2 + f_x^2) = P(f_x^2 > f_y^2 + \rho_y^2 - \rho_x^2)$. Необходимо вычислить эту вер-ть.

см. след. лист №2 ⇓

2) Случайный вектор $(f_y; f_x)$ равномерно распределён на квадрате $[-1;1]\times[-1;1]$

Изобразим на графике:



Рассчитаем область, в которую попадает вер-ть $P(f_x^2 > f_y^2+\rho_y^2-\rho_x^2)$

$$f_x^2 = f_y^2+\rho_y^2-\rho_x^2$$

$$f_x=1 \Rightarrow f_y=\pm\sqrt{1+\rho_x^2-\rho_y^2}$$

$$f_x=-1 \Rightarrow f_y=\pm\sqrt{1+\rho_x^2-\rho_y^2}$$

$$f_x=\pm\sqrt{f_y^2+\rho_y^2-\rho_x^2}$$

На графике видно, что мы имеем 4 симметричные области I; II; III; IV. Рассмотрим область I. Вер-ть попасть в квадрат $[1;1]\times[1;1]$ равна 1, площадь $S=2\cdot 2=4$. Вер-ть попасть в область I $[0;1]\times[0;1]$ равна отношению площадей: $\frac{S_I}{S}=\frac{1}{4}$. При этом в области I f_y изменяется на $[0;\sqrt{1+\rho_x^2-\rho_y^2}]$, f_x изменяется на $[\sqrt{f_y^2+\rho_y^2-\rho_x^2}; 1]$

Найдём площадь данной области:

$$S_{area} = \int_0^{\sqrt{1+\rho_x^2-\rho_y^2}} \int_{\sqrt{f_y^2+\rho_y^2-\rho_x^2}}^{1} \frac{1}{4}\, df_y\, df_x = \left[\text{Пусть } d=\rho_y^2-\rho_x^2\right] =$$

$$= \int_0^{\sqrt{1-d}} \int_{\sqrt{f_y^2+d}}^{1} \frac{1}{4}\, df_y\, df_x = \frac{1}{4}\int_0^{\sqrt{1-d}} 1-\sqrt{f_y^2+d}\, df_y =$$

$$= \frac{1}{4}\cdot\frac{1}{2}\left(\sqrt{1-d} - \operatorname{arsinh}\left(\frac{\sqrt{1-d}}{\sqrt{d}}\right)d\right)$$

см. след лист №3

3) Таких областей четыре =>

$$P(\rho_x^2 > \rho_y^2 + \rho_y^2 - \rho_x^2) = 4 \cdot S_{area} = \frac{1}{2}\left(\sqrt{1-d} - \operatorname{arsinh}\left(\frac{\sqrt{1-d}}{\sqrt{d}}\right) d\right),$$

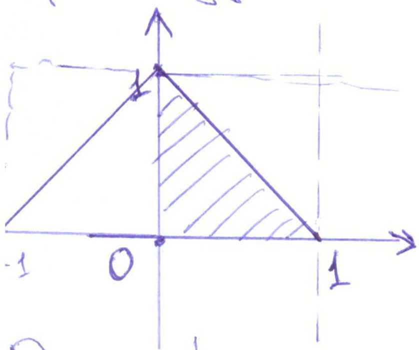
где $d = \rho_y^2 - \rho_x^2$

Таким образом, вероятность того, что объект y станет ближайшим для z вместо x равна:

$$\frac{1}{2}\left(\sqrt{1+\rho_x^2-\rho_y^2} - \operatorname{arsinh}\left(\frac{\sqrt{1+\rho_x^2-\rho_y^2}}{\sqrt{\rho_y^2-\rho_x^2}}\right)(\rho_y^2-\rho_x^2)\right)$$

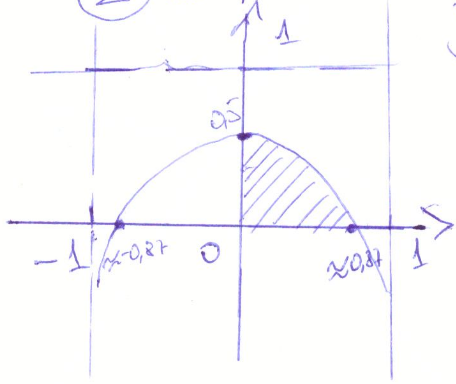
4) Проинтерпретируем полученный результат. Для это вернёмся к обозначению $d = \rho_y^2 - \rho_x^2$ и построим несколько графиков при различном d:

① $d = 0$, в этом случае $\rho_x^2 = \rho_y^2$, объекты равноудалены от z:



При $d = 0$
$P = \frac{1}{2} = 0,5$

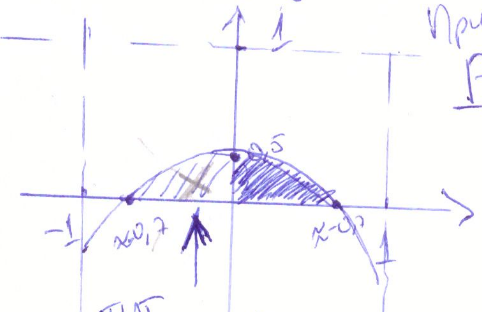
② $d = 0,25$ $\rho_y^2 = \rho_x^2 + 0,25$



При $d = 0,25$
$P \approx 0,268$

③ $d = 0,5$, $\rho_y^2 = \rho_x^2 + 0,5$

При $d = 0,5$
$P \approx 0,13$



тут случайно закрасил.

Таким образом, видно, что чем ближе между собой объекты x и y, тем больше вер-ть того, что при добавлении нового признака y станет ближе к z, чем x.
То есть при $d \to 0$ $P \to \frac{1}{2}$,
при $d \to \infty$, $P \to 0$.

см. след лист № 4

№2 продолжение

В терминах применимости метода ближайших соседей результат можно интерпретировать следующим образом:
Если в пр-ве расстояние между объектами мало (например, в пр-вах большой размерности), то вероятность того, что при добавлении некоторого нового шумового признака объект y ~~нач~~ станет ближе к z, чем x стремится к 0,5. Из этого можно сделать вывод, что алгоритм ближайших соседей неустойчив к шуму.

Ozon Masters ML1
Теор. задание 1

Арешин Станислав
Олегович

Задача №3

$\boxed{\text{стр 1}}$

Дано: $\vec{a}, \vec{b}, \angle(\vec{a}, \vec{b}) = \Theta$, случ. равномерно выбирается прямая l, проходящая через начало координат

Найти: 1. $P(\vec{a}, \vec{b}$ в одном полупр-ве отн. $l)$
2. $P(\vec{a}, \vec{b}$ в разных полупр-вах отн. $l)$

Решение 1) Запишем формулу скалярного произв. векторов $(\vec{a}, \vec{b}) = |\vec{a}||\vec{b}| \cos\alpha$, где
$\alpha = \angle(\vec{a}, \vec{b}) \in [0; \pi]$
$(\vec{a}, \vec{b}) \geq 0$, если $\cos\alpha \geq 0$, т.е. $\alpha \in [0; \frac{\pi}{2}]$
$(\vec{a}, \vec{b}) < 0$, если $\cos\alpha < 0$, т.е. $\alpha \in (\frac{\pi}{2}; \pi]$

Для произвольного вектора $\vec{c}$:

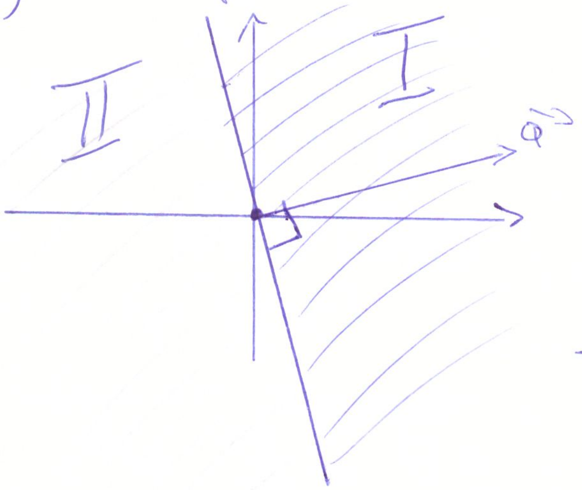
$\vec{a}, \vec{b}$ в одном полупр-ве отн $\vec{c}$, если:

$$\left[\begin{array}{l} \begin{cases} (\vec{a}, \vec{c}) \geq 0 \\ (\vec{b}, \vec{c}) \geq 0 \end{cases} \\ \begin{cases} (\vec{a}, \vec{c}) < 0 \\ (\vec{b}, \vec{c}) < 0 \end{cases} \end{array}\right.$$

$\vec{a}, \vec{b}$ в разных полупр-вах отн. $\vec{c}$, если:

$$\left[\begin{array}{l} \begin{cases} (\vec{a}, \vec{c}) \geq 0 \\ (\vec{b}, \vec{c}) < 0 \end{cases} \\ \begin{cases} (\vec{a}, \vec{c}) < 0 \\ (\vec{b}, \vec{c}) \geq 0 \end{cases} \end{array}\right.$$

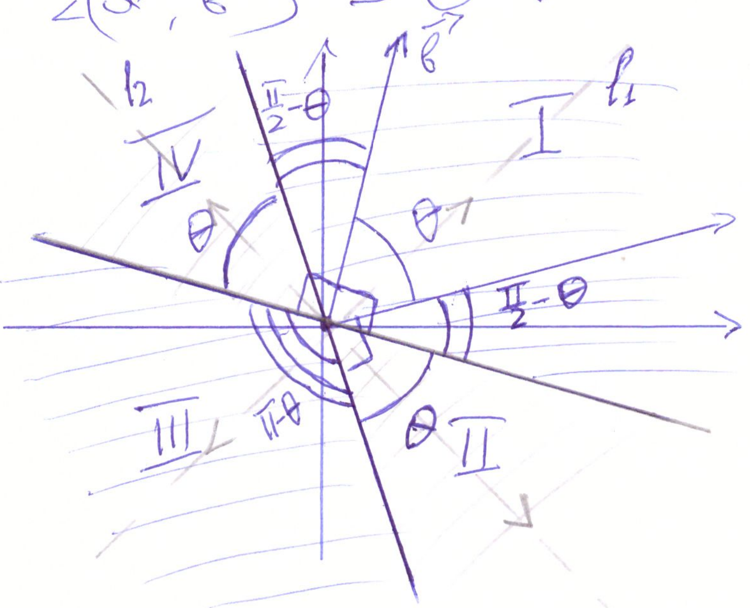
2) Построим вектор $\vec{a}$:



Вектор $\vec{a}$ делит плоскость на две полуплоскости I и II.
для $\forall \vec{c} \in I \Rightarrow \alpha = \angle(\vec{a}, \vec{c}) \in [0; \frac{\pi}{2}]$
$\Rightarrow \cos\alpha \geq 0 \Rightarrow (\vec{a}, \vec{c}) \geq 0$
для $\forall \vec{c} \in II \Rightarrow \alpha = \angle(\vec{a}, \vec{c}) \in (\frac{\pi}{2}; \pi]$
$\Rightarrow \cos\alpha < 0 \Rightarrow (\vec{a}, \vec{c}) < 0$

см. след лист №2

2) Теперь построим вектор $\vec{b}$ так, что $\angle(\vec{a}, \vec{b}) = \theta$:



Вектор $\vec{b}$ в свою очередь аналогично делит плоскость на две полуплоскости.
В результате вектора $\vec{a}, \vec{b}$ делят плоскость на 4 части I, II, III, IV. Вычислим углы.

3.) Теперь случайно проводим прямую l.

Если $l \in$ I, III, то
$$\left[\begin{array}{l} \begin{cases} (\vec{a}, \vec{l}) \geq 0 \\ (\vec{b}, \vec{l}) \geq 0 \end{cases} \\ \begin{cases} (\vec{a}, \vec{l}) < 0 \\ (\vec{b}, \vec{l}) < 0 \end{cases} \end{array}\right.$$
В зависимости от направляющего вектора $\vec{l}$, который мы выберем.

То есть в этом случае $\vec{a}, \vec{b}$ лежат в одном полупл-ве отн. l.

Если $l \in$ II, IV, то
$$\left[\begin{array}{l} (\vec{a}, \vec{l}) \geq 0 \\ (\vec{b}, \vec{l}) < 0 \\ (\vec{a}, \vec{l}) < 0 \\ (\vec{b}, \vec{l}) \geq 0 \end{array}\right.$$
В завис. от выбранного направляющего вектора l.

В этом случае $\vec{a}, \vec{b}$ лежат в разных полупл-вах относительно l.

4) Вычислим вероятности. Одна точка прямой – это начало координат. Весь угол $= 2\pi$.
$\Rightarrow$ ~~P(~~ вероятность попадания второй точки прямой в области II, IV $= \frac{2\theta}{2\pi} = \frac{\theta}{\pi} \Rightarrow$

$P(l \in \text{II}, \text{IV}) = \frac{\theta}{\pi} \Rightarrow P(l \in \text{I}, \text{III}) = 1 - \frac{\theta}{\pi}$

Таким образом, 1. $P(\vec{a}, \vec{b}$ в одном полупл-ве отн. $l) = 1 - \frac{\theta}{\pi}$; 2. $P(\vec{a}, \vec{b}$ в разных полупл-вах отн $l) = \frac{\theta}{\pi}$

см. след. лист №3

5) Проинтерпретируем результат с точки зрения использования семейства $F=\{f_w(x)=\text{sign}((w,x))\}$ в качестве LSH для косинусного расстояния.

В условиях данной задачи $F=\{f_w(x)=\text{sign}((l,x))\}$

То есть $hash(\vec{x}) = \text{sign}((l,\vec{x}))$

$P(l \in I, III) = P(\vec{a},\vec{b}$ лежат в одном полупр-ве отн $l) = P(hash(\vec{a})=hash(\vec{b})) = 1-\frac{\theta}{\pi}$

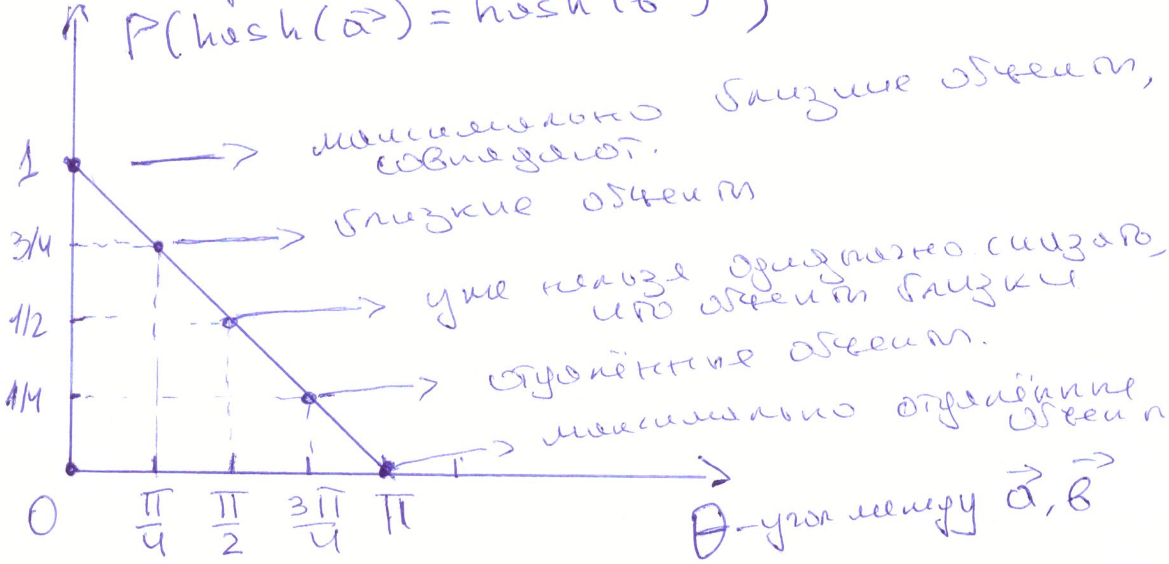
$\theta \in [0;\pi]$

Пусть $\theta = 0 \Rightarrow P(hash(\vec{a})=hash(\vec{b}))=1$

При $\theta = \frac{\pi}{2} \Rightarrow P(hash(\vec{a})=hash(\vec{b}))=\frac{1}{2}$

При $\theta = \pi \Rightarrow P(hash(\vec{a})=hash(\vec{b}))=0$

Построим график:



На графике видно, что чем больше угол между векторами $\vec{a}$ и $\vec{b}$, тем меньше вероятность того, что они окажутся в одном полупр-ве относительно случайной прямой. В этом идея метода приближённого поиска ближайших соседей LSH - с большей вер-тью присваивать одинаковые значения хеша близким объектам, разные - отдалённым

Ozon Masters ML 1

Теор. задание 1

Аремин Станислав Олегович

Задача №4

Дано: $A^TAv = \sigma^2 v$, v - собств. ~~зн~~ вектор A^TA, σ^2 - собств. знач. A^TA

Доказать: $\frac{1}{\sigma}Av$ - собств. знач. AA^T

Док-во: 1) Выполним SVD разложение A:

$A = U\Lambda V^T$ (1), где U, V - ортогональные матрицы, Λ - матрица собств. знач. A: $\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ \vdots & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_r \end{pmatrix}$

2) Тогда $A^T = (U\Lambda V^T)^T = V\Lambda^T U^T$ (2)

Таким образом:

(1), (2)

$U^TU = UU^T = E$

$V^TV = VV^T = E$

$\Lambda = \Lambda^T$

$\Rightarrow$ $A^TA = V\Lambda^T \cancel{U^T \cdot U}\Lambda V^T = V\Lambda^2 V^T$ (3)

$AA^T = U\Lambda \cancel{V^T \cdot V}\Lambda U^T = U\Lambda^2 U^T$ (4)

3) Домножим (3) на V справа, (4) на U справа

$A^TAV = [V\Lambda^2 \cancel{V^TV}] = V\Lambda^2$ (5)

$AA^TU = [U\Lambda^2 \cancel{U^TU}] = U\Lambda^2$ (6)

То есть $A^TAV = V\Lambda^2$ (5); $AA^TU = U\Lambda^2$ (6)

Из (5) следует, что V - матрица собств. векторов A^TA, Λ^2 - матрица собств. знач. A^TA; из (6) следует, что U - матрица собств. векторов AA^T, Λ^2 - матрица собств. знач AA^T

4) Матрица V содержит собств. вектор v, соответствующее собств. знач. которого $\lambda_v^2 = \sigma^2$, откуда выразим $\lambda_i = \sigma_i$, где σ_i^2 - собств. знач. A^TA, λ_i - собств. знач. A.

5) Вернёмся к SVD разложению (1) и домножим на V справа: $AV = U\Lambda$, откуда выразим матрицу ~~вектор~~ собств. знач. AA^T U: $U = \frac{1}{\Lambda}AV$ (7)

Перезапишем (7) в виде: $u_i = \frac{1}{\lambda_i}Av_i \Rightarrow$ $u = \frac{1}{\sigma}Av$ - собственное значение матрицы AA^T

∎

Озон Masters МЛ 1 Аршин Станислав
Теор. задание 1 Олегович.

Задача №5 $\boxed{\text{Стр 1}}$

Дано: $X_1, \dots X_n \in \mathbb{R}$ - выручка маркетплейса Озон в каждый из N дней.

Найти: 1. Формулу оценки выручки на след. день.
2. Формулу оценки дисперсии выручки.

Решение

1) Пусть X_i - выручка в ~~~~ i-й день,
Y_j - выручка с j-ой продажи в этот день.
Предположим, что за i-й день было совершено m продаж, при этом m достаточно велико.
Выручка за один день формируется из суммы выручек с каждой продажи за этот день.
Тогда за i-й день выручка $X_i = \sum_{j=1}^{m} Y_j$
$Y_1, \dots, Y_m$ - независимые случайные величины.
Предположим, что они одинаково распределены и имеют конечные мат. ожидания и дисперсию,
$M_{Y_j} < \infty$, $\sigma^2_{Y_j} < \infty$.
Воспользуемся центральной предельной теоремой: Сумма независ., одинаково распред. случ. величин с конечными мат. ожиданиями и дисперсиями: $X_i = \sum_{j=1}^{m} Y_j$

$$\frac{X_i - M[X_i]}{\sqrt{D[X_i]}} \xrightarrow[m \to \infty]{d} Z \sim N(0,1).$$

При достаточно больших m можно считать, что $X_i \sim N(M[X_i], D[X_i])$.

См. след. лист №2

То есть выручка в каждый из N дней $X_i \sim N(M[X_i], D[X_i])$ для $X_1, \dots X_n$.

2) Проблема в том, что в данном случае каждая величина X_i будет распределена нормально с разными мат. ожиданиями и дисперсиями.

Предположим, что распределение выручки с одной продажи Y_j одинаково для $\forall X_i \in X_1, \dots X_n$.

Пусть в среднем в каждый из N дней совершается n продаж.

То есть $X_i = \sum_{j=1}^{n} Y_j$ для $\forall X_i \in X_1, \dots X_n$

Тогда $M[X_i] = M[\sum_{j=1}^{n} Y_j] = \sum_{j=1}^{n} M[Y_j]$ для $\forall X_i \in X_1, \dots X_n$

$D[X_i] = D[\sum_{j=1}^{n} Y_j] = \sum_{j=1}^{n} D[Y_j]$, т.к Y_j - независ. для $\forall X_i \in X_1, \dots X_n$

Итого получаем, что:

$X_1, \dots, X_n \sim N\left(\sum_{j=1}^{n} M[Y_j], \sum_{j=1}^{n} D[Y_j]\right)$

3) $M[Y_j]$ и $D[Y_j]$ неизвестны, распределения Y_j - неизвестно. Можно оценить параметры нормального распределения $X_1, \dots, X_n$ с помощью ММП. Пусть $\theta_1 = \sum_{j=1}^{n} M[Y_j]$

$\theta_2^2 = \sum_{j=1}^{n} D[Y_j]$

Тогда $X_1, \dots X_n \sim N(\theta_1, \theta_2^2)$

$p(x, \theta) = \frac{1}{\theta_2 \sqrt{2\pi}} e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}}$ - плотность вер-ти

$L(X_1, \dots, X_n, \theta_1, \theta_2) = \prod_{i=1}^{n} p(X_i, \theta_1, \theta_2)$ - функция правдоподобия

см. след лист №3

№5 (продолжение) $\boxed{\text{стр 3}}$

$$L(x_1,\dots,x_n,\theta_1,\theta_2)=\prod_{i=1}^{n}\frac{1}{\theta_2\sqrt{2\pi}}\,e^{-\frac{(x_i-\theta_1)^2}{2\theta_2^2}}$$

$$\ln L(x_1,\dots,x_n,\theta_1,\theta_2)=\sum_{i=1}^{n}\left(\ln\left(\frac{1}{\theta_2\sqrt{2\pi}}\right)-\frac{(x_i-\theta_1)^2}{2\theta_2^2}\right)=$$

$$=-n\ln\left(\sqrt{2\pi}\right)-n\ln(\theta_2)-\sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{2\theta_2^2}$$

(1) $$\frac{\partial L}{\partial\theta_1}=\cancel{-}\sum_{i=1}^{n}\cancel{-}\frac{\cancel{2}(x_i-\theta_1)}{\cancel{2}\theta_2}=\sum_{i=1}^{n}\frac{x_i-\theta_1}{\theta_2}=0$$

$$\sum_{i=1}^{n}(x_i-\theta_1)=0 \iff n\theta_1=\sum_{i=1}^{n}x_i \Rightarrow$$

$$\Rightarrow \hat{\theta}_1=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{X}_n$$ — выборочное среднее.

(2) $$\frac{\partial L}{\partial\theta_2}=-\frac{n}{\theta_2}+\sum_{i=1}^{n}\frac{\cancel{2\theta_2}(x_i-\theta_1)^2}{\cancel{2}\theta_2^3}=-\frac{n}{\theta_2}+\sum_{i=1}^{n}\frac{(x_i-\theta_1)^2}{\theta_2^3}=0$$

$$\frac{1}{\cancel{\theta_2^3}}\sum_{i=1}^{n}(x_i-\theta_1)^2=\frac{n}{\theta_2}\ \Big|\cdot\frac{\theta_2^3}{n}$$

$$\hat{\theta}_2^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X}_n)^2$$

4) Таким образом, значение выручки предсказать на следующий день можно из распределения

$X_1,\dots,X_n \sim N\left(\overline{X}_n,\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X}_n)^2\right)$, где

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Оценка дисперсии выручки будет дисперсия распределения, т.е.

$$D=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X}_n)^2.$$