

Машинное обучение

Антон Андрейцев

Содержание

0.1	Регрессия	3
0.1.1	Задача 1	3
0.1.2	Задача 2	3
0.1.3	Задача 3	4
0.1.4	Задача 4	4
0.2	Perceptron	5
0.2.1	Задача 1	5
0.2.2	Задача 2	9
0.2.3	Задача 3	11
0.3	SVM	13
0.3.1	Задача 1	16
0.3.2	Задача 2	17
0.3.3	Задача 3	18
0.3.4	Задача 4	20
0.4	Naive Bayes	23
0.4.1	Задача 1	23
0.4.2	Задача 2	24
0.5	Логистическая регрессия	26
0.5.1	Задача 1	26
0.5.2	Задача 2	27
0.5.3	Задача 3	29
0.6	Метрики качества	31
0.6.1	Задача 1	31
0.6.2	Задача 2	33
0.7	PCA	34
0.7.1	Задача 1	34
0.7.2	Задача 2	36
0.7.3	Задача 3 (Вероятностная постановка PCA)	39
0.8	ЕМ алгоритм	40
0.8.1	Задача 1	40
0.8.2	Задача 2	42

0.1 Регрессия

0.1.1 Задача 1

Показать, что минимизация суммы квадратов остатков для линейной регрессии ($\|X\beta - y\|_2^2 \rightarrow \min_{\beta}$) эквивалентна максимизации правдоподобия в модели: $y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 \cdot I_n)$

Решение

$$y|X \sim \mathcal{N}(X\beta, \sigma^2 \cdot I_n) \Rightarrow \text{Likelihood} = p(y|X) \max_{\beta}$$

$$p(y|X) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \cdot \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\} \rightarrow \max_{\beta}$$

$$\log L = -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \rightarrow \max_{\beta}$$

Первые два члена не зависят от $\beta \Rightarrow$
 $\Rightarrow \log L \rightarrow \max_{\beta} \sim (y - X\beta)^T(y - X\beta) \rightarrow \min_{\beta}$

$$(y - X\beta)^T(y - X\beta) = \|y - X\beta\|_2^2$$

0.1.2 Задача 2

Показать, что для задачи поиска минимума суммы квадратов в линейной регрессии метод Ньютона за 1 итерацию даёт точное решение при инициализации весов нулевым вектором ($\theta^{(0)} = 0$)

Решение

$$J(\theta) = \|y - X\theta\|_2^2$$

$$\text{Метод Ньютона: } \theta^{(k)} = \theta^{(k-1)} - \nabla_{\theta}^2 J(\theta)^{-1} \cdot \nabla_{\theta} J(\theta)$$

$$\nabla_{\theta} \|y - X\theta\|_2^2 = \nabla_{\theta}(y^T y - 2y^T X\theta + \theta^T X^T X\theta) = -2X^T y + 2X^T X\theta$$

$$\nabla_{\theta}^2 J(\theta) = \nabla_{\theta}(-2X^T y + 2X^T X\theta) = 2X^T X$$

$$\text{Итак: } \theta^{(1)} = 0 - \frac{1}{2}(X^T X)^{-1}(2X^T X\theta^{(0)} - 2X^T y) = (X^T X)^{-1}X^T y$$

0.1.3 Задача 3

Для матрицы X найдено SVD разложение ($X = UDV^T$). Выразить решение задачи МНК через это разложение.

Решение

$$\begin{aligned} \text{Решение задачи МНК имеет вид: } w &= (X^T X)^{-1} X^T y \rightarrow w = \\ (VD \underbrace{U^T U}_I DV^T)^{-1} VD U^T y &= \underbrace{(VD^2 V^T)^{-1}}_{VD^{-2} V^T} VD U^T y = VD^{-2} \underbrace{V^T V}_I DV^T y = \\ VD^{-1} U^T y \end{aligned}$$

0.1.4 Задача 4

Для матрицы X найдено QR разложение ($X = QR$, $Q^T Q = I$). Выразить решение задачи МНК через это разложение.

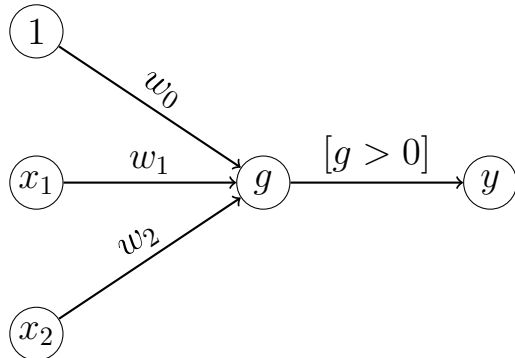
Решение

$$w = (X^T X)^{-1} X^T y = (R^T Q^T Q R)^{-1} R^T Q y = (R^T R)^{-1} R^T y$$

0.2 Perceptron

0.2.1 Задача 1

Рассмотрим простейший персептрон с константой, двумя входами (x_1, x_2) и пороговой функций активации (см. рисунок).



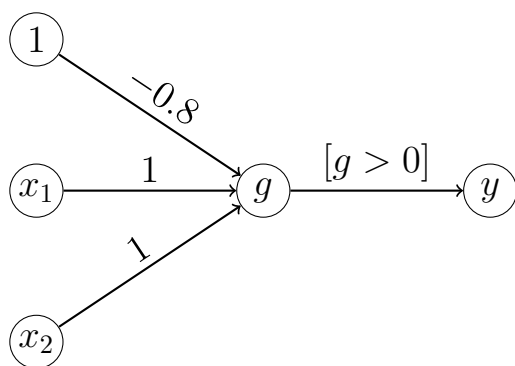
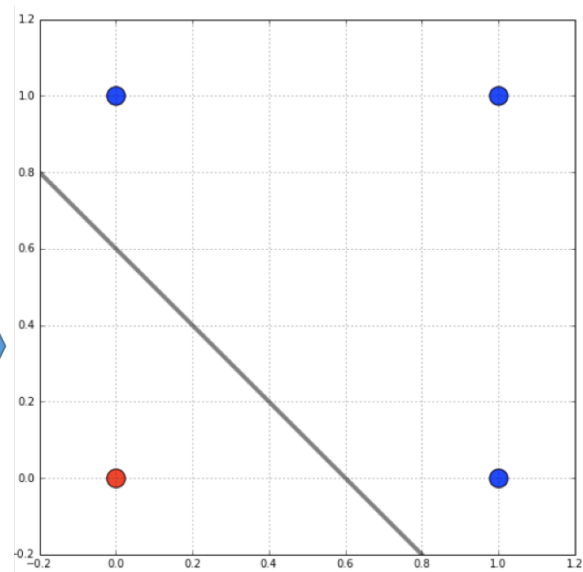
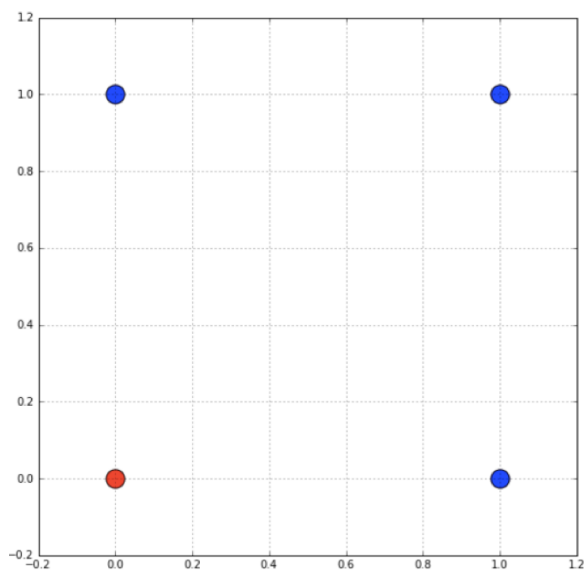
1. Подберите веса персептрона так, чтобы он реализовывал логическое ИЛИ
2. Подберите веса персептрона так, чтобы он реализовывал логическое И
3. Докажите, что веса невозможно подобрать так, чтобы он реализовывал исключающее ИЛИ (XOR)
4. Добавьте персептрону вход $x_3 = x_1 \cdot x_2$ так, чтобы он реализовывал исключающее ИЛИ (XOR)
5. Реализуйте XOR с помощью 3 персептронов с двумя входами и константой

Решение

1) Логическое «ИЛИ»

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	1

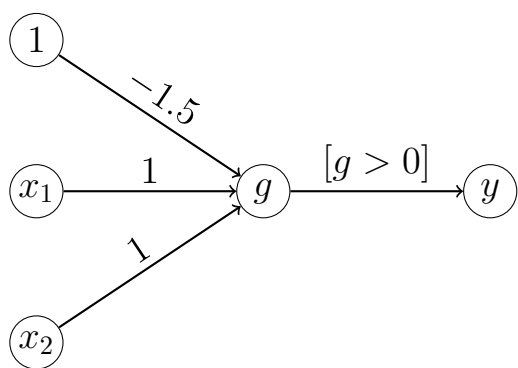
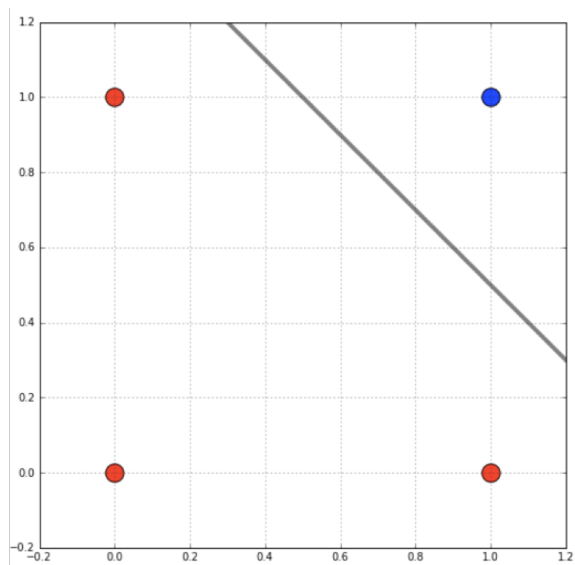
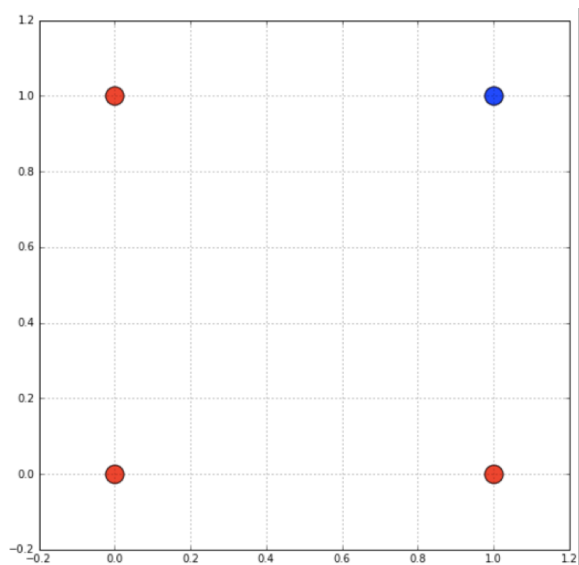
Пороговая функция может выглядеть например $[x_1 + x_2 - 0.8 > 0]$



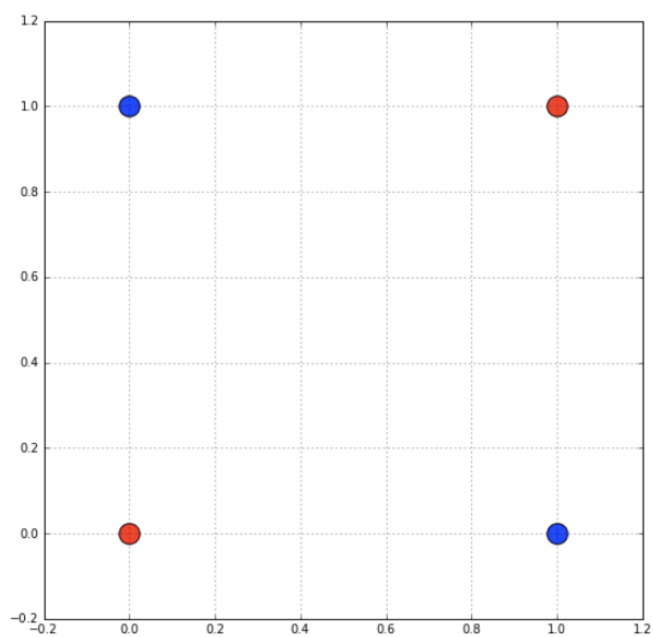
2) Логическое «И»

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Пороговая функция может выглядеть например $[x_1 + x_2 - 1.5 > 0]$



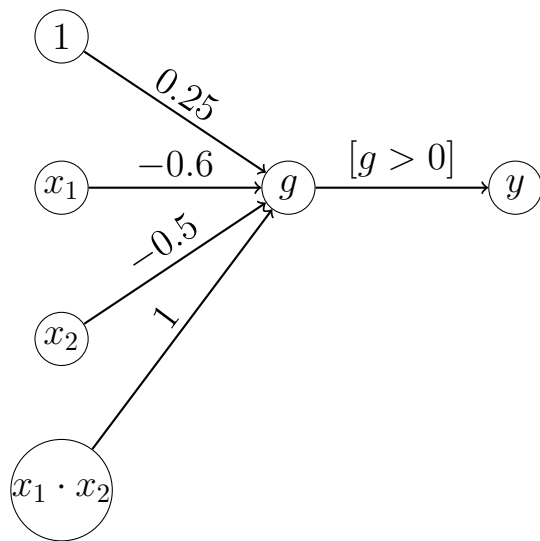
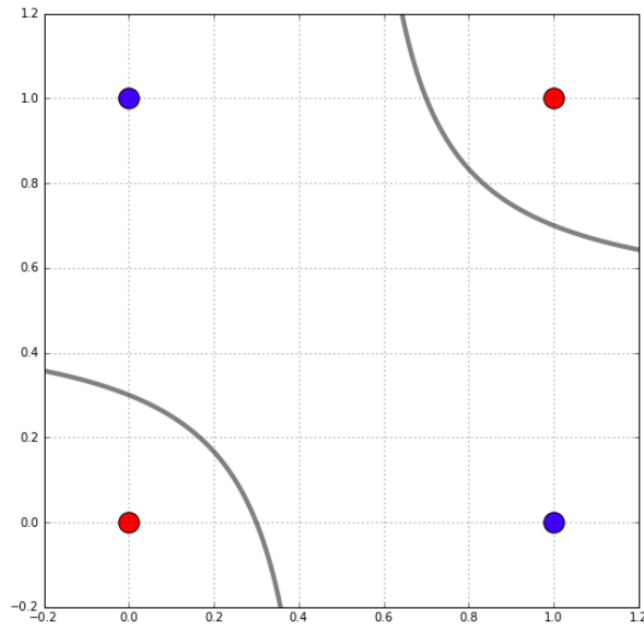
3) Из картинки очевидно, что не существует прямой, верно разделяющей эти точки.



4) Данные точки однако можно разделить гиперболой $x_2 = 0.5 + \frac{0.1}{x_1 - 0.5}$ — остаётся преобразовать её к формату: $a_0 + a_1 \cdot x_1 + a_2 \cdot x_1 \cdot x_2 + a_3 \cdot x_2 = 0$

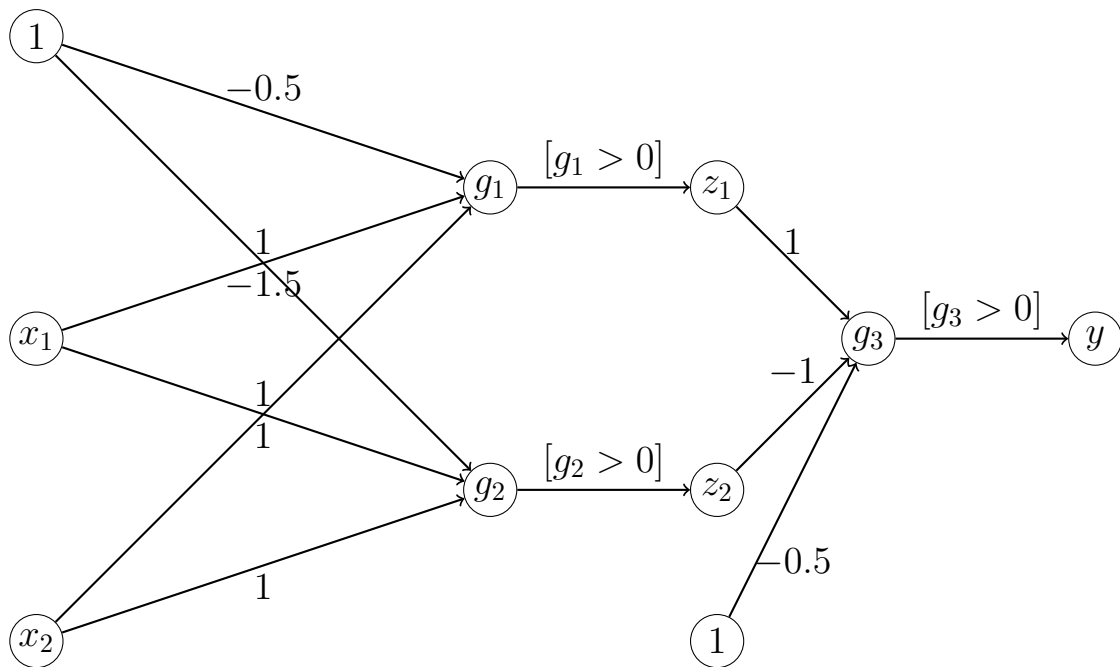
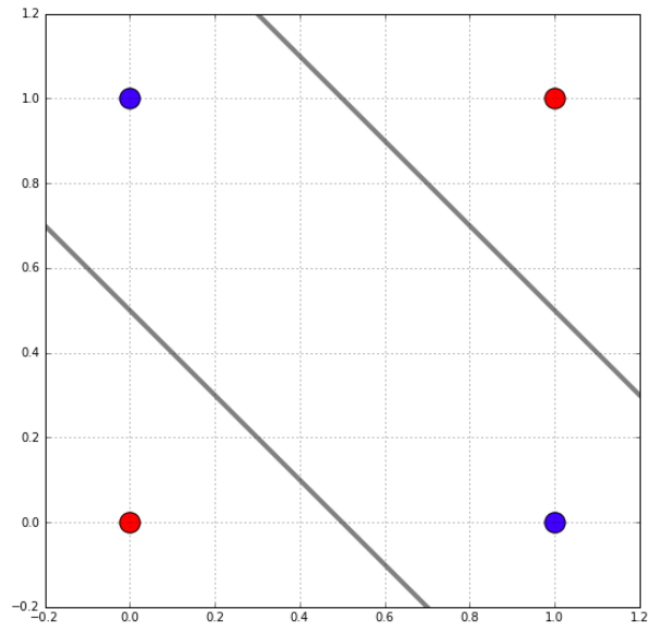
Итак: уравнение разделяющей поверхности принимает вид:

$$[0.25 - 0.6x_1 + x_1 \cdot x_2 - 0.5x_2 > 0]$$



5) Разделение точек с помощью композиции двух персептронов эквивалентно разделению с помощью двух прямых.

$$g_1 : x_1 + x_2 - 0.5, \quad g_2 : x_1 + x_2 - 1.5$$



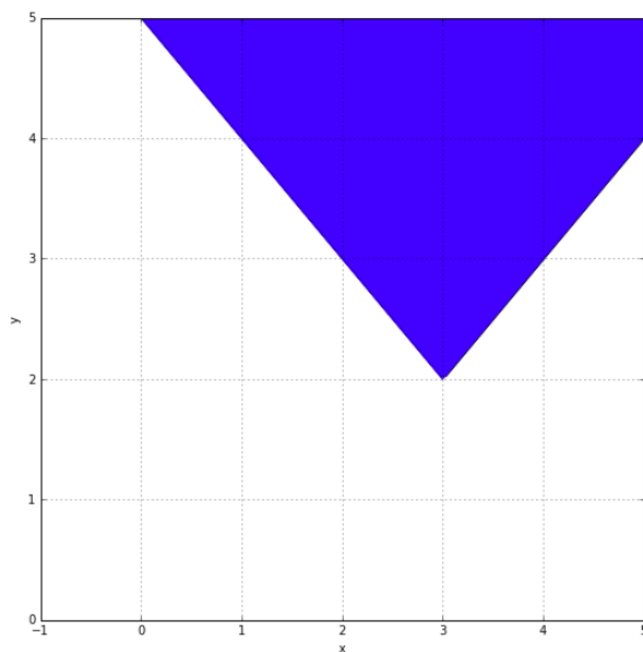
0.2.2 Задача 2

В коробке заваллялось 3 персептрона, у каждого 3 входа (константа, x_1, x_2) и пороговая функция активации. Реализовать с их помощью функцию y

$$y = \begin{cases} 1, & \text{если } x_2 > |x_1 - 3| + 2 \\ 0, & \text{иначе} \end{cases}$$

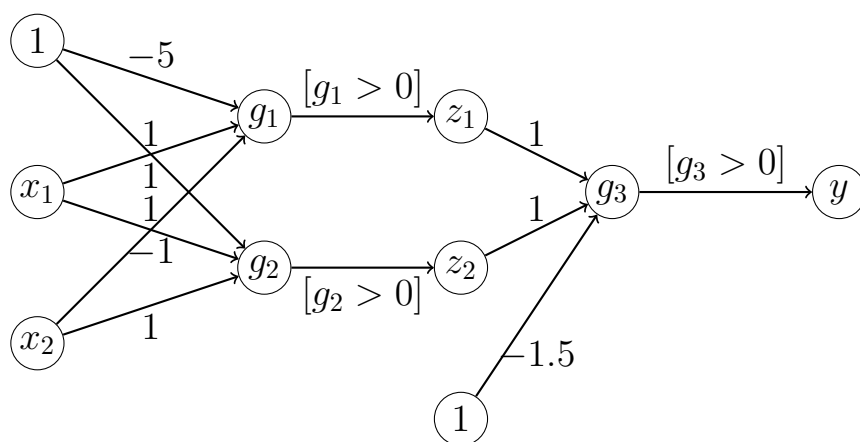
Решение

Нужно задать функцию, которая на синей области будет выдавать 1, а на белой 0



Заметим, что модуль можно представить как 2 отдельные прямые: $x_2 = -x_1 + 5$ и $x_2 = x_1 - 1$. Тогда итоговый алгоритм можно представить, как логическое «И» от двух прямых с порогом:

$$[g_1 + g_2 - 1.5 > 0], \text{ где } g_1 = [x_2 + x_1 - 5 > 0], g_2 = [-x_1 + x_2 + 1 > 0]$$



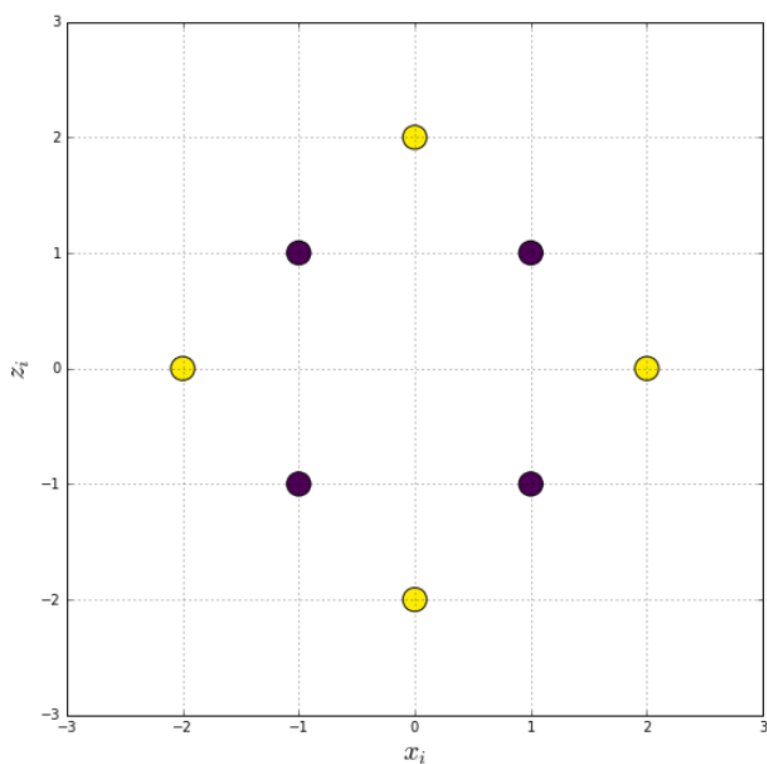
0.2.3 Задача 3

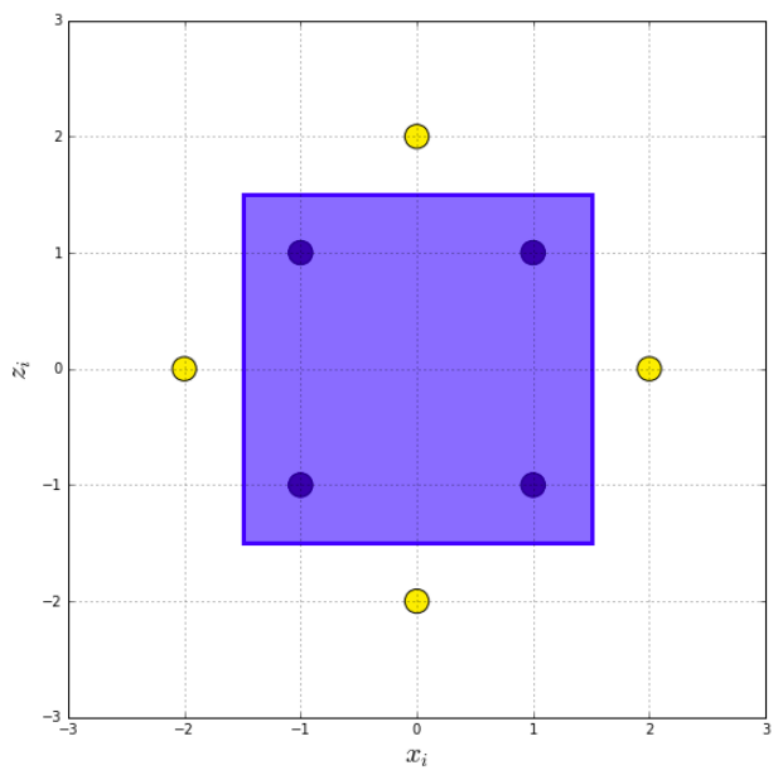
Рассмотрим набор данных:

x_i	z_i	y_i
-1	-1	0
1	-1	0
-1	1	0
1	1	0
0	2	1
2	0	1
0	-2	1
-2	0	1

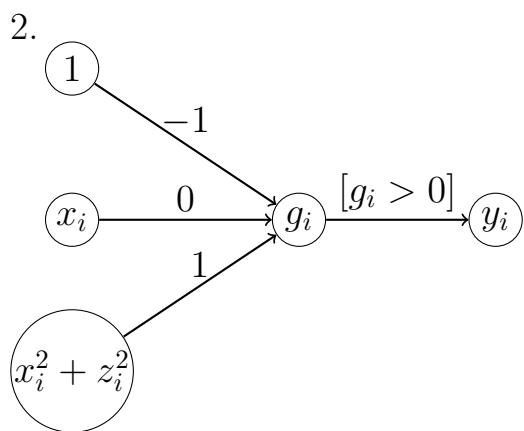
1. Существует ли i персептронов с константой, двумя входами и пороговой функцией активации ($i = 1, 2, 3$), такой что он идеально классифицирует данную выборку.
2. Ввести нелинейное преобразование $h(x_i, z_i)$, такое что хватит даже одного персептрона для идеальной классификации.

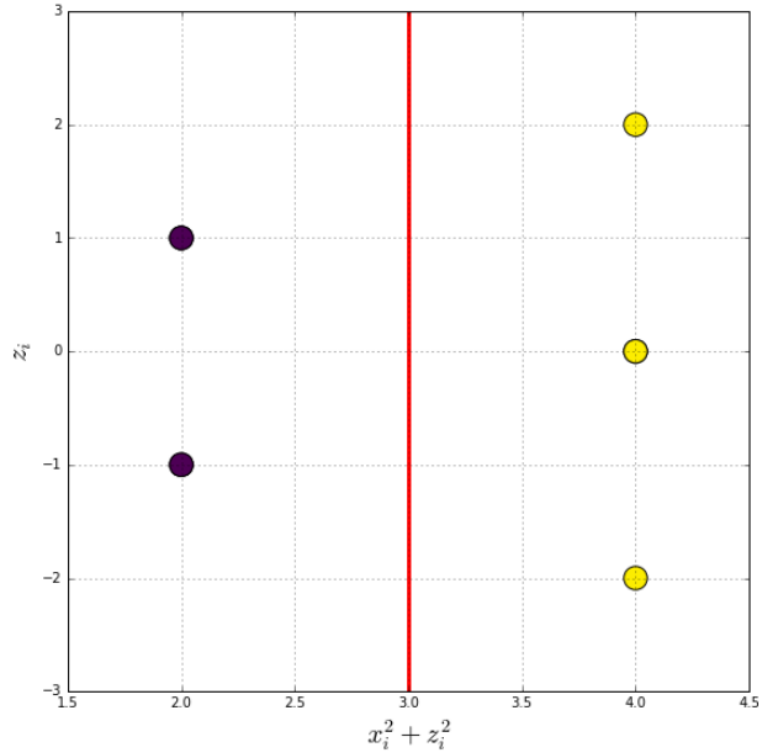
Решение





1. Для того, чтобы разделить данную выборку, необходимо собрать персептрон, который будет внутри синего квадрата выдавать 0, а вне его 1. Это возможно сделать минимум двуслойной нейросетью, а следовательно и минимум 3 персептронами:





0.3 SVM

Постановка и решение задачи SVM в общем виде

$$\begin{cases} \frac{1}{2}w^T w + C \cdot \sum_{i=1}^n \xi_i \rightarrow \min_{w, b, \xi_i} \\ y_i \cdot (x_i^T w + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

$$L = \frac{1}{2}w^T w + C \cdot \sum_{i=1}^n \xi_i - \sum_{i=1}^n \lambda_i \cdot (y_i(x_i^T w + b) - 1 + \xi_i) - \sum_{i=1}^n \mu_i \xi_i$$

Условия Каруша-Куна-Таккера:

$$\begin{cases} \nabla_w L = w - \sum_{i=1}^n \lambda_i y_i x_i = 0 \\ \nabla_b L = - \sum_{i=1}^n \lambda_i y_i = 0 \\ \nabla_{\xi_i} L = C - \lambda_i - \mu_i = 0 \\ \lambda_i \geq 0, \mu_i \geq 0 \\ \lambda_i \cdot (y_i(x_i^T w + b) - 1 + \xi_i) = 0 \\ \mu_i \xi_i = 0 \end{cases} \Rightarrow \begin{cases} w^* = \sum_{i=1}^n y_i \lambda_i x_i \\ \sum_{i=1}^n \lambda_i y_i = 0 \\ \lambda_i + \mu_i = C \end{cases}$$

Отсюда вытекает, что объекты x_i могут быть только 3 типов:
 $(m_i = y_i(x_i^T w + b))$

1) $\lambda_i = 0, \xi_i = 0, \mu_i = C, m_i > 1$ (Эталонные объекты)

2) $0 < \lambda_i < C, \xi_i = 0, 0 < \mu_i < C, m_i = 1$ (Опорные объекты)

3) $\xi_i > 0, \mu_i = 0, \lambda_i = C, m_i < 1$ (Нарушители)

(Причём, если $m_i \in (0, 1)$ – объект внутри полосы, но верно классифицируется, если $m_i = 0$ – объект на границе разделяющей полосы, если $m_i < 0$ – объект неверно классифицируется)

Двойственная задача:

$$q(\lambda, \mu) = \inf_{w, b, \xi} L(w, b, \xi, \lambda, \mu) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j +$$

$$+ \underbrace{C \cdot \sum_{i=1}^n \xi_i - \sum_{i=1}^n \xi_i \overbrace{(\lambda_i + \mu_i)}^C}_{0} - \sum_{i=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j - b \cdot \underbrace{\sum_{i=1}^n \lambda_i y_i}_{0} - \sum_{i=1}^n \lambda_i$$

Итак, двойственная задача выглядит:

$$\left\{ \begin{array}{l} q(\lambda, \mu) = -\frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_{i=1}^n \lambda_i \rightarrow \max_{\lambda} \\ \sum_{i=1}^n \lambda_i y_i = 0 \\ \lambda_i \geq 0 \\ \lambda_i + \mu_i = C \end{array} \right.$$

С учётом последних двух условий систему можно переписать в следующем виде:

$$\left\{ \begin{array}{l} q(\lambda, \mu) = -\frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_{i=1}^n \lambda_i \rightarrow \max_{\lambda} \\ \sum_{i=1}^n \lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C \end{array} \right.$$

Или в матричном виде:

$$(*) \left\{ \begin{array}{l} q = -\frac{1}{2} \lambda^T Q \lambda - \vec{1}^T \lambda \rightarrow \max_{\lambda} \\ y^T \lambda = 0 \\ 0 \leq \lambda \leq C \cdot \vec{1} \end{array} \right.$$

Где $\lambda = (\lambda_1, \dots, \lambda_n)$, $\vec{1} = (1, \dots, 1)$, $y = (y_1, \dots, y_n)$, $Q = (Q_{ij})_{i,j=1}^n$, $Q_{ij} = y_i y_j x_i^T x_j$

Заметим, что всё множество объектов $I = \{1, 2, \dots, n\}$ можно разбить на 3 непересекающихся множества: $I = I_0 \sqcup I_+ \sqcup I_-$

Где I_+ – множество эталонных объектов, I_- – множество нарушителей и I_0 – множество опорных объектов.

Тогда все обозначения можно переписать с учётом этих разделений

$$\lambda = \begin{pmatrix} \lambda_0 \\ \lambda_+ \\ \lambda_- \end{pmatrix} = \begin{pmatrix} \lambda_0 \\ C \cdot 1_+ \\ 0 \end{pmatrix}, \quad y = \begin{pmatrix} y_0 \\ y_+ \\ y_- \end{pmatrix}, \quad \vec{1} = \begin{pmatrix} 1_0 \\ 1_+ \\ 1_- \end{pmatrix}, \quad Q = \begin{pmatrix} Q_{00} & Q_{0+} & Q_{0-} \\ Q_{+0} & Q_{++} & Q_{+-} \\ Q_{-0} & Q_{-+} & Q_{--} \end{pmatrix}$$

Тогда задача (*) переписывается в виде:

$$\begin{cases} q = -\frac{1}{2}\lambda_0^T Q_{00} \lambda_0 - C \cdot 1_+^T Q_{+0} \lambda_0 - \frac{1}{2}C \cdot 1_+^T 1_+ \rightarrow \max_{\lambda_0} \\ y_0^T \lambda_0 + C \cdot y_+^T 1_+ = 0 \end{cases}$$

У этой задачи есть аналитическое решение:

$$L = -\frac{1}{2}\lambda_0^T Q_{00} \lambda_0 - C \cdot 1_+^T Q_{+0} \lambda_0 - \frac{1}{2}C \cdot 1_+^T 1_+ + \gamma \cdot (y_0^T \lambda_0 + C \cdot y_+^T 1_+)$$

Условия Каруша-Куна-Таккера:

$$\begin{cases} \nabla_{\lambda_0} L = -Q_{00} \lambda_0 - C \cdot Q_{0+} 1_+ + \gamma \cdot y_0 = 0 \\ y_0^T \lambda_0 + C \cdot y_+^T 1_+ = 0 \end{cases}$$

$$\boxed{\lambda_0^* = Q_{00}^{-1} (\gamma y_0 - C \cdot Q_{0+} 1_+)}$$

$$y_0^T Q_{00}^{-1} (\gamma y_0 - C \cdot Q_{0+} 1_+) + C \cdot y_+^T 1_+ = 0$$

$$\boxed{\gamma^* = C \cdot \frac{y_+^T 1_+ - y_0^T Q_{00}^{-1} Q_{0+} 1_+}{y_0^T Q_{00}^{-1} y_0}}$$

Теперь можно выписать оценку вектора w :

$$w = \sum_{i=1}^n \lambda_i y_i x_i = y \odot X \lambda =$$

$$\begin{pmatrix} y_0 \\ \vdots \\ y_0 \\ y_+ \\ \vdots \\ y_+ \\ y_- \\ \vdots \\ y_- \end{pmatrix}^T \odot \left(\begin{array}{c|ccc|c|ccc|c} | & \dots & | & | & | & \dots & | & | & | & \dots & | \\ \hline x_0 & \dots & x_0 & | & x_+ & \dots & x_+ & | & x_- & \dots & x_- \\ \hline | & \dots & | & | & | & \dots & | & | & | & \dots & | \end{array} \right) \cdot \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_0 \\ C \cdot 1 \\ \vdots \\ C \cdot 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Итоговый ответ представим в виде:

$$w^* = y \odot X \cdot Q_{00}^{-1} \left(C \cdot \frac{y_+^T 1_+ - y_0^T Q_{00}^{-1} Q_{0+} 1_+}{y_0^T Q_{00}^{-1} y_0} \cdot y_0 - C \cdot Q_{0+} 1_+ \right)$$

$$b^* = \text{med}\{y_i - x_i^T w, \lambda_i > 0, i = 1, \dots, n\}$$

0.3.1 Задача 1

Найти расстояние от точки $x_0 \in \mathbb{R}^d$ до гиперплоскости $w^T x = 0$

Решение

Постановка задачи:

$$\begin{cases} \frac{1}{2} \|x_0 - x\|_2^2 \rightarrow \min_x \\ w^T x = 0 \end{cases}$$

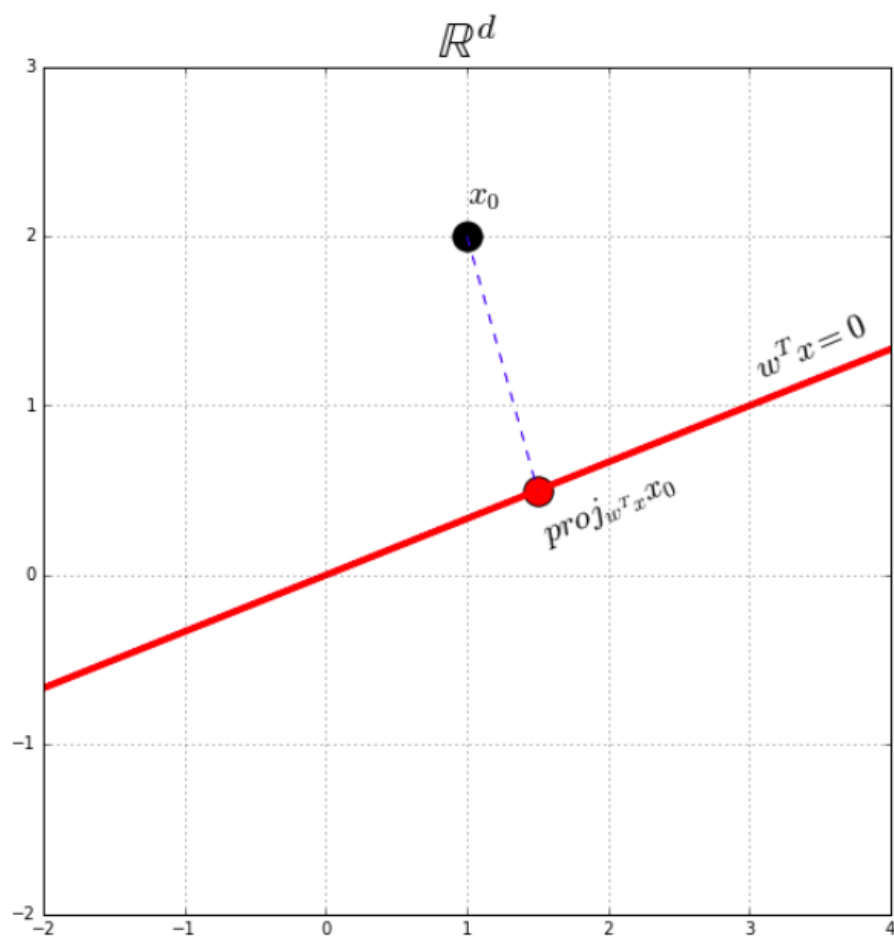
$$L = \frac{1}{2} (x - x_0)^T (x - x_0) + \lambda \cdot w^T x$$

Условия Каруша-Куна-Таккера:

$$\begin{cases} \nabla_x L = x - x_0 + \lambda w = 0 \\ w^T x = 0 \end{cases} \Rightarrow \begin{cases} x = x_0 - \lambda \cdot w \\ w^T x = 0 \end{cases}$$

$$\text{Итак: } w^T x_0 - \lambda \cdot w^T w = 0 \Rightarrow \lambda = \frac{w^T x_0}{w^T w}$$

Тогда, проекция точки x_0 на гиперплоскость $w^T x = 0$ будет $x = x_0 - \frac{w^T x_0}{w^T w} w$



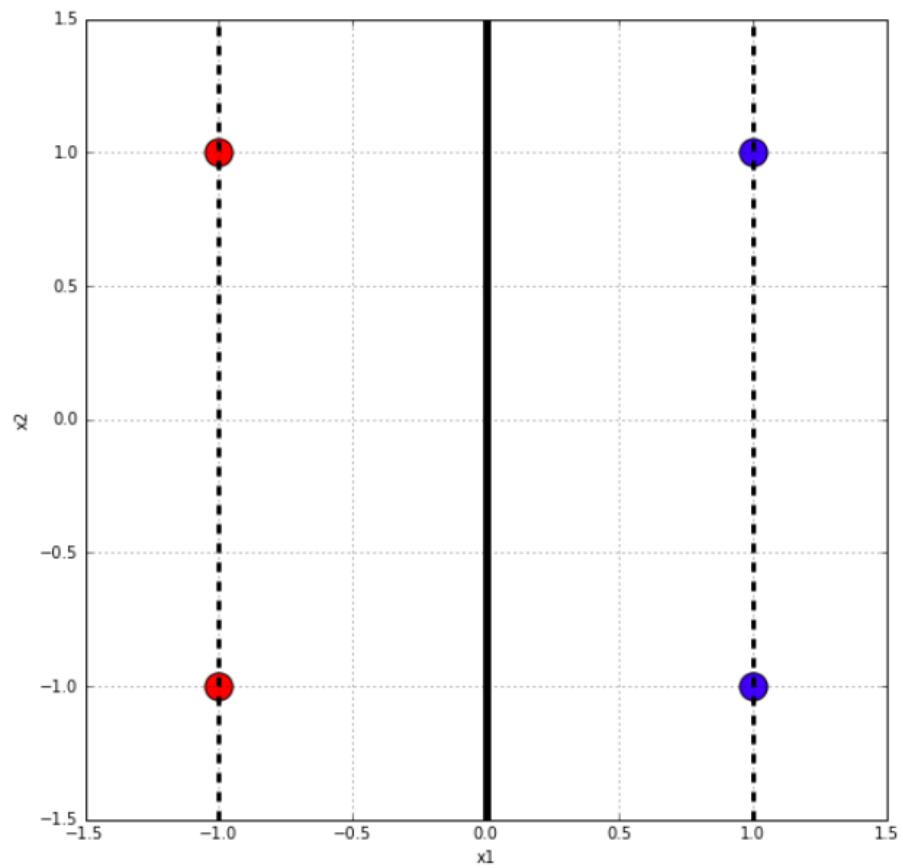
Расстояние от x_0 до $w^T x$ это длина вектора $x_0 - x$

Итак: $\|x_0 - x_0 + \frac{w^T x_0}{w^T w} w\| = \frac{|w^T x_0|}{\|w\|^2} \cdot \|w\| = \frac{|w^T x_0|}{\|w\|}$

0.3.2 Задача 2

На плоскости даны 4 точки: $(1, 1)$, $(1, -1)$ – класса 1, и $(-1, 1)$, $(-1, -1)$ – класса -1. Найти разделяющую гиперплоскость и указать опорные вектора.

Решение



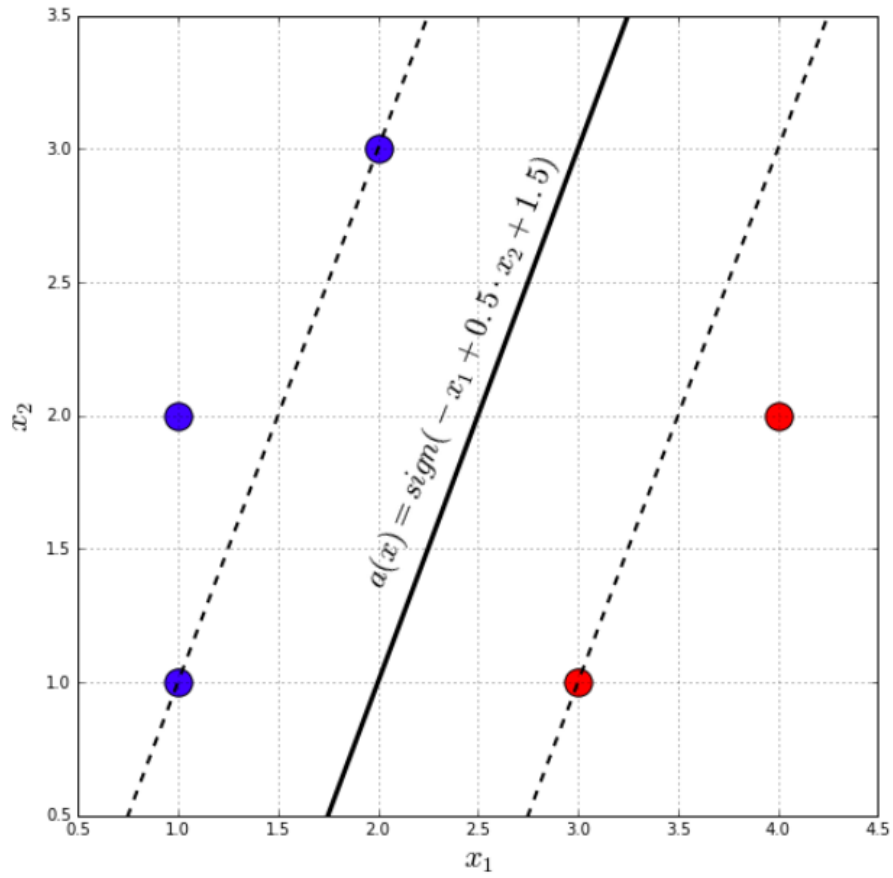
0.3.3 Задача 3

x_1	x_2	y
1	1	1
1	2	1
2	3	1
3	1	-1
4	2	-1

Найти разделяющую гиперплоскость наибольшей ширины.

Решение

Постановка задачи:



$$\left\{ \begin{array}{ll} \frac{1}{2}(w_1^2 + w_2^2) & \rightarrow \min_{w_1, w_2, b} \\ w_1 + w_2 + b & \geq 1 \\ w_1 + 2w_2 + b & \geq 1 \\ 2w_1 + 3w_2 + b & \geq 1 \\ -3w_1 - 1w_2 - b & \geq 1 \\ -4w_1 - 2w_2 - b & \geq 1 \end{array} \right.$$

$$L = \frac{1}{2} \cdot (w_1^2 + w_2^2) - \lambda_1 \cdot (w_1 + w_2 + b - 1) - \lambda_2 \cdot (w_1 + 2w_2 + b - 1) - \lambda_3 \cdot (2w_1 + 3w_2 + b - 1) + \lambda_4 \cdot (3w_1 + w_2 + b + 1) + \lambda_5 \cdot (4w_1 + 2w_2 + b + 1)$$

Условия Каруша-Куна-Таккера:

$$\left\{ \begin{array}{l} L'_{w_1} = w_1 - \lambda_1 - \lambda_2 - 2\lambda_3 + 3\lambda_4 + 4\lambda_5 = 0 \\ L'_{w_2} = w_2 - \lambda_1 - 2\lambda_2 - 3\lambda_3 + \lambda_4 + 2\lambda_5 = 0 \\ L'_b = \lambda_1 + \lambda_2 + \lambda_3 - \lambda_4 - \lambda_5 = 0 \\ \lambda_1 = 0 \quad \text{или} \quad (w_1 + w_2 + b - 1) = 0 \\ \lambda_2 = 0 \quad \text{или} \quad (w_1 + 2w_2 + b - 1) = 0 \\ \lambda_3 = 0 \quad \text{или} \quad (2w_1 + 3w_2 + b - 1) = 0 \\ \lambda_4 = 0 \quad \text{или} \quad (3w_1 + w_2 + b + 1) = 0 \\ \lambda_5 = 0 \quad \text{или} \quad (4w_1 + 2w_2 + b + 1) = 0 \end{array} \right. , \quad \lambda_2, \lambda_5 = 0, \text{ так как обь-}$$

екты $(1, 2)$ и $(4, 2)$ не являются опорными

С учётом этого система принимает вид:

$$\begin{pmatrix} 1 & 0 & 0 & -1 & -2 & 3 \\ 0 & 1 & 0 & -1 & -3 & 1 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ w_2 \\ b \\ \lambda_1 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ -1 \end{pmatrix}$$

Откуда:
$$\begin{pmatrix} w_1 \\ w_2 \\ b \\ \lambda_1 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} = \begin{pmatrix} -1 \\ 0.5 \\ 1.5 \\ 0.375 \\ 0.25 \\ 0.625 \end{pmatrix}$$

И следовательно классификатор принимает вид:

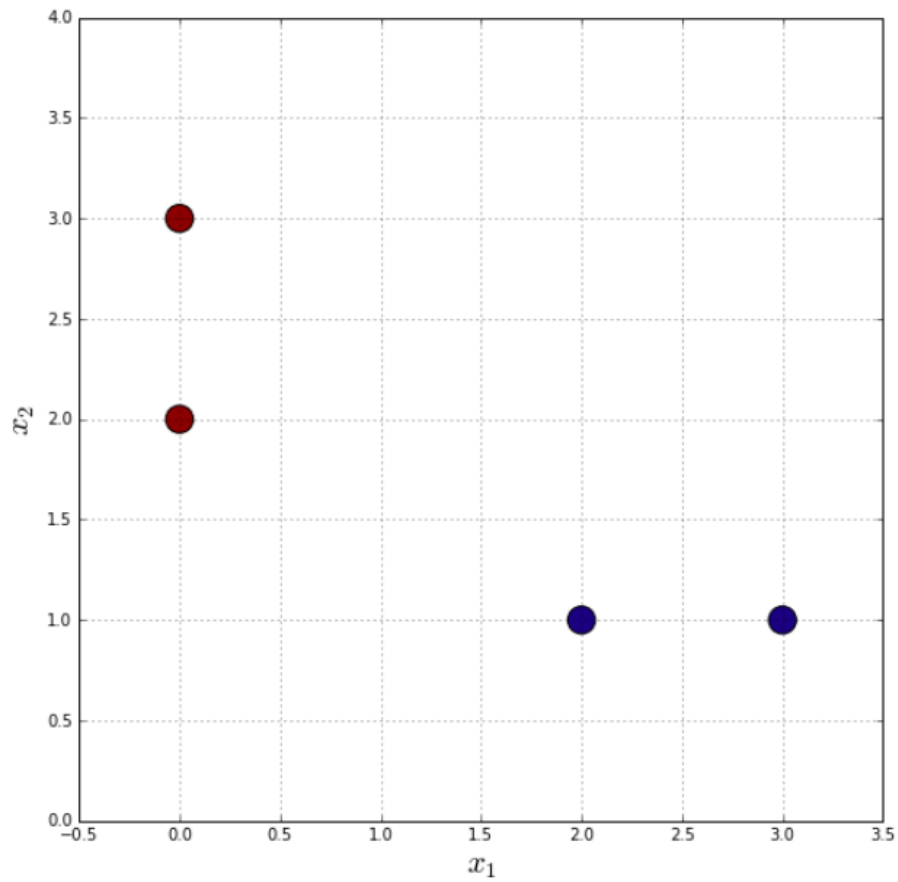
$$a(x) = \text{sign}(-x_1 + 0.5x_2 + 1.5)$$

0.3.4 Задача 4

Найти разделяющую гиперплоскость наибольшей ширины, используя решение двойственной задачи. (В качестве опорных объектов взять точки $(0,2)$, $(2,1)$)

x_1	x_2	y
0	2	1
0	3	1
2	1	-1
3	1	-1

Решение



Известно, что веса в SVM можно найти по формуле: $w = \sum_{i=1}^n \lambda_i \cdot y_i \cdot x_i$, где

$$\lambda = Q_0^{-1} \cdot \left(\vec{1} - \frac{y_0^T Q_0^{-1} \vec{1}}{y_0^T Q_0^{-1} y_0} \cdot y_0 \right)$$

$\vec{1}$ – вектор из единиц длины количество опорных объектов

y_0 – вектор меток класса для опорных объектов

$Q_0 = (y_i y_j < x_i, x_j >)_{i,j=1}^{op}$, op – количество опорных объектов

Тогда: $Q_0 = \begin{pmatrix} y_1 y_1 x_1^T x_1 & y_1 y_3 x_1^T x_3 \\ y_1 y_3 x_1^T x_3 & y_3 y_3 x_3^T x_3 \end{pmatrix} = \begin{pmatrix} 4 & -2 \\ -2 & 5 \end{pmatrix}$

Итак: $Q_0^{-1} = \frac{1}{16} \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix}$

$y_0^T Q_0^{-1} \vec{1} = \frac{1}{16} (1 \quad -1) \cdot \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{16} (3 \quad -2) \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{16}$

$y_0^T Q_0^{-1} y_0 = \frac{1}{16} (3 \quad -2) \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{5}{16}$

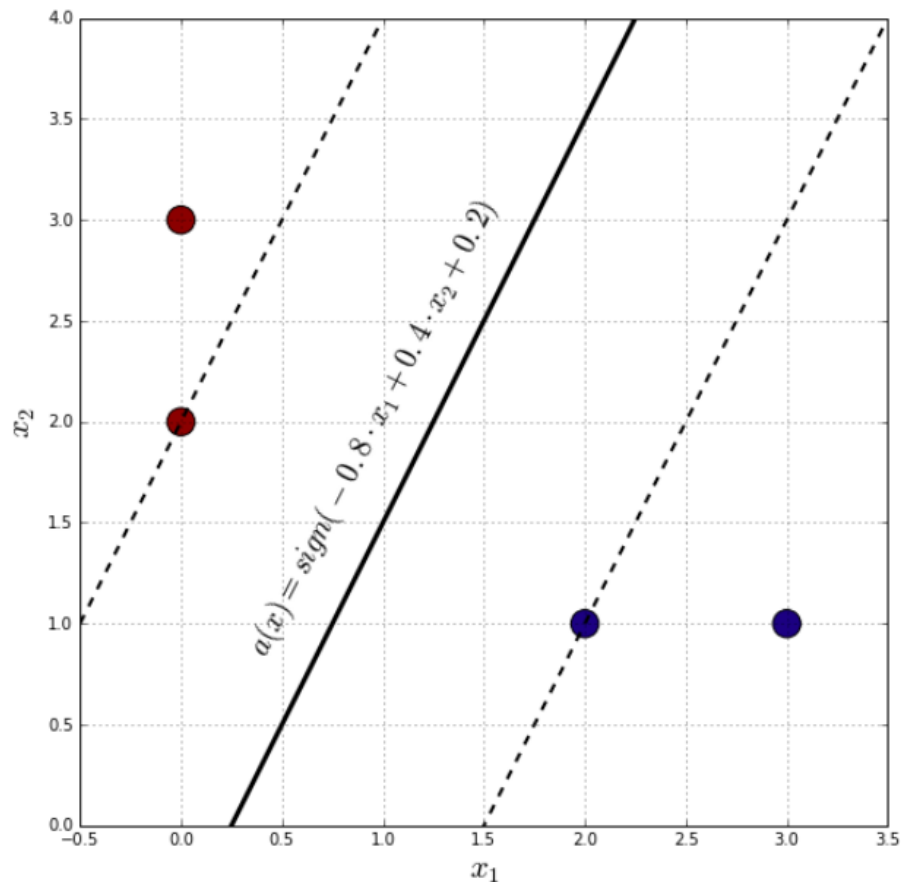
$$\lambda = \frac{1}{16} \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix} \cdot \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{16} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right) = \frac{1}{16} \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} \frac{4}{5} \\ \frac{6}{5} \end{pmatrix} = \frac{1}{80} \begin{pmatrix} 32 \\ 32 \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.4 \end{pmatrix}$$

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = 0.4 \cdot 1 \cdot \begin{pmatrix} 0 \\ 2 \end{pmatrix} - 1 \cdot 0.4 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.8 \\ 0.4 \end{pmatrix}$$

В виду малого количества данных будем искать b как среднее арифметическое $\{y_i - w_1 x_i^1 - w_2 x_i^2, \quad \lambda_i > 0\}$

$$\text{Итак: } b = 0.5 \cdot (1 + 0.8 \cdot 0 - 0.4 \cdot 2 - 1 + 0.8 \cdot 2 - 0.4 \cdot 1) = 0.5 \cdot 0.4 = 0.2$$

Итак, классификатор имеет вид: $a(x) = \text{sign}(-0.8x_1 + 0.4x_2 + 0.2)$



(решение прямой задачи даёт такую-же разделяющую линию)

0.4 Naïve Bayes

0.4.1 Задача 1

Рассмотрим задачу классификации текстов $D = \{d_1, d_2, \dots, d_{|D|}\}$ на K классов $Y = \{1, 2, \dots, K\}$. Каждый документ представляет из себя подмножество множества слов $W = \{w_1, w_2, \dots, w_{|W|}\}$. В качестве признаков для каждого документа выберем индикатор вхождения слов в него. Матрица «объекты-признаки» задаётся, как:

$$x_{ij} = [w_j \in d_i], \quad i = 1, \dots, |D| \quad j = 1, \dots, |W|$$

Для решение задачи воспользуемся наивным байесовским классификатором, который основывается на предположении независимости признаков:

$$p(x_{i1}, \dots, x_{i|W|} | y_i) = p(x_{i1} | y_i) \cdot p(x_{i2} | y_i) \cdot \dots \cdot p(x_{i|W|} | y_i)$$

Будем считать, что при фиксированном классе каждый признак имеет распределение Бернулли, тем самым априорное распределение и функция правдоподобия выглядят, как:

$$p(k | \pi) = \pi_k, \quad k = 1, \dots, K$$

$$p(x_{i,j} | k, \theta) = \theta_{jk}^{x_{ij}} \cdot (1 - \theta_{jk})^{1-x_{ij}}, \quad i = 1, \dots, |D|, \quad j = 1, \dots, |W|, \quad k = 1, \dots, K$$

Распределение одного документа запишется следующим образом:

$$p(d_i, y_i | \pi, \theta) = p(y_i | \pi) \prod_{j=1}^{|W|} p(x_{ij} | \theta, y_i) = \prod_{k=1}^K \pi_k^{[y_i=k]} \cdot \prod_{j=1}^{|W|} \prod_{k=1}^K p(x_{ij} | \theta_{jk}, k)^{[y_i=k]}.$$

Найти оценки максимального правдоподобия для параметров π и θ

Решение

π_k меняются не независимо, они в сумме должны давать 1 – это надо учесть в функционале (Лагранж).

$$L = \prod_{i=1}^{|D|} p(d_i, y_i | \pi, \theta) \rightarrow \max_{\pi, \theta}$$

$$L = \prod_{i=1}^{|D|} \prod_{k=1}^K \left(\pi_k^{[y_i=k]} \cdot \prod_{j=1}^{|W|} \theta_{jk}^{x_{ij} \cdot [y_i=k]} (1 - \theta_{jk})^{(1-x_{ij}) \cdot [y_i=k]} \right)$$

$$\begin{aligned}
\log L &= \sum_{i=1}^{|D|} \sum_{k=1}^K ([y_i = k] \log \pi_k + \sum_{j=1}^{|W|} x_{ij} \cdot [y_i = k] \cdot \log \theta_{jk} + (1 - x_{ij}) \cdot [y_i = k] \cdot \log(1 - \theta_{jk})) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) \\
\log L'_{\pi_k} &= \sum_{i=1}^{|D|} \frac{[y_i=k]}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = -\frac{1}{\lambda} \sum_{i=1}^{|D|} [y_i = k] \\
\sum_{k=1}^K \pi_k &= 1 \Rightarrow \sum_{k=1}^K -\frac{1}{\lambda} \sum_{i=1}^{|D|} [y_i = k] = -\frac{1}{\lambda} \sum_{i=1}^{|D|} \underbrace{\sum_{k=1}^K [y_i = k]}_1 = -|D| \lambda = 1 \Rightarrow \lambda = -\frac{1}{|D|}
\end{aligned}$$

Итак:
$$\hat{\pi}_k = \frac{\sum_{i=1}^{|D|} [y_i = k]}{|D|}$$

$$\begin{aligned}
\log L'_{\theta_{jk}} &= \sum_{i=1}^{|D|} \frac{x_{ij} \cdot [y_i=k]}{\theta_{jk}} - \frac{(1-x_{ij}) \cdot [y_i=k]}{1-\theta_{jk}} = \sum_{i=1}^{|D|} [y_i = k] (x_{ij} - \theta_{jk} \cdot x_{ij} - \theta_{jk} + \theta_{jk} \cdot x_{ij}) = \\
&\sum_{i=1}^{|D|} x_{ij} \cdot [y_i = k] - \theta_{jk} \cdot [y_i = k] = 0
\end{aligned}$$

$$\hat{\theta}_{jk} = \frac{\sum_{i=1}^{|D|} [y_i = k] \cdot x_{ij}}{\sum_{i=1}^{|D|} [y_i = k]}$$

0.4.2 Задача 2

Расширим модель из предыдущей задачи и введём априорное распределение на параметры θ_{jk} . В качестве априорного распределения возьмём Бета распределение:

$$Beta(x|\beta_1, \beta_2) = \frac{1}{Beta(\beta_1, \beta_2)} \cdot x^{\beta_1-1} \cdot (1-x)^{\beta_2-1}, \quad \beta_1, \beta_2 \geq 0$$

Апостериорное распределение на θ_{jk} примет вид:

$$p(\theta_{jk}|D) \propto p(\theta_{jk}) \prod_{i=1}^{|D|} p(y_i, d_i|\theta_{jk}), \text{ где } p(\theta_{jk}) = Beta(\theta_{jk}|\beta_1, \beta_2)$$

1) Выписать $p(\theta_{jk}|D)$ в явном виде

2) Найти $\hat{\theta}_{jk}$, где $\hat{\theta}_{jk} = \mathbb{E}_{p(\theta_{jk}|D)}\theta_{jk} = \int_0^1 \theta_{jk} \cdot p(\theta_{jk}|D) d\theta_{jk}$

Решение

1) $p(\theta_{jk}|D) = \frac{p(D|\theta_{jk}) \cdot p(\theta_{jk})}{\int_0^1 p(D|\theta_{jk}) \cdot p(\theta_{jk}) d\theta_{jk}}$ – формула Байеса

$$\int_0^1 p(D|\theta_{jk}) \cdot p(\theta_{jk}) d\theta_{jk} = \int_0^1 \frac{1}{Beta(\beta_1, \beta_2)} \cdot \theta_{jk}^{\beta_1-1} \cdot (1 - \theta_{jk})^{\beta_2-1} \cdot \prod_{i=1}^{|D|} \prod_{k=1}^K \left(\pi_k^{[y_i=k]} \cdot \prod_{j=1}^{|W|} \theta_{jk}^{x_{ij} \cdot [y_i=k]} (1 - \theta_{jk})^{(1-x_{ij}) \cdot [y_i=k]} \right) d\theta_{jk} =$$

$$\frac{1}{Beta(\beta_1, \beta_2)} \prod_{i=1}^{|D|} \prod_{k=1}^K \left(\pi_k^{[y_i=k]} \cdot \prod_{j=1}^{|W|} \underbrace{\int_0^1 \theta_{jk}^{x_{ij} \cdot [y_i=k] + \beta_1 - 1} \cdot (1 - \theta_{jk})^{(1-x_{ij}) \cdot [y_i=k] + \beta_2 - 1} d\theta_{jk}}_{Beta(x_{ij} \cdot [y_i=k] + \beta_1, (1-x_{ij}) \cdot [y_i=k] + \beta_2)} \right)$$

Заметим, что $\int_0^1 p(D|\theta_{jk}) \cdot p(\theta_{jk}) d\theta_{jk}$ – нормировочная константа распределения $p(D|\theta_{jk}) \cdot p(\theta_{jk})$, поэтому распределение

$$p(\theta_{jk}|D) = \prod_{i=1}^{|D|} \prod_{k=1}^K \prod_{j=1}^{|W|} Beta(\theta_{jk} | x_{ij} \cdot [y_i = k] + \beta_1, (1 - x_{ij}) \cdot [y_i = k] + \beta_2)$$

$$p(\theta_{jk}|D) = \prod_{i=1}^{|D|} \prod_{k=1}^K \prod_{j=1}^{|W|} \frac{1}{Const} \cdot \theta_{jk}^{x_{ij} \cdot [y_i=k] + \beta_1 - 1} (1 - \theta_{jk})^{(1-x_{ij}) \cdot [y_i=k] + \beta_2 - 1}$$

где $Const = Beta(x_{ij} \cdot [y_i = k] + \beta_1, (1 - x_{ij}) \cdot [y_i = k] + \beta_2)$

2) $\hat{\theta}_{jk} = \mathbb{E}_{p(\theta_{jk}|D)}\theta_{jk}$ – матожидание Бета распределения

Для $Beta(x|a, b)$ известно, что $\mathbb{E}x = \frac{a}{a+b}$

$$\text{Итак: } \hat{\theta}_{jk} = \prod_{i=1}^{|D|} \prod_{k=1}^K \prod_{j=1}^{|W|} \frac{x_{ij} \cdot [y_i=k] + \beta_1}{x_{ij} \cdot [y_i=k] + \beta_1 + (1-x_{ij}) \cdot [y_i=k] + \beta_2} =$$

$$\prod_{i=1}^{|D|} \prod_{k=1}^K \prod_{j=1}^{|W|} \frac{x_{ij} \cdot [y_i = k] + \beta_1}{[y_i = k] + \beta_1 + \beta_2}$$

0.5 Логистическая регрессия

0.5.1 Задача 1

Имеется набор данных: $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^m, y_i \in \{0, 1\}$. Предположим, что данные подчиняются следующему закону:

$$p(y) = \phi^y \cdot (1 - \phi)^{1-y}$$

$$p(x|y = k) \sim \mathcal{N}(\mu_k, \Sigma) \Rightarrow p(x|y = k) = \frac{1}{\sqrt{\text{Det}(2\pi \cdot \Sigma)}} \cdot \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\right\}$$

1) Доказать, что апостериорное распределение метки (y), при данном x принимает вид логистической регрессии: $p(y = 1|x) = \frac{1}{1 + \exp\{-\theta^T x\}}$ для какого-то параметра θ .

2) Найти оценки максимального правдоподобия для параметров $\phi, \mu_0, \mu_1, \Sigma$

Решение

$$\begin{aligned} 1) \quad \text{По формуле Байеса:} \quad p(y = 1|x) &= \frac{p(x|y=1) \cdot p(y=1)}{p(x)} = \frac{p(x|y=1) \cdot p(y=1)}{\sum_y p(x,y)} = \frac{p(x|y=1) \cdot p(y=1)}{p(x|y=1) \cdot p(y=1) + p(x|y=0) \cdot p(y=0)} = \\ &= \frac{\exp\{-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\} \cdot \phi}{\exp\{-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\} \cdot \phi + \exp\{-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\} \cdot (1 - \phi)} = \\ &= \frac{1 - \phi}{1 + \underbrace{\frac{\phi}{1 - \phi}}_{\exp\{\log \frac{1 - \phi}{\phi}\}} \cdot \exp\{-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\}} = \\ &= \frac{1}{1 + \exp\{\log \frac{1 - \phi}{\phi} - \frac{1}{2}[2(\mu_1 - \mu_0)^T \Sigma^{-1}x + \mu_0^T \Sigma^{-1}\mu_0 - \mu_1^T \Sigma^{-1}\mu_1]\}} \end{aligned}$$

$$\text{Итак: } x = \begin{bmatrix} 1 \\ x \end{bmatrix}, \quad \theta = \begin{bmatrix} \log \frac{1 - \phi}{\phi} + \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^T \Sigma^{-1}\mu_0 \\ \Sigma^{-1}(\mu_0 - \mu_1) \end{bmatrix}$$

$$2) \quad L = \prod_{i=1}^n p(x_i) \cdot p(x_i|y_i) \rightarrow \max_{\Sigma, \mu_k, \phi}$$

$$\begin{aligned} \log L &= \sum_{i=1}^n \log p(y_i) + \sum_{i=1}^n \log p(x_i|y_i) = \sum_{i=1}^n y_i \cdot \log \phi + (1 - y_i) \cdot \log(1 - \phi) + \\ &+ \sum_{i: y_i=1} -\frac{m}{2} \log 2\pi - \frac{1}{2} \log \text{Det}(\Sigma) - \frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + \sum_{i: y_i=0} -\frac{m}{2} \log 2\pi - \\ &- \frac{1}{2} \log \text{Det}(\Sigma) - \frac{1}{2}(x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0) \rightarrow \max_{\Sigma, \mu_k, \phi} \end{aligned}$$

$$\nabla_{\phi} \log L = \sum_{i=1}^n \frac{y_i}{\phi} - \frac{1-y_i}{1-\phi} = 0 \Rightarrow \sum_{i=1}^n y_i - \phi \cdot y_i - \phi + \phi \cdot y_i = 0 \Rightarrow \hat{\phi} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\nabla_{\mu_k} \log L = \{\nabla_x (x - a)^T B (x - a) = 2 \cdot B (x - a), \quad B = B^T\} =$$

$$\sum_{i:y_i=k} -\Sigma^{-1} (x_i - \mu_k) = 0 \Rightarrow \hat{\mu}_k = \frac{\sum_{i=1}^n [y_i = k] x_i}{\sum_{i=1}^n [y_i = k]}$$

$$\nabla_{\Sigma} \log L = \{\nabla_X a^T X^{-1} b = -X^{-1} a b^T X^{-1}, \nabla_X \log \text{Det}(X) = X^{-1}, X = X^T\} =$$

$$= \sum_{i:y_i=1} \left(-\frac{1}{2} \Sigma^{-1} - \frac{1}{2} \Sigma^{-1} (x_i - \mu_1)^T (x_i - \mu_1) \Sigma^{-1} \right) +$$

$$\sum_{i:y_i=0} \left(-\frac{1}{2} \Sigma^{-1} - \frac{1}{2} \Sigma^{-1} (x_i - \mu_0)^T (x_i - \mu_0) \Sigma^{-1} \right) = 0 \mid \cdot -2 \Sigma^2$$

$$n \cdot \Sigma = \sum_{i:y_i=1} (x_i - \mu_1)^T (x_i - \mu_1) + \sum_{i:y_i=0} (x_i - \mu_0)^T (x_i - \mu_0)$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n [y_i = 1] \cdot (x_i - \mu_0)^T (x_i - \mu_0) + [y_i = 0] \cdot (x_i - \mu_0)^T (x_i - \mu_0)$$

0.5.2 Задача 2

Винни-Пух знает, что мёд бывает правильный $honey_i = 1$ и неправильный $honey_i = 0$. Пчёлы так же бывают правильные $bee_i = 1$ и неправильные $bee_i = 0$. По 100 попыткам добыть мёд Винни-Пух составил таблицу сопряжённости:

	$honey_i = 1$	$honey_i = 0$
$bee_i = 1$	12	36
$bee_i = 0$	32	20

Он использует логистическую регрессию с константой для прогнозирования правильности мёда с помощью информации и правильности пчёл.

- 1) Какие коэффициенты получит Винни-Пух.
- 2) Какой прогноз выдаст логистическая регрессия при столкновении с неправильными пчёлами

Решение

1) Логистическая регрессия настраивается с помощью максимизации правдоподобия по выборке:

$$\begin{aligned}
 p(\text{honey}_i = 1 | \text{bee}_i) &= \sigma(w_0 + w_1 \cdot \text{bee}_i) \\
 L &= \prod_{i=1}^{100} p(\text{honey}_i | \text{bee}_i) \rightarrow \max_{w_0, w_1} \\
 L &= \prod_{i=1}^{100} p(\text{honey}_i = 1 | \text{bee}_i = 1) \cdot p(\text{honey}_i = 1 | \text{bee}_i = 0) \cdot p(\text{honey}_i = 0 | \text{bee}_i = 1) \cdot p(\text{honey}_i = 0 | \text{bee}_i = 0) = \\
 &= \prod_{i=1}^{100} \sigma(w_0 + w_1) \cdot \sigma(w_0) \cdot (1 - \sigma(w_0 + w_1)) \cdot (1 - \sigma(w_0)) = \\
 &= \sigma(w_0 + w_1)^{12} \cdot \sigma(w_0)^{32} \cdot (1 - \sigma(w_0 + w_1))^{36} \cdot (1 - \sigma(w_0))^{20} \rightarrow \max_{w_0, w_1}
 \end{aligned}$$

$$\log L = 12 \cdot \log \sigma(w_0 + w_1) + 32 \cdot \log \sigma(w_0) + 36 \cdot \log(1 - \sigma(w_0 + w_1)) + 20 \cdot \log(1 - \sigma(w_0)) \rightarrow \max_{w_0, w_1}$$

$$\begin{cases} \log L'_{w_0} = \frac{12 \cdot \sigma'_{w_0}(w_0 + w_1)}{\sigma(w_0 + w_1)} + \frac{32 \cdot \sigma'_{w_0}(w_0)}{\sigma(w_0)} - \frac{36 \cdot \sigma'_{w_0}(w_0 + w_1)}{1 - \sigma(w_0 + w_1)} - \frac{20 \cdot \sigma'_{w_0}(w_0)}{1 - \sigma(w_0)} = 0 \\ \log L'_{w_1} = \frac{12 \cdot \sigma'_{w_1}(w_0 + w_1)}{\sigma(w_0 + w_1)} - \frac{36 \cdot \sigma'_{w_1}(w_0 + w_1)}{1 - \sigma(w_0 + w_1)} = 0 \end{cases}$$

Докажем, что $\sigma'_{w_0}(w_0 + w_1) = \sigma'_{w_1}(w_0 + w_1)$:

$$\begin{aligned}
 \sigma'_{w_0}(w_0 + w_1) &= \left(\frac{1}{1 + e^{-(w_0 + w_1)}} \right)'_{w_0} = \frac{-e^{-(w_0 + w_1)}}{(1 + e^{-(w_0 + w_1)})^2} = \left(\frac{1}{1 + e^{-(w_0 + w_1)}} \right)'_{w_1} = \\
 &= \sigma'_{w_1}(w_0 + w_1)
 \end{aligned}$$

Тогда предыдущая система перепишется, как:

$$\begin{cases} \log L'_{w_0} = \frac{32 \cdot \sigma'_{w_0}(w_0)}{\sigma(w_0)} - \frac{20 \cdot \sigma'_{w_0}(w_0)}{1 - \sigma(w_0)} = 0 & (1) \\ \log L'_{w_1} = \frac{12 \cdot \sigma'_{w_1}(w_0 + w_1)}{\sigma(w_0 + w_1)} - \frac{36 \cdot \sigma'_{w_1}(w_0 + w_1)}{1 - \sigma(w_0 + w_1)} = 0 & (2) \end{cases}$$

$$(1) : (1 - \sigma(w_0)) \cdot 32 \cdot \sigma'(w_0) - 20 \cdot \sigma(w_0) \cdot \sigma'(w_0) = 0 \mid : \sigma'(w_0) \neq 0$$

$$32 - 20 \cdot \sigma(w_0) = 0 \Rightarrow \boxed{\sigma(w_0) = \frac{32}{52}}$$

$$\frac{1}{1 + e^{-w_0}} = \frac{32}{52} \rightarrow e^{-w_0} = \frac{52}{32} - 1 = \frac{20}{32} \rightarrow \boxed{w_0 = \ln \frac{32}{20}}$$

$$(2) : 12 \cdot (1 - \sigma(w_0 + w_1)) \cdot \sigma'(w_0 + w_1) - 36 \cdot \sigma(w_0 + w_1) \cdot \sigma'(w_0 + w_1) = 0 \mid : \sigma'(w_0 + w_1) \neq 0$$

$$12 - 48 \cdot \sigma(w_0 + w_1) = 0 \rightarrow \frac{1}{1+e^{-w_0-w_1}} = \frac{1}{4} \rightarrow \underbrace{e^{-w_0}}_{\frac{20}{32}} \cdot e^{-w_1} = 3 \rightarrow e^{-w_1} = \frac{96}{20} \rightarrow$$

$$w_1 = \ln \frac{5}{24}$$

$$\text{Итак: } \hat{p}(\text{honey}_i = 1 | \text{bee}_i) = \sigma(\ln \frac{32}{20} + \ln \frac{5}{24} \cdot \text{bee}_i) = \frac{1}{1+e^{-\ln \frac{32}{20} \cdot e^{-\ln \frac{5}{24} \cdot \text{bee}_i}}} = \frac{1}{1+\frac{20}{32} \cdot \left(\frac{24}{5}\right)^{\text{bee}_i}}$$

$$2) \hat{p}(\text{honey}_i = 1 | \text{bee}_i = 0) = \frac{1}{1+\frac{20}{32}} = \frac{32}{52}$$

0.5.3 Задача 3

Рассмотрим целевую функцию логистической регрессии с константой:
 $x_i, w \in \mathbb{R}^d$

$$Q(w) = \frac{1}{n} \sum_{i=1}^n L(y_i, b_i)$$

$$b_i = b(x_i) = \frac{1}{1+\exp(-x_i^T w)}, \quad L(y_i, b_i) = \begin{cases} -\log b_i & , \text{если } y_i = 1 \\ -\log(1 - b_i) & , \text{если } y_i = -1 \end{cases}$$

- 1) Найдите $\nabla Q(w), \nabla^2 Q(w)$
- 2) Найдите $\nabla Q(0), \nabla^2 Q(0)$
- 3) Найдите квадратичную аппроксимацию $Q(w)$ в районе $w = 0$
- 4) Найти w^* , минимизирующий квадратичную аппроксимацию

Решение

$$\text{Заметим, что: } 1 - b_i = 1 - \frac{1}{1+\exp(-a)} = \frac{\exp(-a)}{1+\exp(-a)} = \frac{1}{1+\exp(a)}$$

$$\text{Распишем } Q(w) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-x_i^T w)) \cdot [y_i = 1] + \log(1 + \exp(x_i^T w)) \cdot$$

$$[y_i = -1] = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \cdot x_i^T w))$$

$$1) \quad \nabla Q = \frac{1}{n} \sum_{i=1}^n -\frac{\exp(-y_i \cdot x_i^T w)}{1+\exp(-y_i \cdot x_i^T w)} \cdot y_i x_i = \frac{1}{n} \sum_{i=1}^n \frac{-y_i}{1+\exp(y_i \cdot x_i^T w)} \cdot x_i =$$

$$\frac{1}{n} \sum_{i=1}^n -y_i \cdot \sigma(-y_i \cdot x_i^T w) \cdot x_i$$

$$\nabla^2 Q = \left(Q''_{w_k, w_p} \right)_{k,p=1}^d \rightarrow \left(\frac{1}{n} \sum_{i=1}^n -\frac{\exp(-y_i x_i^T w)}{1+\exp(-y_i x_i^T w)} \cdot y_i x_i^k \right)'_{w_p} =$$

$$\left(\frac{1}{n} \sum_{i=1}^n \frac{-y_i x_i^k}{1 + \exp(y_i x_i^T w)} \right)'_{w_p} = \frac{1}{n} \sum_{i=1}^n \frac{\exp(y_i x_i^T w)}{\underbrace{(1 + \exp(y_i x_i^T w))^2}_{\sigma(-y_i x_i^T w) \cdot (1 - \sigma(-y_i x_i^T w))}} \underbrace{y_i^2}_1 x_i^k x_i^p$$

$$\text{Итак: } \nabla^2 Q = \left(\frac{1}{n} \sum_{i=1}^n \sigma(-y_i x_i^T w) \cdot (1 - \sigma(-y_i x_i^T w)) x_i^k x_i^p \right)_{k,p=1}^d =$$

$$\boxed{\frac{1}{n} \sum_{i=1}^n \sigma(-y_i x_i^T w) \cdot (1 - \sigma(-y_i x_i^T w)) x_i x_i^T}$$

$$2) \nabla Q(0) = -\frac{1}{2n} \sum_{i=1}^n y_i x_i, \quad \nabla^2 Q(0) = \frac{1}{4n} \sum_{i=1}^n x_i x_i^T$$

$$3) Q(w) = Q(0) + \nabla Q(0)^T w + \frac{1}{2} w^T \nabla^2 Q(0) w + o(\|w\|)$$

$$\text{Итак: } Q(w) = \log 2 - \frac{1}{2n} \sum_{i=1}^n y_i x_i^T w + \frac{1}{8n} \sum_{i=1}^n w^T x_i x_i^T w + o(\|w\|)$$

$$4) Q(w) = Q(0) + \nabla Q(0)^T w + \frac{1}{2} w^T \nabla^2 Q(0) w \rightarrow \min_w$$

$$\nabla_w Q(w) = \nabla Q(0) + \nabla^2 Q(0) w = 0 \rightarrow w^* = -(\nabla^2 Q)^{-1} \nabla Q(0)$$

$$w^* = \boxed{2 \left(\sum_{i=1}^n x_i^T x_i \right)^{-1} \cdot \sum_{i=1}^n y_i x_i}$$

0.6 Метрики качества

0.6.1 Задача 1

Дана выборка из 8 объектов и дан классификатор, предсказывающий вероятность принадлежности объекта положительному классу $b(x_i)(x) = p(y_i = 1|x_i)$

y_i	$b(x_i)$
1	0.1
1	0.8
-1	0.2
-1	0.25
1	0.9
1	0.3
-1	0.6
1	0.95

1. Постройте ROC кривую
2. Посчитайте AUC-ROC
3. Постройте PR кривую (точность-полнота)
4. Посчитайте площадь под PR кривой

Решение

Confusion matrix:

	y_+	y_-
\hat{y}_+	a	b
\hat{y}_-	c	d

$$TP = a, \quad FP = b, \quad TN = d, \quad FN = c$$

1) ROC кривая отражает график функции $TPR(FRP)$

$$FPR = \frac{FP}{|y_-|} = \frac{b}{b+d} - \text{False Positive Rate}$$

$$TPR = \frac{TP}{|y_+|} = \frac{a}{a+c} - \text{True Positive Rate}$$

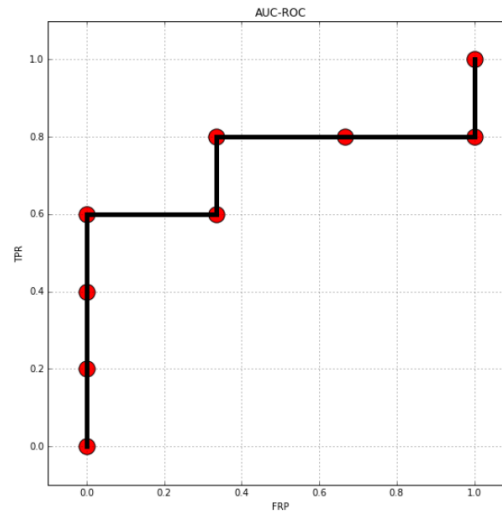
Отсортируем все значения по убыванию функционала $b_i(x)$:

$(b_i, y_i) : (0.1, 1), (0.2, -1), (0.25, -1), (0.3, 1), (0.6, -1), (0.8, 1), (0.9, 1), (0.95, 1)$

$$|y_-| = 3, \quad |y_+| = 5$$

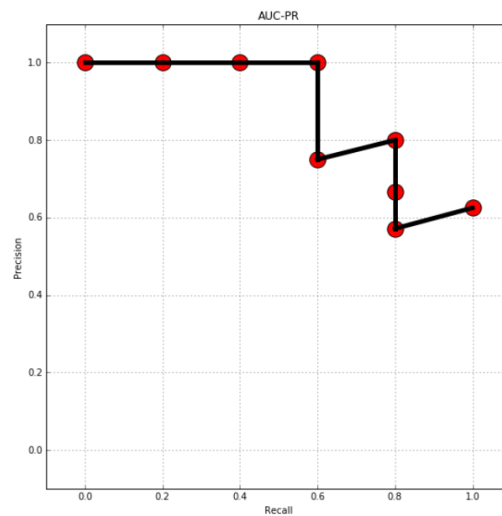
Будем двигаться справа налево и считать значения TPR и FPR

y_i	1		-1		-1		1		-1		1		1		1	
TRP FRP	1 1	$\frac{4}{5}$ 1	-	$\frac{4}{5}$ $\frac{2}{5}$	-	$\frac{4}{5}$ $\frac{1}{5}$	-	$\frac{3}{5}$ $\frac{1}{5}$	-	$\frac{3}{5}$ 0	-	$\frac{2}{5}$ 0	-	$\frac{1}{5}$ 0	-	0 0
Pr Rec	$\frac{5}{8}$ 1	$\frac{4}{7}$ $\frac{4}{5}$	-	$\frac{3}{6}$ $\frac{3}{5}$	-	$\frac{2}{5}$ $\frac{2}{5}$	-	$\frac{1}{2}$ $\frac{2}{5}$	-	$\frac{1}{3}$ $\frac{1}{5}$	-	$\frac{1}{2}$ $\frac{1}{5}$	-	$\frac{1}{1}$ $\frac{1}{5}$	-	0 0



$$2) \text{ AUC-ROC} = \frac{3}{5} \cdot \frac{1}{3} + \frac{4}{5} \cdot \frac{2}{3} = \frac{11}{15} \approx 0.73$$

$$3) \text{ Precision} = \frac{TP}{|\hat{y}_+|} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{|y_+|} = \frac{TP}{TP+FN}$$



4) AUC-PR ≈ 0.87

0.6.2 Задача 2

Рассмотрим функционал вида: $f^{-1} \left(\frac{1}{n} \sum_{i=1}^n f(x_i) \right)$

1. Подберите функцию $f(x)$ так, чтобы получились: среднее арифметическое, гармоническое и геометрическое.
2. При использовании какого среднего в качестве меры качества классификации будут выходить самые качественные и некачественные прогнозы

Решение

1. Среднее арифметическое: $f(x) = x$, $f^{-1}(x) = x$

$$f^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \right) = \frac{x_1 + \dots + x_n}{n}$$

2. Среднее геометрическое: $f(x) = \ln x$, $f^{-1}(x) = e^x$

$$f^{-1} \left(\frac{1}{n} \sum_{i=1}^n \ln x_i \right) = e^{\frac{1}{n} \ln(x_1 \dots x_n)} = \sqrt[n]{x_1 \cdot \dots \cdot x_n}$$

3. Среднее гармоническое: $f(x) = \frac{1}{x}$, $f^{-1}(x) = \frac{1}{x}$

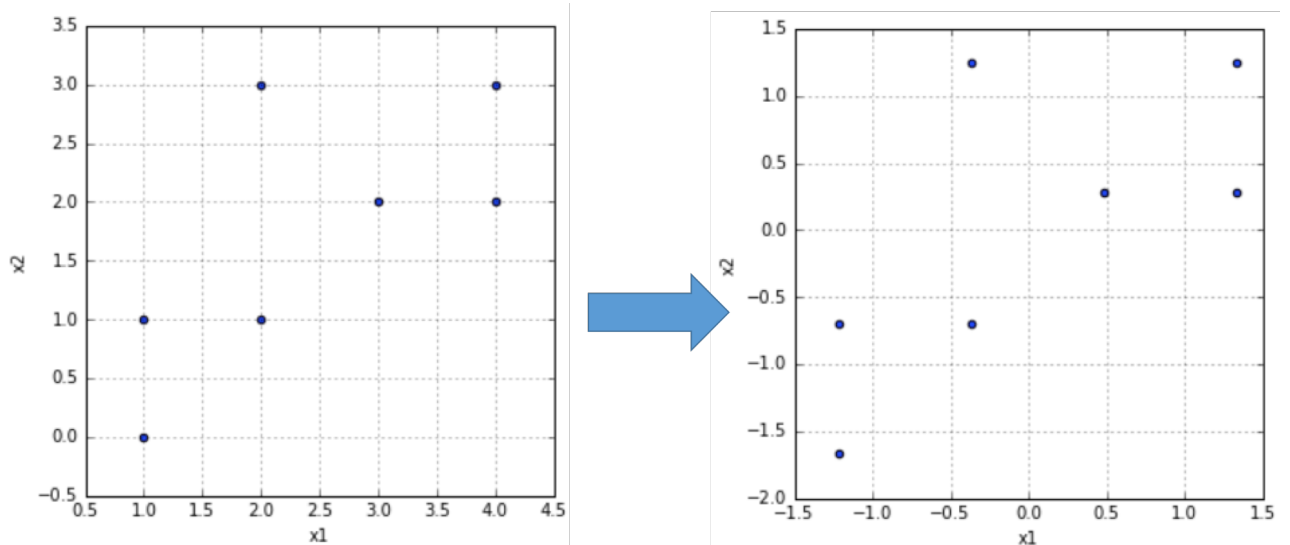
$$f^{-1} \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right) = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$

В вопросе про качество прогноза по сути требуется вспомнить классическое неравенство о средних, которое выглядит так: $x_{garm} \leq x_{geom} \leq x_{arithm}$.

0.7 PCA

0.7.1 Задача 1

Найти две главные компоненты для набора данных: $X = \{(1, 0), (1, 1), (2, 3), (4, 2), (4, 3), (3, 2), (2, 1)\}$



Решение

Перед непосредственным решением задачи данные нужно нормировать и центрировать $X \rightarrow X'$.

$$\begin{aligned}\mathbb{E}x^{(1)} &= 2.43, & \sigma(x^{(1)}) &= 1.18 \\ \mathbb{E}x^{(2)} &= 1.71, & \sigma(x^{(2)}) &= 1\end{aligned}$$

$$X' = \{x | x = \frac{x - \mathbb{E}x}{\sigma(x)}\} = \begin{bmatrix} -1.21 & -1.66 \\ -1.21 & -0.69 \\ -0.36 & 1.25 \\ 1.33 & 0.28 \\ 1.33 & 1.25 \\ 0.49 & 0.28 \\ -0.36 & -0.69 \end{bmatrix}$$

Задача покомпонентного нахождения главных компонент ставится следующим образом:

$$\begin{cases} \|Xa\|_2^2 \rightarrow \max_a, \text{ где } a - \text{первая главная компонента} \\ \|a\|_2^2 = 1 \end{cases}$$

$$L = a^T X^T X a + \lambda(a^T a - 1)$$

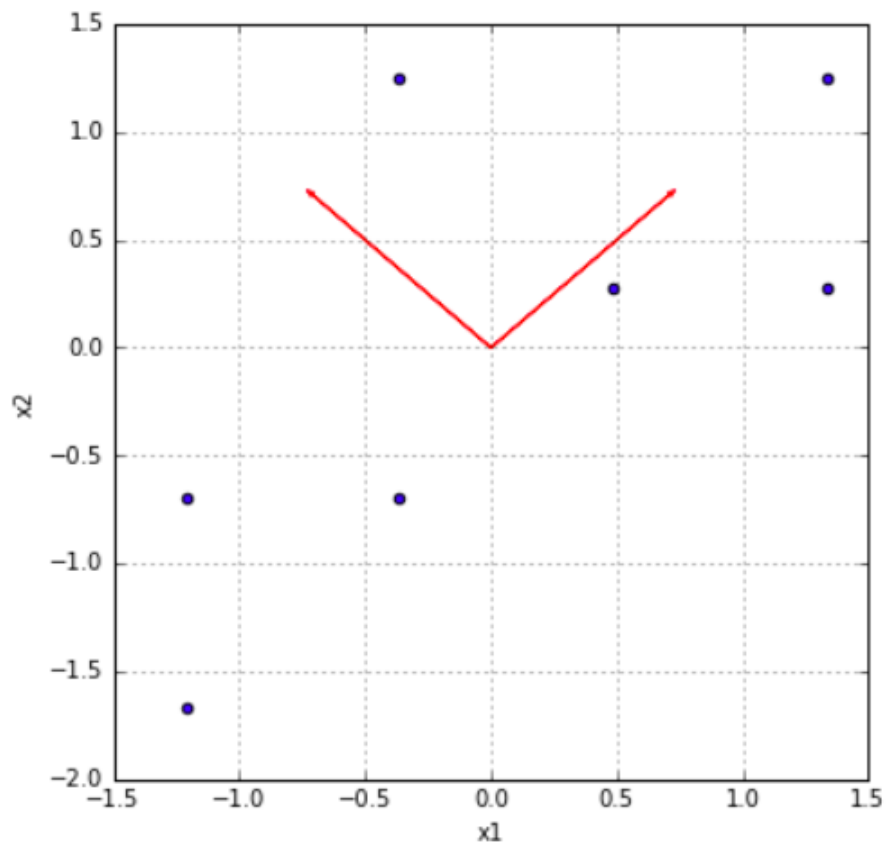
$\nabla_a L = 2X^T X a + 2\lambda a = 0 \Rightarrow X^T X a = -\lambda a \Rightarrow$ решением этого уравнения является собственный вектор, отвечающий наибольшему (так как мы ищем максимум) собственному значению матрицы $X^T X$

Не сложно показать, что остальные главные компоненты это будут собственные вектора матрицы $X^T X$, следовательно нам нужно найти спектральное разложение матрицы $X^T X$.

Так как матрица $X^T X$ – симметричная, то для неё существует спектральное разложение: $X^T X = UDU^T$

Итак: $X^T X = \begin{bmatrix} 7 & 4.83 \\ 4.83 & 7 \end{bmatrix} = \begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix} \cdot \begin{bmatrix} 11.83 & 0 \\ 0 & 2.17 \end{bmatrix} \cdot \begin{bmatrix} 0.71 & 0.71 \\ -0.71 & 0.71 \end{bmatrix}$
(с точностью до округлений).

Тогда $a_1 = (0.71, 0.71), \quad a_2 = (-0.71, 0.71)$

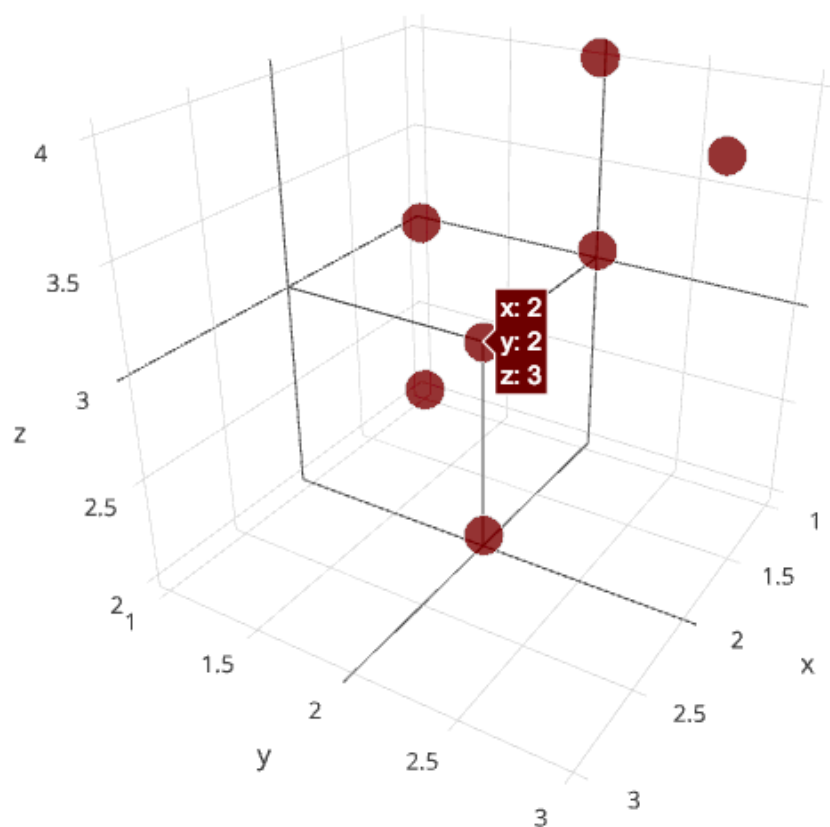


0.7.2 Задача 2

Изобразить данные в проекции первых двух главных компонент

x_1	x_2	x_3
1	1	2
1	1	3
1	2	4
2	3	4
2	2	3
2	2	2
3	3	4

Решение



- 1) Нормируем и центрируем данные: $X \rightarrow X'$, где X' :
- 2) Найдём ковариационную матрицу для данного набора: $X'^T X'$ и представим её в виде собственного разложения (eigendecomposition):
$$X'^T X' = S \cdot D \cdot S^T$$

```

[-1.02062073, -1.32287566, -1.37198868]
[-1.02062073, -1.32287566, -0.17149859]
[-1.02062073, 0., 1.02899151]
[ 0.40824829, 1.32287566, 1.02899151]
[ 0.40824829, 0., -0.17149859]
[ 0.40824829, 0., -1.37198868]
[ 1.83711731, 1.32287566, 1.02899151]

```

Рис. 1: X'

$$\begin{aligned}
X'^T X' &= \begin{bmatrix} 7. & 5.67064811, & 2.20544113 \\ 5.67064811, & 7. & 4.76429737 \\ 2.20544113, & 4.76429737, & 7. \end{bmatrix} \\
&= \\
&\begin{matrix} \begin{bmatrix} -0.56214007, & -0.62809246, & 0.53805055 \\ -0.65322416, & -0.06182077, & -0.75463659 \\ -0.50724426, & 0.77567909, & 0.37553324 \end{bmatrix} & \begin{bmatrix} 15.57953725, & 0. & 0. \\ 0. & 4.83447369, & 0. \\ 0. & 0. & 0.58598906 \end{bmatrix} & \begin{bmatrix} -0.56214007, & -0.65322416, & -0.50724426 \\ -0.62809246, & -0.06182077, & 0.77567909 \\ 0.53805055, & -0.75463659, & 0.37553324 \end{bmatrix} \\
S & & D & S^T
\end{matrix}
\end{aligned}$$

Рис. 2: $X'^T X' = S \cdot D \cdot S^T$

Первые 2 столбца матрицы S это и есть первые 2 главные компоненты.

Заметим, что если бы мы просто хотели представить данные в новом базисе из собственных векторов, то тогда любой вектор x мог бы быть представлен, как: $x = \begin{bmatrix} | & | & | \\ s_1 & | & s_2 & | & s_3 \\ | & | & | \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$, где s_1, s_2, s_3 – собственные вектора, а α_i – и будут как раз координатами вектора x в новом базисе.

Но у нас задача, представить данные, используя только 2 главные компоненты. Тогда, вполне может оказаться так, что для какого-то вектора y из исходного пространства (в данном случае \mathbb{R}^3) не будет выполняться равенство: $y = \alpha_1 \cdot s_1 + \alpha_2 \cdot s_2$.

Таким образом, переходим к задаче: найти проекцию вектора x на первые 2 главные компоненты, то есть: $\|x - \alpha_1 \cdot s_1 - \alpha_2 \cdot s_2\|_2 \rightarrow \min_{\alpha_i}$

Её можно представить в следующем эквивалентном виде:

$\|x - S \cdot \alpha\|_2^2 \rightarrow \min_{\alpha}$, где S – матрица, по столбцам которой записаны

собственные вектора, α – вектор, отвечающий за координаты вектора x в базисе s_j

$$\|x - S \cdot \alpha\|_2^2 = (x^T - \alpha^T S^T)(x - S\alpha) = x^T x - 2x^T S\alpha + \alpha^T S^T S\alpha \rightarrow \min_{\alpha}$$

$$\nabla_{\alpha} = -2S^T x + 2S^T S\alpha = 0 \Rightarrow \alpha = (S^T S)^{-1} S^T x$$

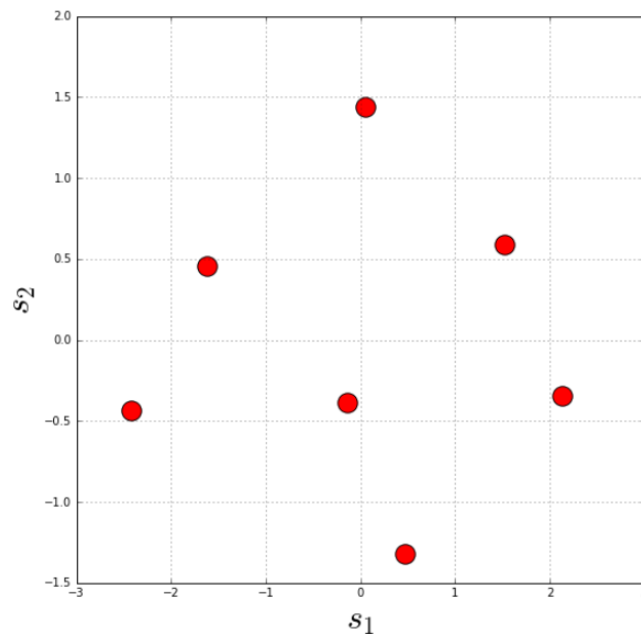
Заметим, что матрица S – ортогональная, а значит: $S^T S = I$ и следовательно: $\alpha = S^T x$

Итого получаем: $X'_{projection} = S^T \cdot X'^T$:

$$\begin{bmatrix} -0.56214007, & -0.65322416, & -0.50724426 \\ -0.62809246, & -0.06182077, & 0.77567909 \end{bmatrix} \cdot \begin{bmatrix} -1.0206 & -1.0206 & -1.0206 & 0.4082 & 0.4082 & 0.4082 & 1.8371 \\ -1.3229 & -1.3229 & 0. & 1.3229 & 0. & 0. & 1.3229 \\ -1.372 & -0.1715 & 1.029 & 1.029 & -0.1715 & -1.372 & 1.029 \end{bmatrix} =$$

$$\begin{bmatrix} 2.13379953, & -0.34139756 \\ 1.52485782, & 0.58979751 \\ 0.05178178, & 1.43921138 \\ -1.6155771, & 0.45996834 \\ -0.14250105, & -0.38944554 \\ 0.46644065, & -1.32064061 \\ -2.41880163, & -0.43749353 \end{bmatrix}$$

Рис. 3:



0.7.3 Задача 3 (Вероятностная постановка РСА)

Имеется набор точек $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^D$, распределённых по нормальному закону распределения: $x_i \in \mathcal{N}(x_i|A \cdot z_i + b, \sigma^2 \cdot I)$, $A \in \mathbb{R}^{D \times d}$, $z_i \in \mathbb{R}^d$, $b \in \mathbb{R}^D$. Требуется найти оценки параметров A , z_i , b методом максимального правдоподобия.

Решение

$$\theta = \{A, z_i, b\}$$

$$L = \prod_{i=1}^n p(x_i|\theta) = \prod_{i=1}^n \frac{1}{(2\pi)^{\frac{D}{2}} \cdot \sigma^D} \cdot \exp\left\{-\frac{1}{2\sigma^2}(x_i - Az_i - b)^T(x_i - Az_i - b)\right\} \rightarrow \max_{\theta}$$

$$\log L = -\frac{Dn}{2} \log 2\pi - D \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^T x_i - 2x_i^T A z_i - 2x_i^T b + z_i^T A^T A z_i + 2b^T A z_i + b^T b)$$

$$\{\nabla_X a^T X b = ab^T, \quad \nabla_X a^T X^T X b = Xab^T + Xba^T\}$$

$$\nabla_{z_i} \log L = -2A^T x_i + 2A^T A z_i + 2A^T b = 0 \Rightarrow \boxed{z_i = (A^T A)^{-1} A^T (x_i - b)}$$

$$\nabla_b \log L = \sum_{i=1}^n (-2x_i + 2A z_i + 2b) = 0 \Rightarrow \boxed{b = \frac{1}{n} \sum_{i=1}^n (A z_i - x_i)}$$

$$\nabla_A \log L = \sum_{i=1}^n (-2x_i z_i^T + 2A z_i z_i^T + 2b z_i^T) = 0 \Rightarrow$$

$$\Rightarrow \boxed{A = \left(\sum_{i=1}^n (x_i - b) z_i^T \right) \left(\sum_{i=1}^n z_i z_i^T \right)^{-1}}$$

0.8 ЕМ алгоритм

Теория:

Е - шаг: оцениваем $\mathbb{P}(Z|X, \Theta)$

М - шаг: $\Theta^* = \arg \max_{\Theta} \mathbb{E}_{Z \sim \mathbb{P}(Z|X, \Theta^{old})} \log \mathbb{P}(X, Z|\Theta)$

0.8.1 Задача 1

Пусть у нас имеется 2 монеты A, B , мы выбираем случайно одну монету, делаем 10 бросков (записываем результат), затем случайно выбираем еще монету (возможно опять первую монету), снова делаем 10 бросков. Результаты представлены на рисунке 1. Вопрос, который перед нами стоит: возможно ли лишь по результатам бросков определить, какая это была монета и с какой вероятностью?

a Maximum likelihood

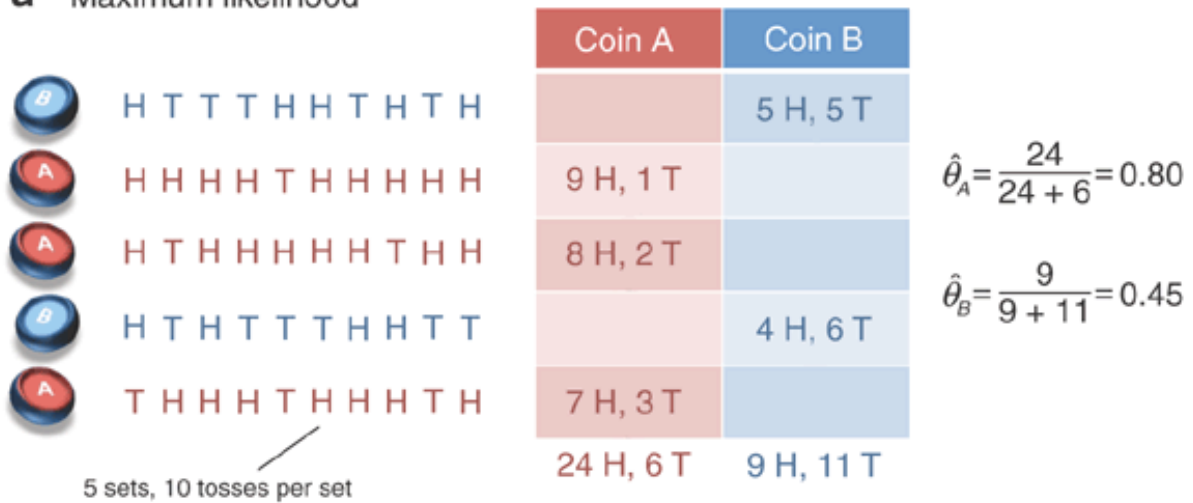


Рис. 4: Пример определения монеты

Решение

Введём параметры: θ_A, θ_B – вероятности выпадения орла у монеты A и B соответственно. Инициализируем каким-то образом эти вероятности, например, равновероятно: $\theta_A^{(0)} = 0.6, \theta_B^{(0)} = 0.5$

Н Т Т Т Н Н Т Н Т Н
Н Н Н Н Т Н Н Н Н Н
Н Т Н Н Н Н Н Т Н Н
Н Т Н Т Т Т Н Н Т Т
Т Н Н Н Т Н Н Н Т Н

Рис. 5: Имеем данные

Посчитаем с какой вероятностью реализуется последовательность 1, если это монета А:

$$p_A(5 \text{ орлов}, 5 \text{ решек}) = C_{10}^5 0.6^5 0.4^5 = 0.2$$

$$p_B(5 \text{ орлов}, 5 \text{ решек}) = C_{10}^5 \theta_A^5 (1 - \theta_A)^5 = C_{10}^5 0.5^{10} = 0.246$$

$$\text{Отсюда: } p_A(1) = \frac{0.2}{0.2+0.246} = 0.45, \quad p_B(1) = \frac{0.246}{0.2+0.246} = 0.55$$

Для второй последовательности:

$$p_A(9 \text{ орлов}, 1 \text{ решка}) = C_{10}^9 0.6^9 0.4^1 = 0.04$$

$$p_B(9 \text{ орлов}, 1 \text{ решка}) = C_{10}^9 0.5^9 0.5^1 = 0.0098$$

$$\text{Отсюда: } p_A(2) = \frac{0.04}{0.04+0.0098} = 0.804, \quad p_B(2) = \frac{0.0098}{0.04+0.0098} = 0.195$$

Для третьей последовательности:

$$p_A(8 \text{ орлов}, 2 \text{ решки}) = C_{10}^8 0.6^8 0.4^2 = 0.121$$

$$p_B(8 \text{ орлов}, 2 \text{ решки}) = C_{10}^8 0.5^8 0.5^2 = 0.044$$

$$\text{Отсюда: } p_A(3) = \frac{0.121}{0.121+0.044} = 0.73, \quad p_B(3) = \frac{0.044}{0.044+0.121} = 0.27$$

$$\text{Для 4ой строки получаем: } p_A(4) = 0.35, \quad p_B(4) = 0.65$$

$$\text{Для 5ой строки получаем: } p_A(5) = 0.65, \quad p_B(5) = 0.35$$

После того, как мы посчитали вероятности для каждой строки оказаться полученной из каждой монеты: посчитаем наивероятнейшее число орлов и

решек для каждой строки и для каждой монеты:

Строка	Монета А (орёл, решка)	Монета В, (орёл, решка)
1	$(0.45 \cdot 5 = 2.25, 0.55 \cdot 5 = 2.75)$	$(0.55 \cdot 5 = 2.75, 0.45 \cdot 5 = 2.25)$
2	$(0.804 \cdot 9 = 7.24, 0.196 \cdot 1 = 0.196)$	$(0.195 \cdot 9 = 1.755, 0.81 \cdot 1 = 0.81)$
3	$(0.73 \cdot 8 = 5.84, 0.27 \cdot 2 = 0.54)$	$(0.27 \cdot 8 = 2.16, 0.73 \cdot 2 = 1.46)$
4	$(0.35 \cdot 4 = 1.4, 0.65 \cdot 6 = 3.9)$	$(0.65 \cdot 4 = 2.6, 0.35 \cdot 6 = 2.1)$
5	$(0.65 \cdot 7 = 4.55, 0.35 \cdot 3 = 1.05)$	$(0.35 \cdot 7 = 2.45, 0.65 \cdot 3 = 1.95)$
Сумма	(21.28, 8.44)	(11.715, 8.57)

Получаем новые оценки $\theta_A^{(1)} = \frac{21.28}{21.28+8.44} = 0.72$, $\theta_B^{(1)} = \frac{11.715}{11.715+8.57} = 0.58$

Теперь опять рассчитываем вероятности для каждой строки, но уже используя новые оценки вероятностей.

После 5ти итераций такого алгоритма мы получим оценки: $\theta_A^{(5)} \approx 0.82$, $\theta_B^{(5)} \approx 0.52$

Теперь мы можем понять, с какой вероятностью каждая строка из какой монеты получилась, например для первой строки:

$$p_A(5 \text{ орлов}, 5 \text{ решек}) = C_{10}^5 0.82^5 0.18^5 = 0.0177$$

$$p_B(5 \text{ орлов}, 5 \text{ решек}) = C_{10}^5 0.52^5 0.48^5 = 0.244$$

$$\text{Итак: } P(\text{первая строка из монеты А}) = \frac{0.0177}{0.0177+0.244} = 0.07$$

$$P(\text{первая строка из монеты В}) = \frac{0.244}{0.0177+0.244} = 0.93$$

Значит, скорее всего первая строка получилась в результате подбрасывания монеты В

0.8.2 Задача 2

Пусть x_1, \dots, x_N – независимая выборка из смеси $p(x) = \gamma \cdot p_1(x) + (1 - \gamma) \cdot p_2(x)$, где:

x	1	2	3
$p_1(x)$	α	$1 - \alpha$	0

x	1	2	3
$p_2(x)$	0	$1 - \beta$	β

Выборка \mathcal{X} состоит из 30 единиц, 20 двоек и 60 троек. Провести первые 2 итерации ЕМ-алгоритма из начального приближения $\gamma^{(0)} = \alpha^{(0)} = \beta^{(0)} = \frac{1}{2}$

Решение:

$p_1(x)$ можно представить как $\alpha^{[x_i=1]} \cdot (1 - \alpha)^{[x_i=2]}$

$p_2(x)$ можно представить как $\beta^{[x_i=3]} \cdot (1 - \beta)^{[x_i=2]}$

$$\mathbb{P}(\mathbb{X}, \mathbb{Z} | \Theta) = \prod_{i=1}^N [\alpha^{[x_i=1]} \cdot (1 - \alpha)^{[x_i=2]}]^{z_i} \cdot [\beta^{[x_i=3]} \cdot (1 - \beta)^{[x_i=2]}]^{1-z_i}$$

Е-шаг:

Оцениваем $\mathbb{P}(\mathbb{Z} | \mathbb{X}, \Theta) = \prod_{i=1}^N \mathbb{P}(z_i | x_i, \theta)$, где $\theta = \{\gamma, \alpha, \beta\}$

$$\begin{aligned} \mathbb{P}(z_i = 1 | x_i, \theta) &= \frac{\mathbb{P}(z_i=1) \cdot \mathbb{P}(x_i | z_i=1, \theta)}{\sum_{k=0}^1 \mathbb{P}(z_i=k) \cdot \mathbb{P}(x_i | z_i=k, \theta)} = \{\text{в силу формулы Байеса}\} = \\ &= \frac{\gamma \cdot \alpha^{[x_i=1]} \cdot (1 - \alpha)^{[x_i=2]}}{\underbrace{\gamma \cdot \alpha^{[x_i=1]} \cdot (1 - \alpha)^{[x_i=2]} + (1 - \gamma) \cdot \beta^{[x_i=3]} \cdot (1 - \beta)^{[x_i=2]}}_{g_{i1}}} \\ \mathbb{P}(z_i = 0 | x_i, \theta) &= \frac{(1 - \gamma) \cdot \beta^{[x_i=3]} \cdot (1 - \beta)^{[x_i=2]}}{\underbrace{(1 - \gamma) \cdot \beta^{[x_i=3]} \cdot (1 - \beta)^{[x_i=2]} + \gamma \cdot \alpha^{[x_i=1]} \cdot (1 - \alpha)^{[x_i=2]}}_{g_{i0}}} \end{aligned}$$

М-шаг:

$$\begin{aligned} Q &= \mathbb{E}_z \log \mathbb{P}(\mathbb{X}, \mathbb{Z} | \Theta) \rightarrow \max_{\gamma, \alpha, \beta} \\ Q &= \mathbb{E}_z \left(\sum_{i=1}^N z_i \cdot \{\log \gamma + [x_i = 1] \log \alpha + [x_i = 2] \log(1 - \alpha)\} \right) + \\ &+ \mathbb{E}_z \left(\sum_{i=1}^N (1 - z_i) \cdot \{\log(1 - \gamma) + [x_i = 2] \log(1 - \beta) + [x_i = 3] \log \beta\} \right) = \\ &= \{\mathbb{E}_{z_i}(z_i) = \mathbb{P}(z_i = 1 | x_i, \theta) = g_{i1}, \mathbb{E}_{z_i}(1 - z_i) = \mathbb{P}(z_i = 0 | x_i, \theta) = g_{i0}\} = \\ &= \sum_{i=1}^N g_{i1} \cdot \{\log \gamma + [x_i = 1] \log \alpha + [x_i = 2] \log(1 - \alpha)\} + \sum_{i=1}^N g_{i0} \cdot \{\log(1 - \gamma) + [x_i = \\ &2] \log(1 - \beta) + [x_i = 3] \log \beta\} \rightarrow \max_{\alpha, \beta, \gamma} \end{aligned}$$

Итак:

$$Q'_\gamma = \sum_{i=1}^N \frac{g_{i1}}{\gamma} - \frac{g_{i0}}{1-\gamma} = \frac{1}{\gamma(1-\gamma)} \cdot \sum_{i=1}^N (g_{i1} - \gamma(g_{i1} + g_{i0})) = 0 \Rightarrow \hat{\gamma} = \frac{\sum_{i=1}^N g_{i1}}{\sum_{i=1}^N \underbrace{g_{i1} + g_{i0}}_1} \Rightarrow$$

$$\hat{\gamma} = \frac{1}{N} \sum_{i=1}^N g_{i1}$$

$$Q'_\alpha = \sum_{i=1}^N g_{i1} \cdot \left\{ \frac{[x_i=1]}{\alpha} - \frac{[x_i=2]}{1-\alpha} \right\} = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^N g_{i1} \cdot$$

$$([x_i = 1] - \alpha([x_i = 1] + [x_i = 2])) = 0 \Rightarrow \hat{\alpha} = \frac{\sum_{i=1}^N g_{i1} \cdot [x_i=1]}{\sum_{i=1}^N g_{i1} \cdot ([x_i=1] + [x_i=2])} \Rightarrow$$

$$\hat{\alpha} = \frac{\sum_{x_i=1} g_{i1}}{\sum_{x_i=1} g_{i1} + \sum_{x_i=2} g_{i1}}$$

$$Q'_\beta = \sum_{i=1}^N g_{i0} \cdot \left\{ \frac{[x_i=3]}{\beta} - \frac{[x_i=2]}{1-\beta} \right\} = \frac{1}{\beta(1-\beta)} \sum_{i=1}^N g_{i0} \cdot$$

$$([x_i = 3] - \beta([x_i = 3] + [x_i = 2])) = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^N g_{i0} \cdot [x_i=3]}{\sum_{i=1}^N g_{i0} \cdot ([x_i=3] + [x_i=2])} \Rightarrow$$

$$\hat{\beta} = \frac{\sum_{x_i=3} g_{i0}}{\sum_{x_i=2} g_{i0} + \sum_{x_i=3} g_{i0}}$$

После того, как формулы для Е и М шагов получены – осталось проделать 2 шага ЕМ алгоритма. Для этого, выразим необходимые суммы через исходные параметры:

$$\sum_{i=1}^N g_{i1} = \underbrace{\sum_{x_i=1} g_{i1}}_a + \underbrace{\sum_{x_i=2} g_{i1}}_b + \underbrace{\sum_{x_i=3} g_{i1}}_c$$

$$a = 30 \cdot \frac{\gamma\alpha}{\gamma\alpha + (1-\gamma)}$$

$$b = 20 \cdot \frac{\gamma(1-\alpha)}{\gamma(1-\alpha) + (1-\gamma)(1-\beta)} = \frac{20\gamma(1-\alpha)}{\gamma(\beta-\alpha) + 1-\beta}$$

$$c = 60 \cdot \frac{\gamma}{\gamma + (1-\gamma)\beta}$$

$$a|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}} = \frac{30 \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{4} + \frac{1}{2}} = 10$$

$$b|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}} = \frac{20 \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}} = 10$$

$$c|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}} = \frac{60 \cdot \frac{1}{2}}{\frac{1}{2} + \frac{1}{4}} = 40$$

$$\text{Итак: } \hat{\gamma}^{(1)} = \frac{1}{110} \cdot (10 + 10 + 40) = \frac{6}{11}$$

$$\hat{\alpha}^{(1)} = \frac{a|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}}}{a|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}} + b|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}}} = \frac{10}{10+10} = \frac{1}{2}$$

$$\sum_{i=1}^N g_{i0} = \underbrace{\sum_{x_i=1} g_{i0}}_d + \underbrace{\sum_{x_i=2} g_{i0}}_e + \underbrace{\sum_{x_i=3} g_{i0}}_f$$

$$e = 20 \cdot \frac{(1-\gamma)(1-\beta)}{(1-\gamma)(1-\beta) + \gamma(1-\alpha)} = 20 \cdot \frac{(1-\gamma)(1-\beta)}{\gamma(\beta-\alpha) + 1 - \beta}$$

$$f = 60 \cdot \frac{(1-\gamma)\beta}{(1-\gamma)\beta + \gamma}$$

$$e|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}} = \frac{20 \cdot \frac{1}{4}}{\frac{1}{2}} = 10$$

$$f|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}} = 60 \cdot \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = 20$$

$$\text{Итак: } \hat{\beta}^{(1)} = \frac{f|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}}}{f|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}} + e|_{\gamma^{(0)}, \alpha^{(0)}, \beta^{(0)}}} = \frac{20}{20+10} = \frac{2}{3}$$

После первой итерации:

$$\boxed{\hat{\gamma}^{(1)} = \frac{6}{11}, \alpha^{(1)} = \frac{1}{2}, \beta^{(1)} = \frac{2}{3}}$$

Вторая итерация:

Пересчет формул Е-шага:

$$\sum_{i=1}^N g_{i1} = \underbrace{\sum_{x_i=1} g_{i1}}_a + \underbrace{\sum_{x_i=2} g_{i1}}_b + \underbrace{\sum_{x_i=3} g_{i1}}_c$$

$$a = 30 \cdot \frac{\gamma\alpha}{\gamma\alpha + (1-\gamma)}$$

$$b = 20 \cdot \frac{\gamma(1-\alpha)}{\gamma(1-\alpha) + (1-\gamma)(1-\beta)} = \frac{20\gamma(1-\alpha)}{\gamma(\beta-\alpha) + 1 - \beta}$$

$$c = 60 \cdot \frac{\gamma}{\gamma + (1-\gamma)\beta}$$

$$a|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}} = \frac{30 \cdot \frac{6}{11} \cdot \frac{1}{2}}{\frac{6}{11} \cdot \frac{1}{2} + \frac{5}{11}} = \frac{45}{4}$$

$$b|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}} = \frac{20 \cdot \frac{6}{11} \cdot \frac{1}{2}}{\frac{6}{11} \cdot (\frac{2}{3} - \frac{1}{2}) + \frac{1}{3}} = \frac{90}{7}$$

$$c|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}} = 60 \cdot \frac{\frac{6}{11}}{\frac{6}{11} + \frac{5}{11} \cdot \frac{2}{3}} = \frac{270}{7}$$

$$\sum_{i=1}^N g_{i0} = \underbrace{\sum_{x_i=1} g_{i0}}_d + \underbrace{\sum_{x_i=2} g_{i0}}_e + \underbrace{\sum_{x_i=3} g_{i0}}_f$$

$$e = 20 \cdot \frac{(1-\gamma)(1-\beta)}{(1-\gamma)(1-\beta) + \gamma(1-\alpha)} = 20 \cdot \frac{(1-\gamma)(1-\beta)}{\gamma(\beta-\alpha) + 1 - \beta}$$

$$f = 60 \cdot \frac{(1-\gamma)\beta}{(1-\gamma)\beta + \gamma}$$

$$e|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}} = \frac{20 \cdot \frac{5}{11} \cdot \frac{1}{3}}{\frac{6}{11} \cdot \frac{1}{6} + \frac{1}{3}} = \frac{50}{7}$$

$$f|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}} = \frac{60 \cdot \frac{5}{11} \cdot \frac{2}{3}}{\frac{5}{11} \cdot \frac{2}{3} + \frac{6}{11}} = \frac{150}{7}$$

Пересчёт формул М-шага:

$$\hat{\gamma}^{(2)} = \frac{1}{110} \cdot \left(\frac{45}{4} + \frac{90}{7} + \frac{270}{7} \right) = \frac{1755}{28 \cdot 110} \approx 0.5698$$

$$\hat{\alpha}^{(2)} = \frac{a|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}}}{a|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}} + b|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}}} = \frac{\frac{45}{4}}{\frac{45}{4} + \frac{90}{7}} = \frac{63}{135}$$

$$\hat{\beta}^{(2)} = \frac{e|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}}}{e|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}} + f|_{\gamma^{(1)}, \alpha^{(1)}, \beta^{(1)}}} = \frac{\frac{50}{7}}{\frac{50}{7} + \frac{150}{7}} = \frac{1}{4}$$

После второй итерации:

$$\boxed{\gamma^{(2)} = \frac{1755}{3080}, \hat{\alpha}^{(2)} = \frac{63}{135}, \hat{\beta}^{(2)} = \frac{1}{4}}$$