Машинное обучение

Антон Андрейцев

Содержание

0.1 Оптимизация

0.1.1 Задача 1

Показать, что минимизация эмпирического риска с l_2 – регуляризацией эквивалентна раннему останову в градиентном спуске.

Решение

Разложим $L(w)$ по формуле Тейлора в окрестности точки оптимума w^*

$L(w) = L(w^*) + \frac{1}{2}(w - w^*)H(w - w^*) + o(\|w\|^2)$, H – гессиан функции $L(w)$ в точке w^*

$\nabla_w L(w) = H(w - w^*)$

$w^t = w^{t-1} - \epsilon \cdot \nabla_w L(w) = w^{t-1} - \epsilon \cdot H(w^{t-1} - w^*)$

$w^t - w^* = (I - \epsilon \cdot H)(w^{t-1} - w^*)$

Так как H – матрица гессиан, то она неотрицательно определена (для выпуклой функции), а значит её можно представить через спектральное разложение: $H = Q\Lambda Q^T$, где Q – матрица из ортогональных столбцов, Λ – диагональная матрицая.

$w^t - w^* = (I - \epsilon \cdot Q\Lambda Q^T)(w^{t-1} - w^*)$

$Q^T(w^t - w^*) = (I - \epsilon \cdot \Lambda)Q^T(w^{t-1} - w^*)$

Рассмотрим еще одну итерацию градиентного спуска:

$Q^T(w^{t+1} - w^*) = (I - \epsilon \cdot \Lambda)Q^T(w^t - w^*) = (I - \epsilon \cdot \Lambda)^2 Q^T(w^{t-1} - w^*)$

Тогда, если мы стартовали из точки $w^0 = 0$, то через t итераций мы получим веса:

$$Q^T(w^t - w^*) = (I - \epsilon \cdot \Lambda)^t Q^T(0 - w^*) \Rightarrow \boxed{Q^T w^t = \left[I - (I - \epsilon \cdot \Lambda)^t\right] Q^T w^*}$$

В случае регуляризованного функционала:

$L(w) = L(w^*) + \frac{1}{2}(w - w^*)H(w - w^*) + \frac{\alpha}{2}w^T w$

Необходимое условие минимума функции – $\nabla_w L(w) = 0$

$\nabla_w L(w) = H(w - w^*) + \alpha \cdot w = 0 \Rightarrow w^{opt} = (H + \alpha I)^{-1} H w^*$

Отсюда:

$$Q^T w^{opt} = Q^T (Q\Lambda Q^T + \alpha \underbrace{Q^T Q}_{I})^{-1} Q\Lambda Q^T w^* = Q^T Q(\Lambda + \alpha I)^{-1} Q^T Q\Lambda Q^T w^*$$

Итак:

$$\boxed{Q^T w^{opt} = (\Lambda + \alpha I)^{-1} \Lambda Q^T w^*}$$

Пользуясь частным случаем тождества Вудбери: $\left\{(I + AB)^{-1} = I - A(I + BA)^{-1}B\right\}$

$$(\Lambda + \alpha I)^{-1}\Lambda = \left[\Lambda^{-1}(\Lambda + \alpha I)\right]^{-1} = \left(I + \underbrace{\alpha I}_{A} \cdot \underbrace{\Lambda^{-1}}_{B}\right)^{-1} =$$

$I - \alpha(I + \alpha\Lambda^{-1})^{-1}\Lambda^{-1} = I - \alpha(\Lambda + \alpha I)^{-1}$

Итак:

$$\boxed{Q^T w^{opt} = \left[I - \alpha(\Lambda + \alpha I)^{-1}\right] Q^T w^*}$$

Сравнивая $Q^T w^t$ и $Q^T w^{opt}$ получаем:

$(I - \epsilon\Lambda)^t = \alpha(\alpha I + \Lambda)^{-1}$

Откуда и получаем, что при определённом подборе параметра регуляризации α точка оптимума $L(w) + \frac{\alpha}{2}w^T w$ совпадает со значением веса на итерации t.

$t \log(1 - \epsilon \cdot \lambda_i) = -\log(1 + \frac{\lambda_i}{\alpha})$

При длине шага стремящемся к нулю ($\epsilon \to 0$) и параметре регуляризации к бесконечности ($\alpha \to \infty$) можно сделать такую оценку на соотношение количества итераций с параметром регуляризации:

$-t \cdot \epsilon \cdot \lambda_i = -\frac{\lambda_i}{\alpha}$

$t = \frac{1}{\epsilon \cdot \alpha}$ – то есть количество итераций обратно величине регуляризации

0.2 Регрессия

0.2.1 Задача 1

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

Решение

$y|X \sim \mathcal{N}(X\beta, \sigma^2 \cdot I_n) \Rightarrow Likelyhood = p(y|X) \max\limits_{\beta}$

$p(y|X) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \cdot \exp\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\} \to \max\limits_{\beta}$

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

Первые два члена не зависят от $\beta \Rightarrow$
$\Rightarrow \log L \to \max\limits_{\beta} \sim (y - X\beta)^T(y - X\beta) \to \min\limits_{\beta}$

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

0.2.2 Задача 2

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

Решение

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

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

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

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

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

0.2.3 Задача 3

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

Решение

Решение задачи МНК имеет вид: $w = (X^TX)^{-1}X^Ty \rightarrow w = (VD\underbrace{U^TU}_{I}DV^T)^{-1}VDU^Ty = \underbrace{(VD^2V^T)^{-1}}_{VD^{-2}V^T}VDU^Ty = VD^{-2}\underbrace{V^TV}_{I}DU^Ty = VD^{-1}U^Ty$

0.2.4 Задача 4

Докажите, что для линейной регрессии выполнено свойство:
$\sum_{i=1}^{n} \hat{y}_i \cdot (y_i - \hat{y}_i) = 0$
Где $\hat{y} = X \cdot \hat{\alpha}, \quad \hat{\alpha} = (X^TX)^{-1}X^Ty$

Решение

Заметим, что $\sum_{i=1}^{n} \hat{y}_i \cdot (y_i - \hat{y}_i)$ можно переписать в виде $\hat{y}^T(y - \hat{y})$

Тогда: $\hat{y}^T(y - \hat{y}) = \left(X(X^TX)^{-1}X^Ty\right)^T(y - X(X^TX)^{-1}X^Ty) = y^T\left(\underbrace{X(X^TX)^{-1}X^T}_{P}\right)(I - X(X^TX)^{-1}X^T)y$

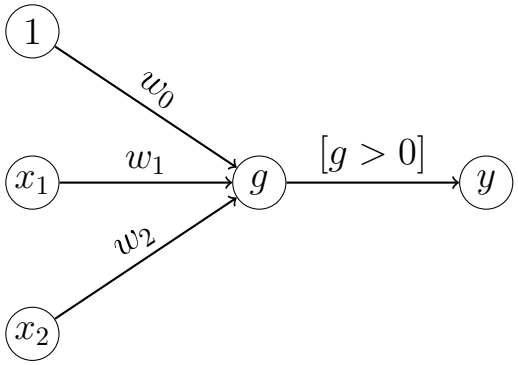
Матрица P – идемпотентная, так как: $P^2 = P$, $X(X^TX)^{-1}X^T \cdot X(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T$

Тогда: $P \cdot (I - P) = P - P^2 = P - P = 0 \Rightarrow y^T0y = 0$
(ч.т.д)

0.3 Perceptron

0.3.1 Задача 1

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



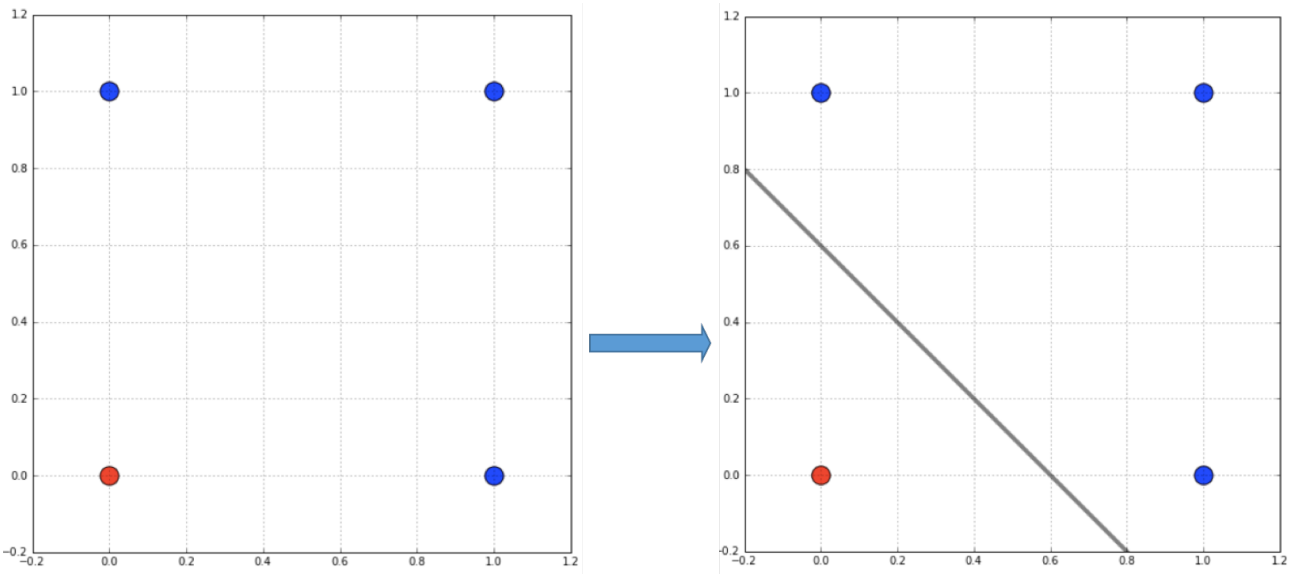
1. Подберите веса персептрона так, чтобы он реализовывал логическое ИЛИ
2. Подберите веса персептрона так, чтобы он реализовывал логическое И
3. Докажите, что веса невозможно подобрать так, чтобы он реализовывал исключающее ИЛИ (XOR)
4. Добавьте персептрону вход $x_3 = x_1 \cdot x_2$ так, чтобы он реализовывал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

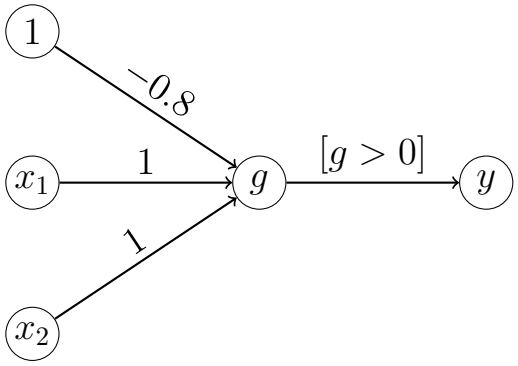
Решение

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 $[x_1 + x_2 - 0.8 > 0]$

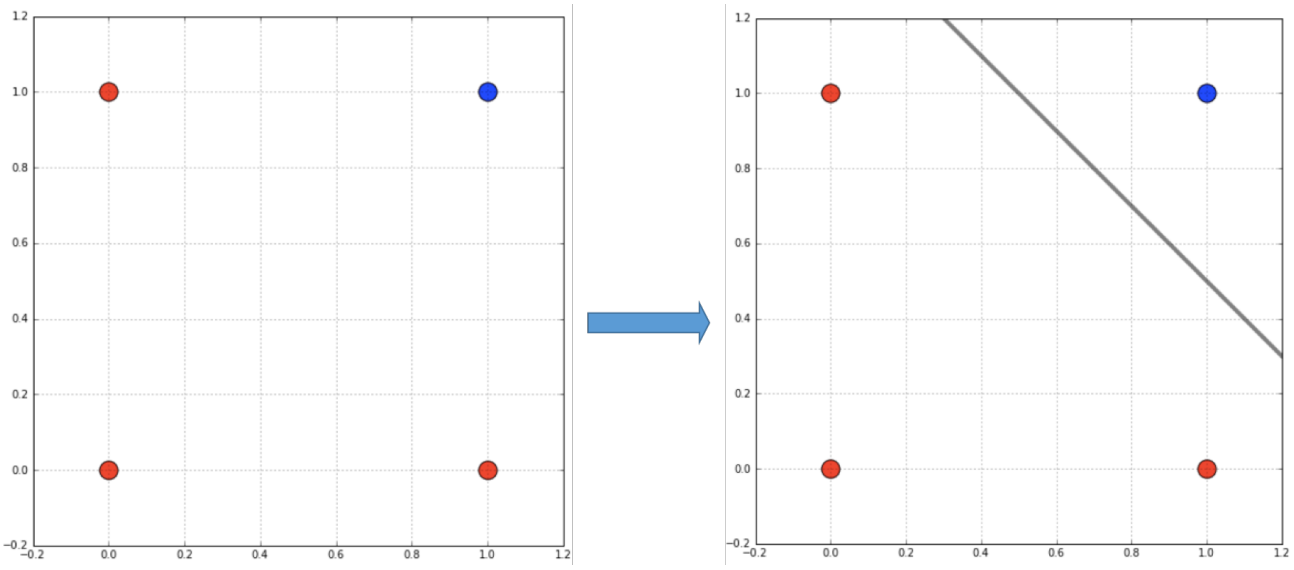


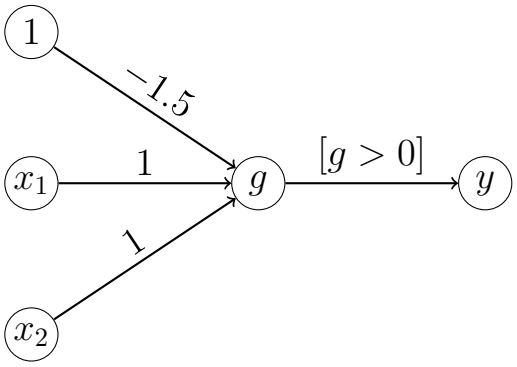


2) Логическое «И»

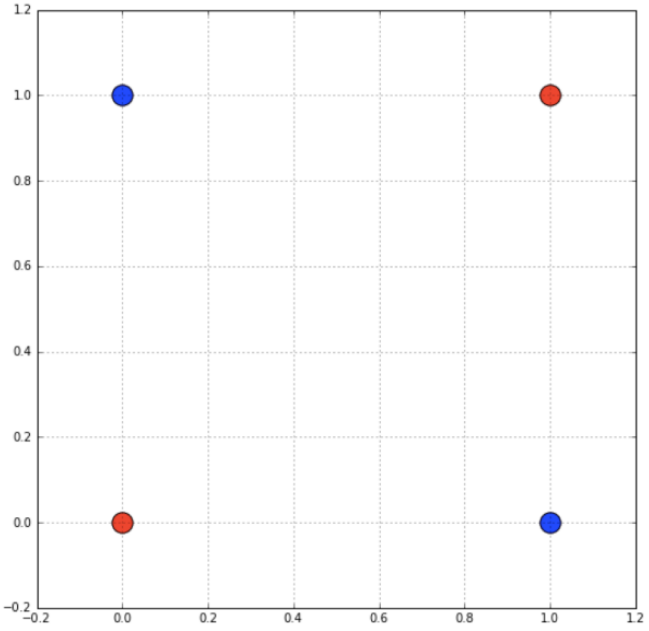
x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

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





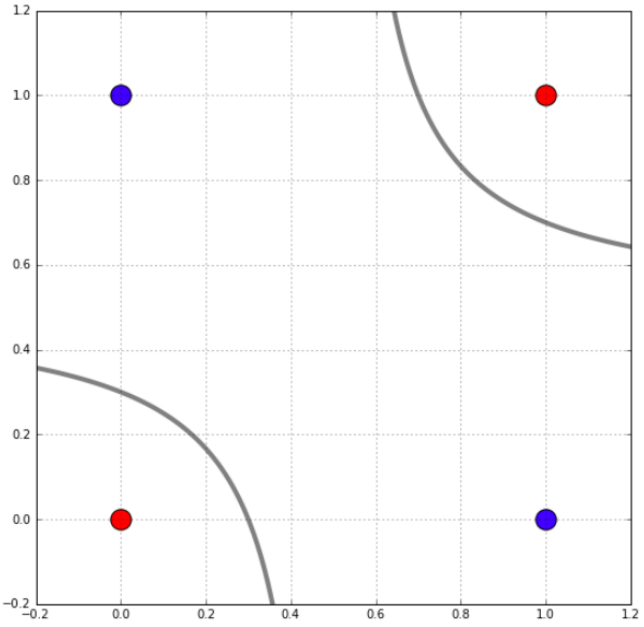
3) Из картинки очевидно, что не существует прямой, верно разделяющей эти точки.

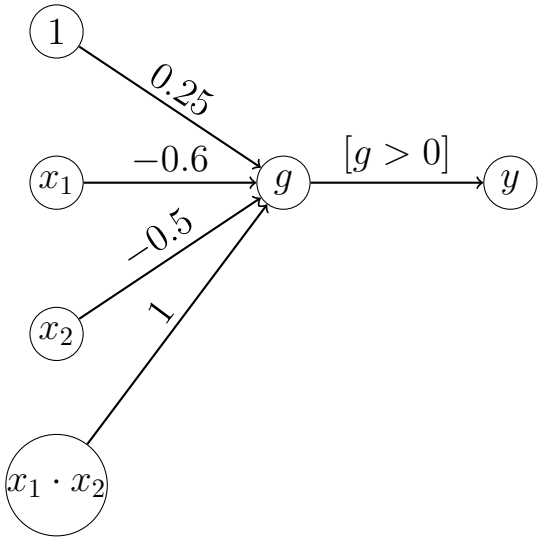


4) Данные точки однако можно разделить гиперболой $x_2 = 0.5 + \frac{0.1}{x_1 - 0.5}$ – остаётся преобразовать её к формату: $a_0 + a_1 \cdot x_1 + a_2 \cdot x_1 \cdot x_2 + a_3 \cdot x_2 = 0$

Итак: уравнение разделяющей поверхности принимает вид:

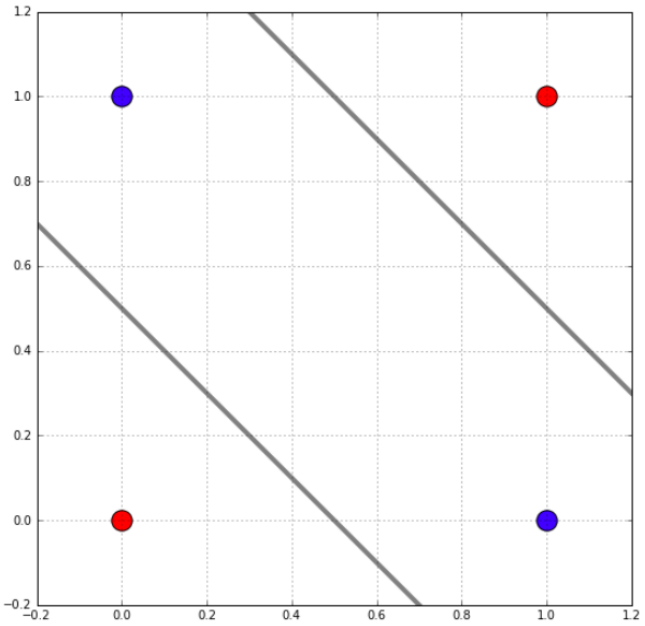
$[0.25 - 0.6x_1 + x_1 \cdot x_2 - 0.5x_2 > 0]$

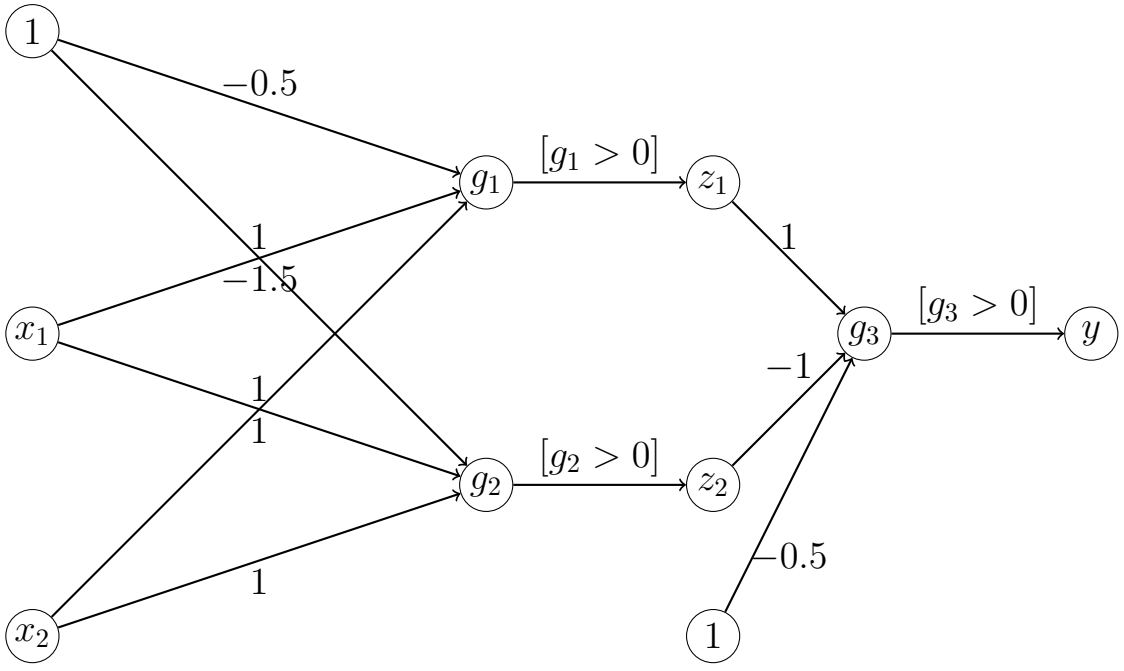




5) Разделение точек с помощью композиции двух персептронов эквивалентно разделению с помощью двух прямых.

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





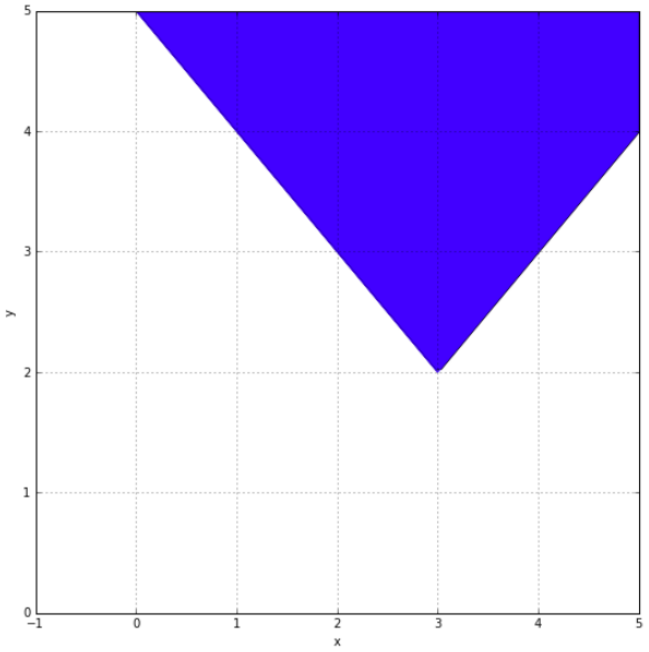
0.3.2 Задача 2

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

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

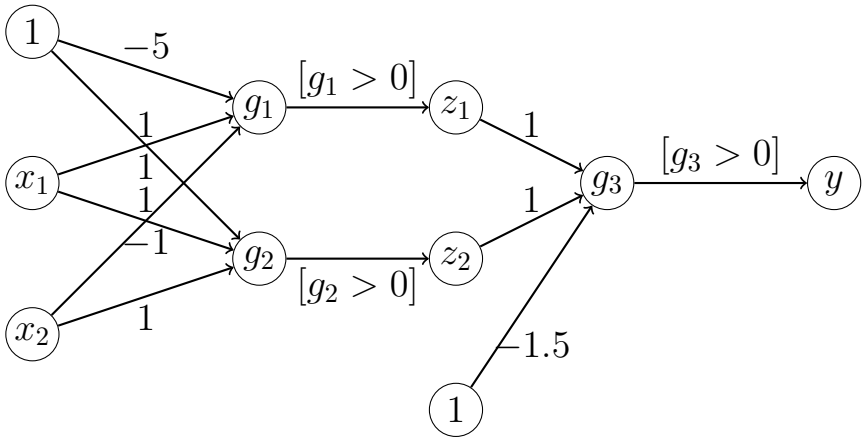
Решение

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:

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



0.3.3 Задача 3

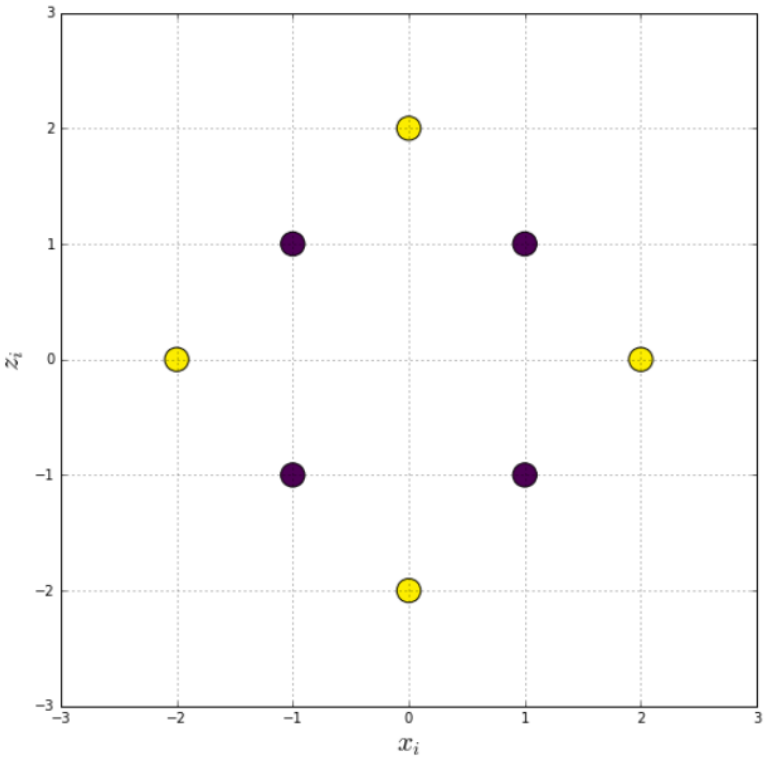
Рассмотрим набор данных:

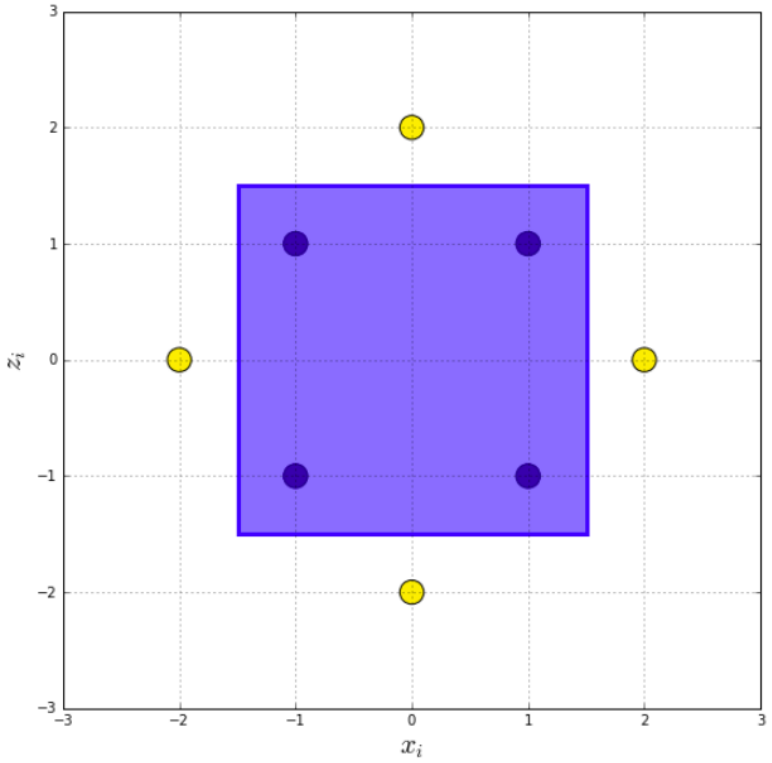
x_i	z_i	y_i
-1	-1	0
1	-1	0
-1	1	0
1	1	0
0	2	1
2	0	1
0	-2	1
-2	0	1

1. Существует ли i персептронов с константой, двумя входами и пороговой фунцкией 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.

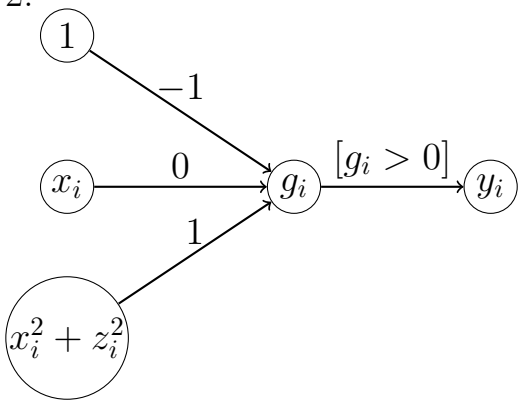
Решение

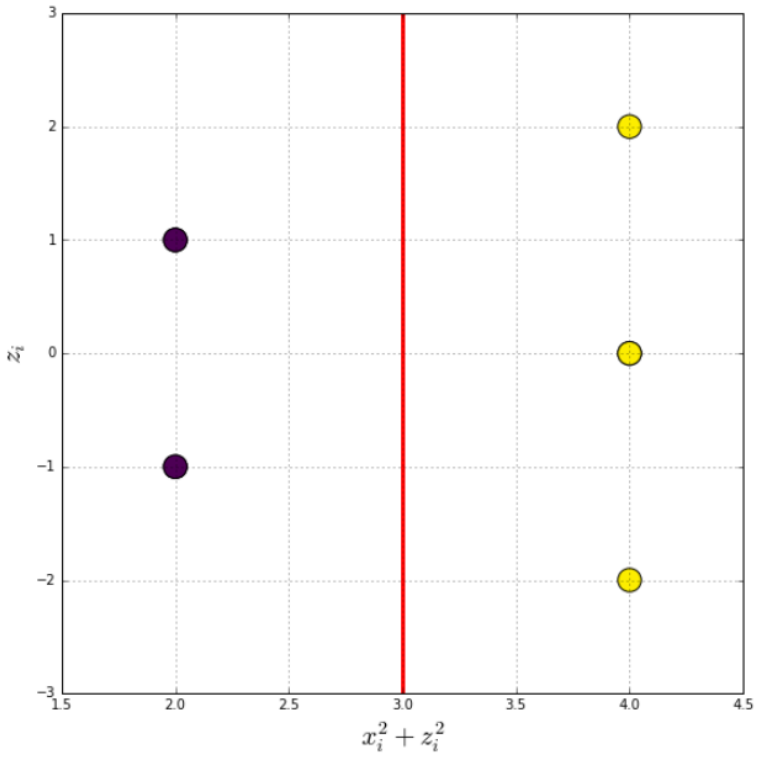




1. Для того, чтобы разделить данную выборку, необходимо собрать персептрон, который будет внутри синего квадрата выдавать 0, а вне его 1. Это возможно сделать минимум двуслойной нейросетью, а следовательно и минимум 3 персептронами:

2.





0.4 SVM

Постановка и решение задачи SVM в общем виде

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

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

Условия Каруша-Куна-Таккера:

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

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:

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

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

Итак, двойственная задача выглядит:

$$\begin{cases} q(\lambda,\mu) = \quad -\frac{1}{2}\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{n}\lambda_i\lambda_j y_i y_j x_i^T x_j - \sum\limits_{i=1}^{n}\lambda_i \to \max\limits_{\lambda} \\ \sum\limits_{i=1}^{n}\lambda_i y_i = 0 \\ \lambda_i \geq 0 \\ \lambda_i + \mu_i = C \end{cases}$$

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

$$\begin{cases} q(\lambda,\mu) = \quad -\frac{1}{2}\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{n}\lambda_i\lambda_j y_i y_j x_i^T x_j - \sum\limits_{i=1}^{n}\lambda_i \to \max\limits_{\lambda} \\ \sum\limits_{i=1}^{n}\lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C \end{cases}$$

Или в матричном виде:

$$(*)\begin{cases} q = -\frac{1}{2}\lambda^T Q\lambda - \vec{1}^T \quad \lambda \to \max\limits_{\lambda} \\ y^T\lambda = 0 \\ 0 \leq \lambda \leq C\cdot\vec{1} \end{cases}$$

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

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

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 $(*)$ перепишется в виде:

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

У этой задачи есть аналитическое решение:

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

Условия Каруша-Куна-Таккера:

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

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

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

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

Теперь можно выписать оценку вектора w:

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

$$\begin{pmatrix} y_0 \\ \vdots \\ y_0 \\ y_+ \\ \vdots \\ y_+ \\ y_- \\ \vdots \\ y_- \end{pmatrix}^T \odot \begin{pmatrix} | & \dots & | & | & | & \dots & | & | & | & \dots | \\ x_0 & \dots & x_0 & | & x_+ & \dots & x_+ & | & x_- & \dots x_- \\ | & \dots & | & | & | & \dots & | & | & | & \dots | \end{pmatrix} \cdot \begin{pmatrix} \lambda_0 \\ \vdots \\ \lambda_0 \\ C \cdot 1 \\ \vdots \\ C \cdot 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Итоговый ответ представим в виде:

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

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

0.4.1 Задача 1

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

Решение

Постановка задачи:

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

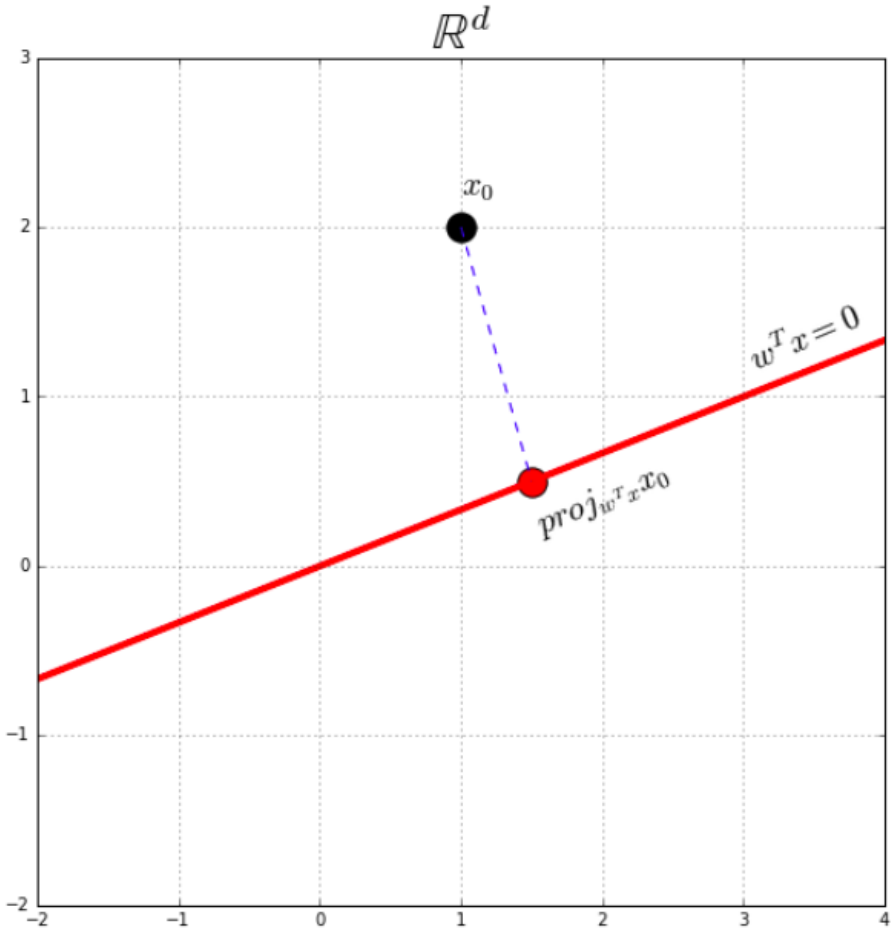
$L = \frac{1}{2}(x - x_0)^T (x - x_0) + \lambda \cdot w^T x$

Условия Каруша-Куна-Таккера:

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

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

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



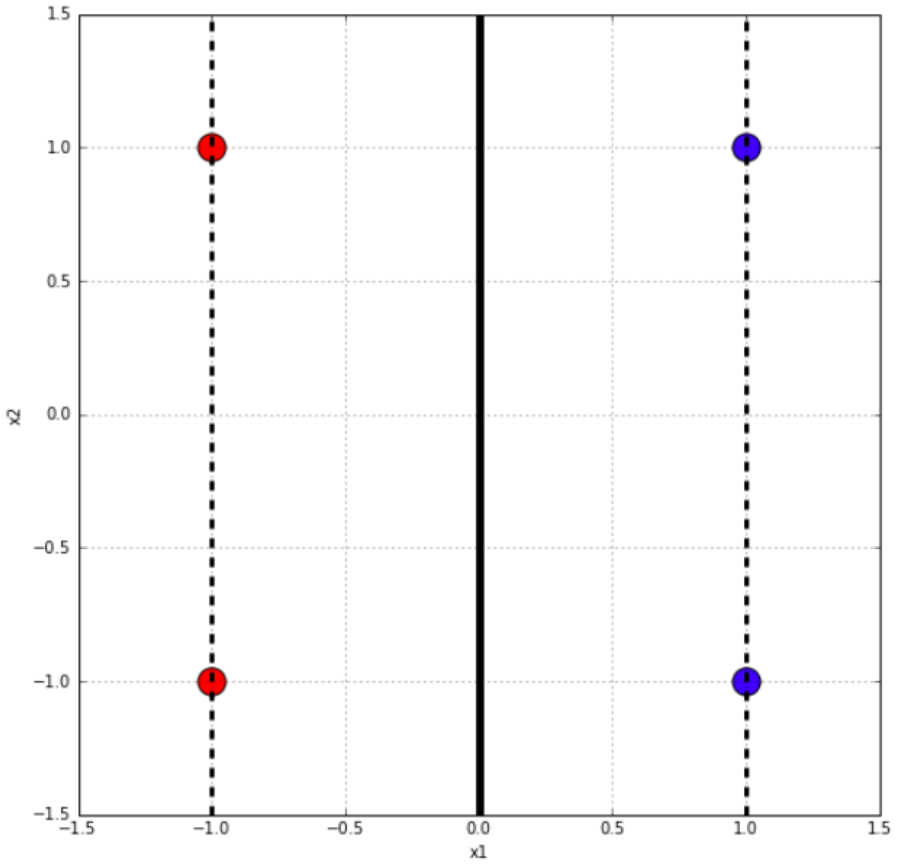
Расстояние от x_0 до $w^T x$ это длина вектора $x_0 - x$

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

0.4.2 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



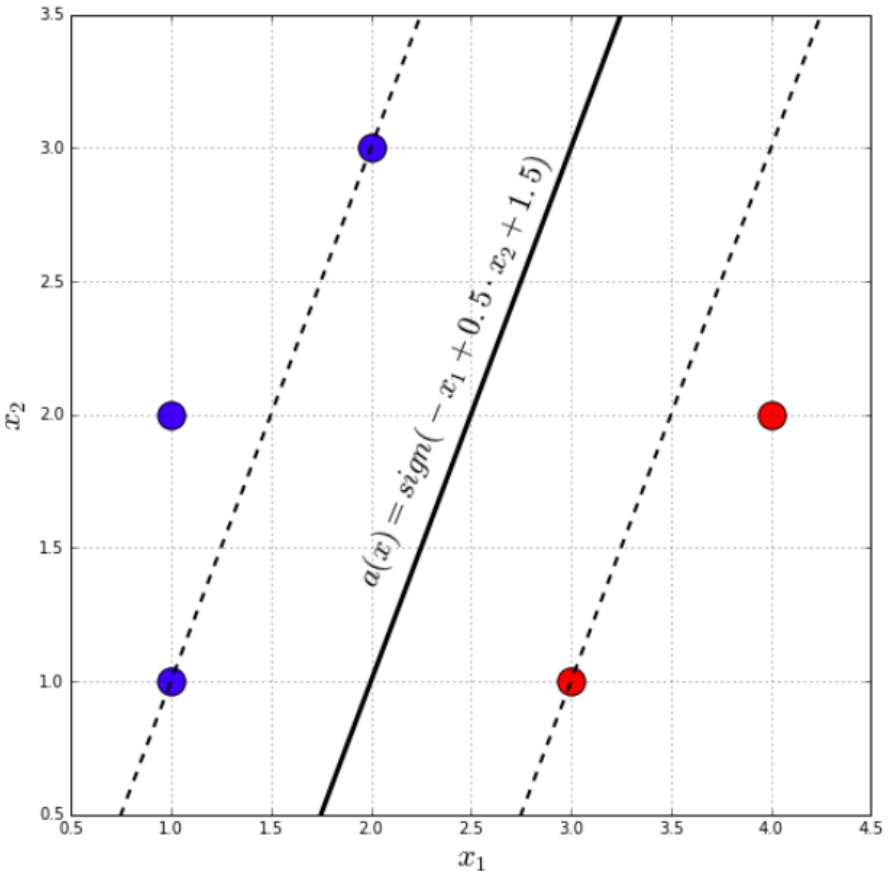
0.4.3 Задача 3

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:



$$\begin{cases} \frac{1}{2}(w_1^2 + w_2^2) & \to \min\limits_{w_1,w_2,b} \\ w_1 + w_2 + b & \geq 1 \\ w_1 + 2w_2 + b & \geq 1 \\ 2w_1 + 3w_2 + b & \geq 1 \\ -3w_1 - 1w_2 - b & \geq 1 \\ -4w_1 - 2w_2 - b & \geq 1 \end{cases}$$

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

Условия Каруша-Куна-Таккера:

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

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:

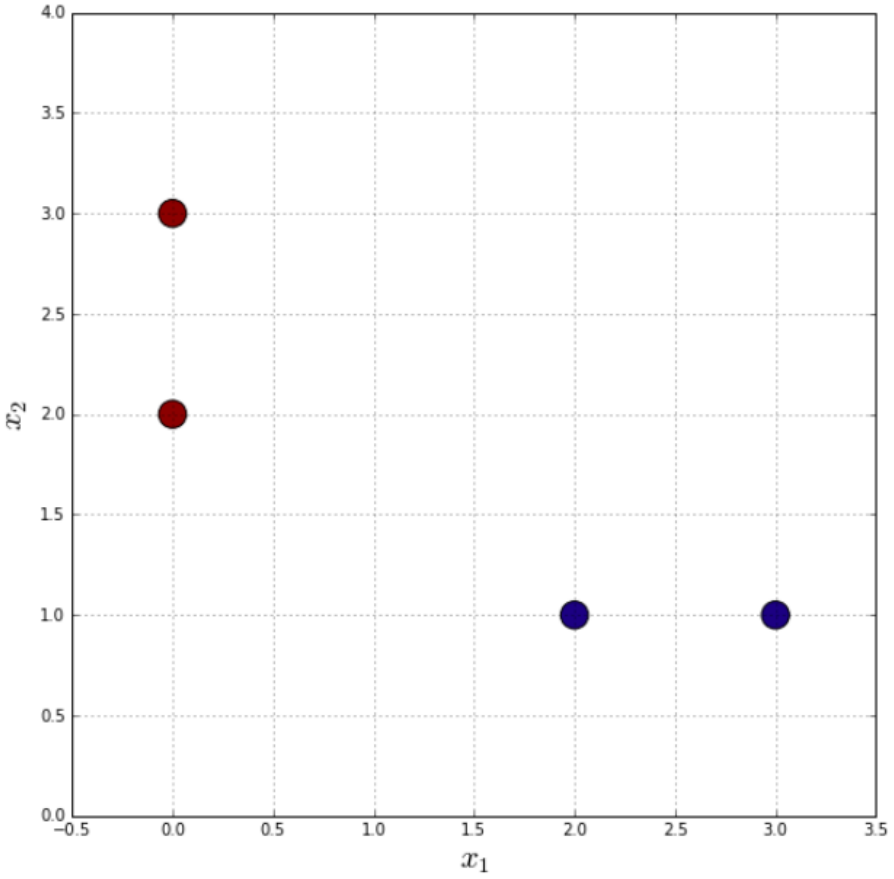
$a(x) = sign(-x_1 + 0.5x_2 + 1.5)$

0.4.4 Задача 4

Найти разделяющую гиперплоскость наибольшей ширины, используя решение двойственной задачи. (В качестве опорных обектов 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: $Q_0^{-1} = \frac{1}{16} \begin{pmatrix} 5 & 2 \\ 2 & 4 \end{pmatrix}$

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

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

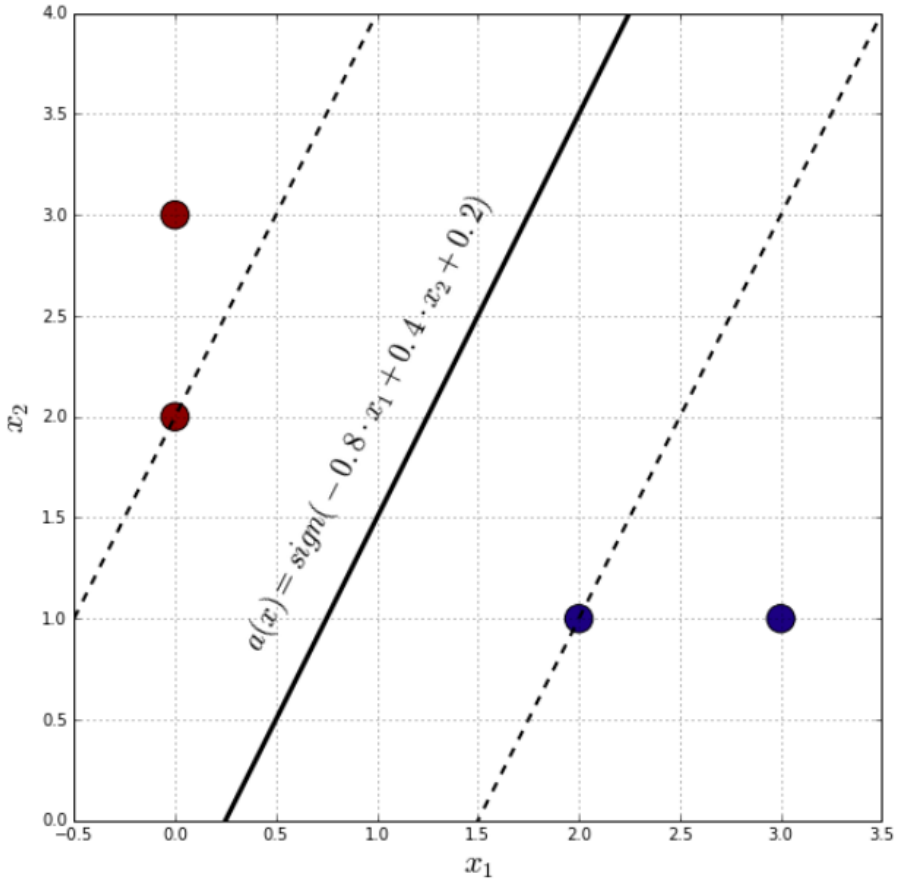
$$\lambda = \frac{1}{16}\begin{pmatrix}5 & 2\\ 2 & 4\end{pmatrix}\cdot\left(\begin{pmatrix}1\\1\end{pmatrix} - \frac{\frac{1}{16}}{\frac{5}{16}}\cdot\begin{pmatrix}1\\-1\end{pmatrix}\right) = \frac{1}{16}\begin{pmatrix}5 & 2\\ 2 & 4\end{pmatrix}\begin{pmatrix}\frac{4}{5}\\ \frac{6}{5}\end{pmatrix} =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

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

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

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: $a(x) = sign(-0.8x_1 + 0.4x_2 + 0.2)$



(решение прямой задачи даёт такую-же разделяющую линию)

0.5 Naive Bayes

0.5.1 Задача 1

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

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

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

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

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

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

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

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).

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

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

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

$$\log L'_{\pi_k} = \sum_{i=1}^{|D|}\frac{[y_i=k]}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = -\frac{1}{\lambda}\sum_{i=1}^{|D|}[y_i = k]$$

$$\sum_{k=1}^{K}\pi_k = 1 \Rightarrow \sum_{k=1}^{K} -\frac{1}{\lambda}\sum_{i=1}^{|D|}[y_i = k] = -\frac{1}{\lambda}\sum_{i=1}^{|D|}\underbrace{\sum_{k=1}^{K}[y_i = k]}_{1} = -|D|\lambda = 1 \Rightarrow \lambda = -\frac{1}{|D|}$$

Итак: $\boxed{\hat{\pi}_k = \dfrac{\sum\limits_{i=1}^{|D|}[y_i = k]}{|D|}}$

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

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

0.5.2 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 $\hat{\theta}_{jk} = \mathbb{E}_{p(\theta_{jk}|D)}\theta_{jk} = \int\limits_0^1 \theta_{jk} \cdot p(\theta_{jk}|D)d\theta_{jk}$

Решение

1) $p(\theta_{jk}|D) = \frac{p(D|\theta_{jk})\cdot p(\theta_{jk})}{\int\limits_0^1 p(D|\theta_{jk})\cdot p(\theta_{jk})d\theta_{jk}}$ – формула Байеса

$$\int\limits_0^1 p(D|\theta_{jk}) \cdot p(\theta_{jk})d\theta_{jk} = \int\limits_0^1 \frac{1}{Beta(\beta_1,\beta_2)} \cdot \theta_{jk}^{\beta_1-1} \cdot (1-\theta_{jk})^{\beta_2-1} \cdot \prod_{i=1}^{|D|}\prod_{k=1}^{K}\left(\pi_k^{[y_i=k]} \cdot \prod_{j=1}^{|W|}\theta_{jk}^{x_{ij}\cdot[y_i=k]}(1-\theta_{jk})^{(1-x_{ij})\cdot[y_i=k]}\right)d\theta_{jk} =$$

$$\frac{1}{Beta(\beta_1,\beta_2)}\prod_{i=1}^{|D|}\prod_{k=1}^{K}\left(\pi_k^{[y_i=k]} \cdot \prod_{j=1}^{|W|}\underbrace{\int\limits_0^1 \theta_{jk}^{x_{ij}\cdot[y_i=k]+\beta_1-1}\cdot(1-\theta_2)^{(1-x_{ij})\cdot[y_i=k]+\beta_2-1}d\theta_{jk}}_{Beta(x_{ij}\cdot[y_i=k]+\beta_1,(1-x_{ij})\cdot[y_i=k]+\beta_2)}\right)$$

Заметим, что $\int\limits_0^1 p(D|\theta_{jk}) \cdot p(\theta_{jk})d\theta_{jk}$ – 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 $\mathbb{E}x = \frac{a}{a+b}$

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

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

0.6 Логистическая регрессия

0.6.1 Задача 1

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

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

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

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

1) По формуле Байеса: $p(y=1|x) = \frac{p(x|y=1)\cdot p(y=1)}{p(x)} = \frac{p(x|y=1)\cdot p(y=1)}{\sum\limits_y p(x,y)} = \frac{p(x|y=1)\cdot p(y=1)}{p(x|y=1)\cdot p(y=1)+p(x|y=0)\cdot p(y=0)} =$

$\frac{\exp\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\}\cdot\phi}{\exp\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\}\cdot\phi+\exp\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\}\cdot(1-\phi)} =$

$\frac{1}{1+\underbrace{\frac{1-\phi}{\phi}}_{\exp\{\log\frac{1-\phi}{\phi}\}}\cdot\exp\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1^T\Sigma^{-1}(x-\mu_1))\}} =$

$\frac{1}{1+\exp\{\log\frac{1-\phi}{\phi}-\frac{1}{2}[2(\mu_1-\mu_0)^T\Sigma^{-1}x+\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1]\}}$

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) $L = \prod\limits_{i=1}^n p(x_i)\cdot p(x_i|y_i) \to \max\limits_{\Sigma,\mu_k,\phi}$

$logL = \sum\limits_{i=1}^n \log p(y_i) + \sum\limits_{i=1}^n \log p(x_i|y_i) = \sum\limits_{i=1}^n y_i\cdot\log\phi + (1-y_i)\cdot\log(1-\phi) + \sum\limits_{i:y_i=1} -\frac{m}{2}\log 2\pi - \frac{1}{2}\log Det(\Sigma) - \frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) + \sum\limits_{i:y_i=0} -\frac{m}{2}\log 2\pi - \frac{1}{2}\log Det(\Sigma) - \frac{1}{2}(x_i-\mu_0)^T\Sigma(x_i-\mu_0) \to \max\limits_{\Sigma,\mu_k,\phi}$

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

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

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

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

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

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

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

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

0.6.2 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 выбоке:

$p(honey_i = 1|bee_i) = \sigma(w_0 + w_1 \cdot bee_i)$

$L = \prod_{i=1}^{100} p(honey_i|bee_i) \to \max\limits_{w_0,w_1}$

$L = \prod_{i=1}^{100} p(honey_i = 1|bee_i = 1) \cdot p(honey_i = 1|bee_i = 0) \cdot p(honey_i = 0|bee_i = 1) \cdot p(honey_i = 0|bee_i = 0) = \prod_{i=1}^{100} \sigma(w_0+w_1) \cdot \sigma(w_0) \cdot (1-\sigma(w_0+w_1)) \cdot (1-\sigma(w_0)) = \sigma(w_0 + w_1)^{12} \cdot \sigma(w_0)^{32} \cdot (1 - \sigma(w_0 + w_1))^{36} \cdot (1 - \sigma(w_0))^{20} \to \max\limits_{w_0,w_1}$

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

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

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

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

Тогда предыдущая система перепишется, как:

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

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

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

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

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

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

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

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

2) $\hat{p}(honey_i = 1|bee_i = 0) = \frac{1}{1+\frac{20}{32}} = \frac{32}{52}$

0.6.3 Задача 3

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

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

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

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

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

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

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

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

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

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

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) $Q(w) = Q(0) + \nabla Q(0)^T w + \frac{1}{2} w^T \nabla^2 Q(0) w \to \min\limits_{w}$

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

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

0.7 Метрики качества

0.7.1 Задача 1

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

$\mathbf{y_i}$	$\mathbf{b(x_i)}$
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
5. Как связан AUC-ROC с долей неправильно классифицированных пар
6. Посчитать AUC-ROC по формуле из пункта 5

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 $TPR(FRP)$

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

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

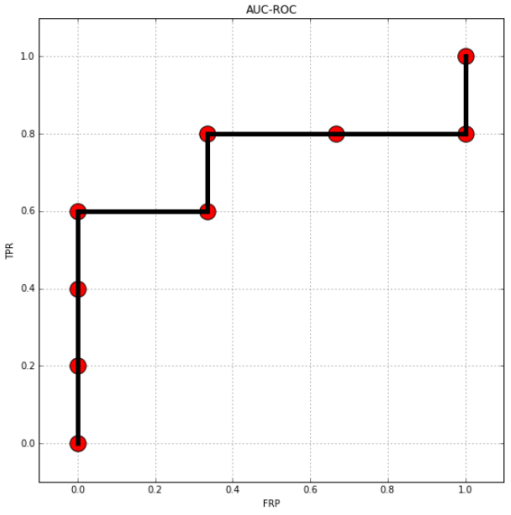
Отсортируем все значения по неубыванию функционала $b_i(x)$:

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

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

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



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

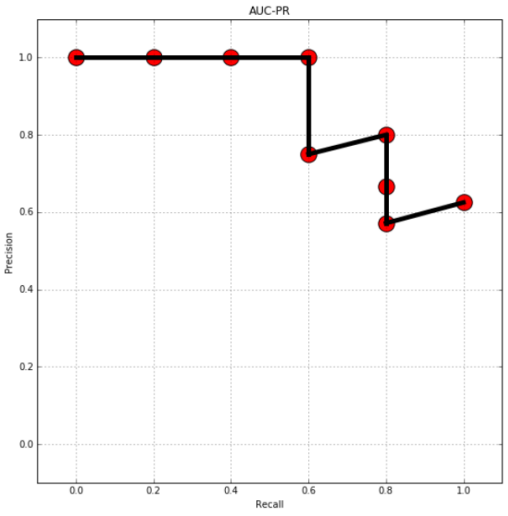
3) $Precision = \frac{TP}{|\hat{y}_+|} = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{|y_+|} = \frac{TP}{TP+FN}$

4) AUC-PR ≈ 0.87

5) Введём функционал $DP = \frac{1}{C_l^2} \sum\limits_{i<j}^{l} [y_i > y_j]$ – доля неправильно классифицированных пар

Докажем, что

$$\boxed{AUC - ROC = 1 - \frac{C_l^2}{l_+ \cdot l_-} \cdot DP = 1 - \frac{1}{l_+ \cdot l_-} \sum_{i<j} [y_i > y_j]}$$, где l_+, l_- – количество объектов положительного и отрицательного классов соответственно



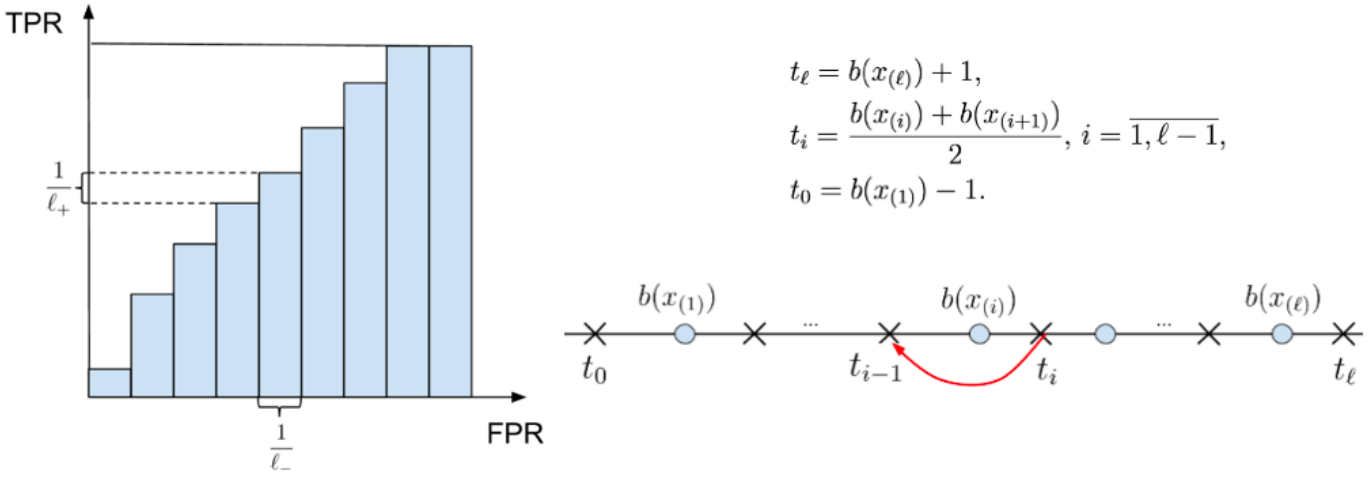


Рис. 1: AUC-ROC

Рассмотрим (рисунок 1). Заметим, что чтобы нарисовать ROC-крисую – достатоно перебрать $l + 1$ значение порога $t_i, \quad i = 1, 2, \ldots, l + 1$

На текущем объекте x_i возможны 2 варианта развития событий:

1) $y_i = +1$ – следовательно, при смене порога с t_i на t_{i+1} TPR возрастает на $\frac{1}{l_+}$, а FPR не меняется.

2) $y_i = -1$ – следовательно, при смене порога с t_i на t_{i+1} TPR не меняется, а FPR возрастает на $\frac{1}{l_-}$.

На текущем объекте x_i AUC-ROC возрастает, если у него изменился FPR. Площадь возрастает на площадь прямоугольника, стороны которого равны соответственно: $\frac{1}{l_-}$ – длина, $\underbrace{\sum_{j=i+1}^{l}[y_j=+1]}_{\texttt{кол-во подъёмов}} \cdot \underbrace{\frac{1}{l_+}}_{\texttt{высота 1 подъёма}}$ – ширина (суммирование ведётся от $i+1$ до l, так как мы перебираем объекты от x_l до x_0)

Тогда AUC-ROC – сумма всех таких прямоугольников:

$$AUC-ROC = \frac{1}{l_-l_+}\sum_{i=1}^{l}[y_i=-1]\left(\sum_{j=i+1}^{l}[y_j=+1]\right) = \frac{1}{l_-l_+}\sum_{i=1}^{l}\sum_{j=i+1}^{l}[y_i=-1]\cdot[y_j=+1] = \frac{1}{l_-l_+}\underbrace{\sum_{i=1}^{l}\sum_{j=i+1}^{l}}_{\sum\limits_{i<j}}[y_i<y_j] = \frac{1}{l_-l_+}\sum_{i<j}(1-[y_i=y_j]-[y_i>y_j]) =$$

$$\frac{1}{l_-l_+}\underbrace{\sum_{i<j}1}_{\frac{l(l-1)}{2}} - \frac{1}{l_-l_+}\sum_{i<j}[y_i=y_j] - \frac{1}{l_-l_+}\sum_{i<j}[y_i>y_j] = \{\sum_{i<j}[y_i=y_j] \text{ – кол-во способов выбрать 2 объекта из } l_+ \text{ объектов или выбрать 2 объекта из } l_- \text{ объектов } = \frac{l_+(l_+-1)}{2}+\frac{l_-(l_--1)}{2}\} =$$

$$= \frac{l^2-l}{2l_+l_-} - \frac{l_+^2-l_++l_-^2-l_-}{2l_+l_-} - \frac{1}{l_-l_+}\sum_{i<j}[y_i>y_j] = \frac{l_+^2+2l_+l_-+l_-^2-l_+^2-l_-^2}{2l_+l_-} - \frac{1}{l_-l_+}\sum_{i<j}[y_i>y_j] = 1-\frac{1}{l_-l_+}\sum_{i<j}[y_i>y_j]$$

6) AUC-ROC $= 1-\frac{1}{3\cdot5}(3+1) = \frac{11}{15}$

0.7.2 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 прогно-

зы

Решение

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

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

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

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

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

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

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

0.7.3 Задача 3

Дата саентист Саша настроил бинарный классификатор $a(x_i)$ в задаче с двумя классами. Ночью во сне ему открылась истина по поводу распределения его данных: данные первого класса распределены с плотностью $p_1(a) = C_1 \cdot (-10a^2 + 20a - 7.5) \cdot [0.5 \leq a \leq 1]$, а данные нулевого класса с плотностью $p_0(a) = C_2 \cdot e^{-2a}$. Помогите Саше вычислить AUC-ROC.

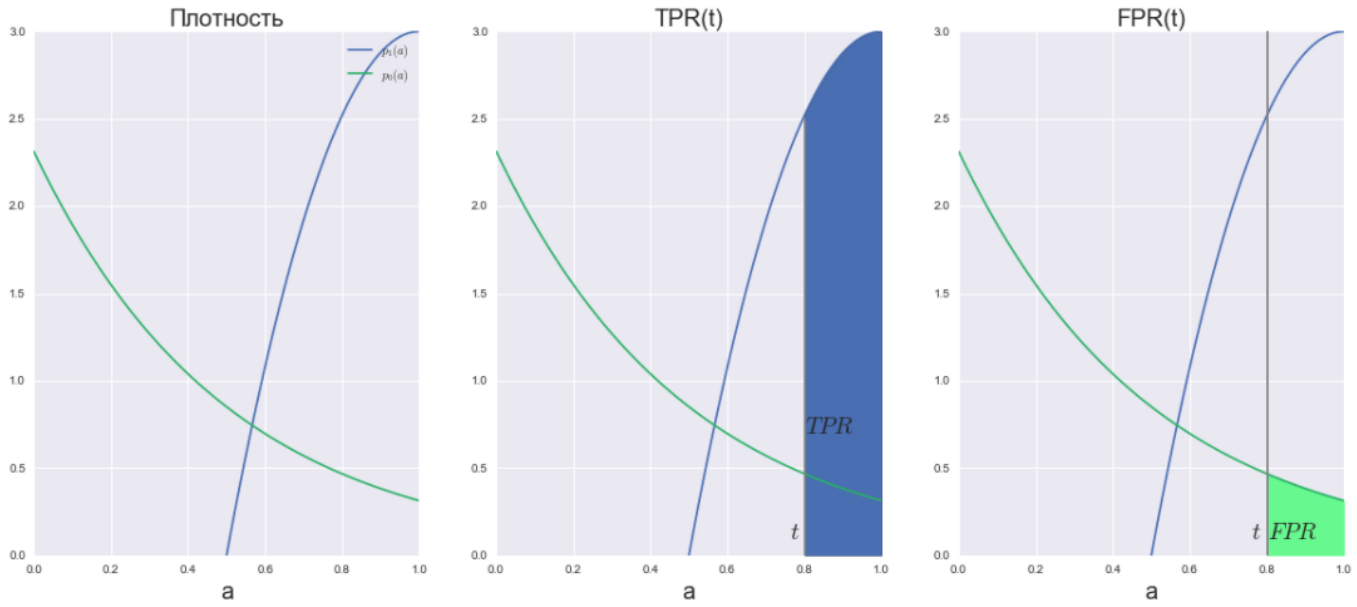
Решение

Для начала найдем плотности, с которыми распределены данные обоих классов:

$\int_{0.5}^{1} C_1 \cdot (-10a^2 + 20a - 7.5)da = C_1 \cdot \left(\frac{5}{6}\right) = 1 \Rightarrow p_1(a) = \frac{6}{5}(-10a^2 + 20a - \frac{15}{2})$

$\int_{0}^{1} C_0 \cdot e^{-2a}da = C_0 \cdot \left(-\frac{1}{2}e^{-2a}|_0^1\right) = -\frac{1}{2}C_0(e^{-2} - 1) = C_0 \cdot (\frac{1}{2} - \frac{e^{-2}}{2}) \Rightarrow p_0(a) = \frac{2}{1-e^{-2}}e^{-2a}$

$\text{TPR(t)} = \int_{t}^{1} \frac{6}{5}(-10a^2 + 20a - 7.5)da = 4t^3 - 12t^2 + 9t - 1, \quad \texttt{при } t \in [0.5, 1]$



$$\text{FPR(t)} = \int_t^1 \frac{2}{1-e^{-2}} \cdot e^{-2a} da = \frac{1}{e^2-1} \cdot (e^{2-2t} - 1)$$

$$\text{AUC-ROC} = \int TPR(t) d(FPR(t)) = \int_1^{0.5} TPR(t) \cdot FPR'(t) dt = \frac{e^{2 \cdot \frac{-12}{5}}}{e^2-1} \int_1^{0.5} (-10a^2 + 20a - \frac{15}{2}) e^{-2t} dt \approx -2.78 \cdot (-0.17) = 0.47$$

0.7.4 Задача 4

Пусть бинарный классификатор $b(x_i)$ выдает в качестве прогноза случайное число между 0 и 1, то есть $b(x_i) \sim Uniform[0, 1]$. Чему равно матожидание значения AUC-ROC для такого классификатора.

Решение

$$AUC - ROC = \frac{1}{l_- l_+} \sum_{i<j} [y_i = -1] \cdot [y_j = +1] \Rightarrow \mathbb{E}(AUC - ROC) = \frac{1}{l_+ l_-} \sum_{i<j} \mathbb{E} \left([y_i = -1] \cdot [y_j = +1] \right)$$

Заметим, что случайная величина $[y_i = -1] \cdot [y_j = +1]$ принимает всего 2 значения: 0 и 1. Рассмотрим матожидание для произвольной бинарной случайной величины: $g \in \{0, 1\}$

$$\mathbb{E}g = p(g = 1) \cdot 1 + p(g = 0) \cdot 0 = p(g = 1)$$

Тогда $\mathbb{E}\left([y_i = -1] \cdot [y_j = +1]\right) = p\left([y_i = -1 \cdot [y_j = +1] = 1\right) = p(y_i = -1, y_j = +1) = \frac{\overbrace{l_+}^{\text{на 1 месте любой из } l_+} \quad \overbrace{l_-}^{\text{на 2ом месте любой из } l_-} \quad \overbrace{(l-2)!}^{\text{остальные } l-2 \text{ элементов}}}{\underbrace{l!}_{\text{все перестановки}}}$

$\mathbb{E}(AUC - ROC) = \frac{1}{l_+ l_-} \sum\limits_{i<j} \frac{l_+ l_- (l-2)!}{l!} = \frac{1}{l_+ l_-} \cdot \frac{l(l-1)}{2} \cdot \frac{l_+ l_-}{l(l-1)} = \frac{1}{2}$

0.8 Деревья

0.8.1 Задача 1

Постройте регрессионное дерево для прогнозирования y по x на обучающей выборке. Критерий деления узла на два – минимизация квадрата отклонения. Узел делится, до тех пор пока в нём больше 2 наблюдений.

$\mathbf{y_i}$	$\mathbf{x_i}$
100	1
102	2
103	3
50	4
55	5
61	6
70	7

Решение

$H(R) = \frac{1}{|R|} \sum\limits_{(x_i,y_i)\in R} \left(y_i - \frac{1}{|R|} \sum\limits_{(x_j,y_j)\in R} y_j \right)^2$, где R – множество объектов, попавших в вершину R

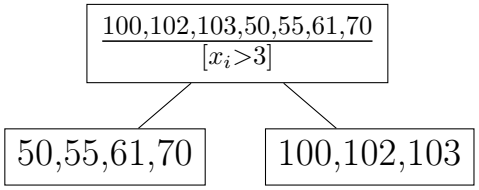
Общий критерий: $Q(R_{now}) = H(R_{now}) - \frac{|R_l|}{|R_{now}|} \cdot H(R_l) - \frac{|R_r|}{|R_{now}|} \cdot H(R_r) \to \max\limits_t$

R_{now} – текущий лист, R_r – множество объектов в правом подлисте, R_l – множество объектов в левом подлисте.

Критерий будет максимизироваться, если мы каждый раз будем разбирать по порогу, который близок к среднему значению y в текущем листе.

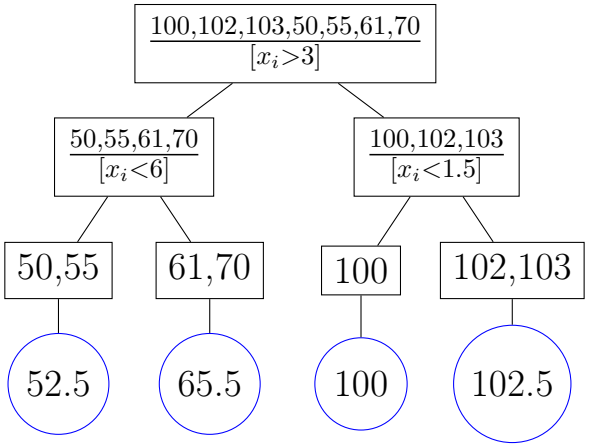
1) $\bar{y} = \frac{541}{7} = 77.29 \Rightarrow$ первым критерием разбиения может стать например $[y_i < 78] \Rightarrow [x_i > 3]$

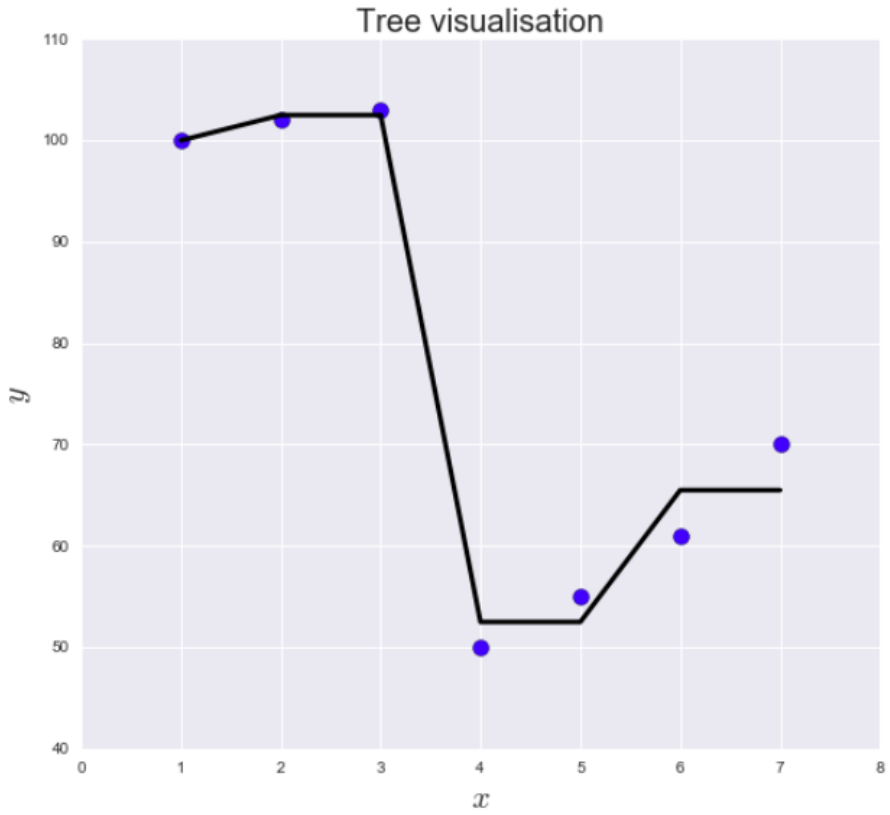
Договоримся, что в левый лист будут просеиваться объекты для которых выполнен придикат, а в правый для которых не выполнен.



2) В левом листе: $\bar{y}_l = \frac{236}{4} = 59, \quad \bar{y}_r = \frac{305}{3} = 101.67$

Тогда критерием разделения для левого листа может стать например $[y_i < 59] \Rightarrow [x_i < 6]$, а для правого $[y_i < 101] \Rightarrow [x_i < 1.5]$

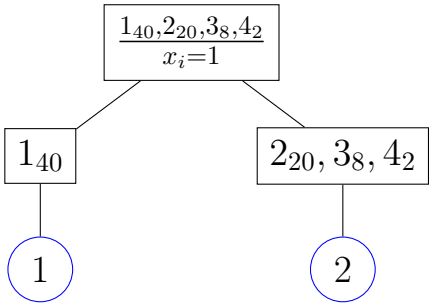




0.8.2 Задача 2

Дон Жуан предпочитает брюнеток. Он посчитал, что в его записной книжке 20 блондинок, 40 брюнеток, 2 рыжих и 8 шатенок. С нового года он решил записать всех на 2 записные книжки – в одну брюнеток, во вторую всех остальных. Как в результате этого события изменится индекс Джинни и энтропия.

Решение



В случае задачи классификации в листе выдается в качестве прогноза самый часто встречающийся класс.

Посчитаем энтропию и индекс Джинни в корне дерева:

Обозначения: $p_{mk} = \frac{1}{|R_m|} \sum\limits_{x_i \in R_m} [y_i = k]$, p_{mk} – доля объектов класса k, попавших в вершину m

Тогда индекс Джини: $F_G(R_m) = 1 - \sum\limits_{k=1}^{K} p_{mk}^2$

Энтропия: $F_H(R_m) = -\sum\limits_{k=1}^{K} p_{mk} \cdot \log p_{mk}$

$F_H(R_{root}) = -\left(\frac{40}{70}\log\frac{40}{70} + \frac{20}{70}\log\frac{20}{70} + \frac{8}{70}\log\frac{8}{70} + \frac{2}{70}\log\frac{2}{70}\right) = 1.03$

$F_H(R_{left}) = 0, \quad F_H(R_{right}) = -\left(\frac{20}{30}\log\frac{20}{30} + \frac{8}{30}\log\frac{8}{30} + \frac{2}{30}\log\frac{2}{30}\right) = 0.8$

Среднее значение энтропии: $\bar{H} = \frac{40}{70} \cdot 0 + \frac{30}{70} \cdot 0.8 = 0.6$

$F_G(R_{root}) = 1 - \left(\left(\frac{40}{70}\right)^2 + \left(\frac{20}{70}\right)^2 + \left(\frac{8}{70}\right)^2 + \left(\frac{2}{70}\right)^2\right) = 0.578$

$F_G(R_{left}) = 0, \quad F_G(R_{right}) = 1 - \left(\left(\frac{20}{30}\right)^2 + \left(\frac{8}{30}\right)^2 + \left(\frac{2}{30}\right)^2\right)^2 = 0.39$

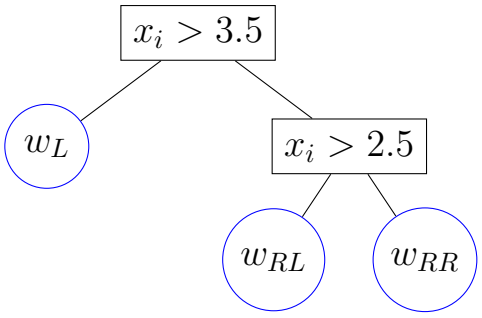
Среднее значение Джини: $\bar{G} = \frac{40}{70} \cdot 0 + \frac{30}{70} \cdot 0.39 = 0.165$

0.8.3 Задача 3

Пятачок собрал данные о визитах Винни-Пуха к Кролику. x_i – количество съеденного Пухом мёда, y_i – факт застревания Винни-Пуха при выходе. Пятачок получил следующие данные:

$\mathbf{y_i}$	$\mathbf{x_i}$
0	1
1	4
1	2
0	3
1	3
0	1

Пятачок использует модель:



Где в качестве криетрия качества используется квадратичная аппроксимация логистической функции потерь:

$$Q(w) = \sum_{i=1}^{n} \left(loss(y_i, 0) + loss(y_i, 0)' \cdot w_i + \frac{1}{2} \cdot loss(y_i, 0)'' \cdot w_i^2 \right) + \frac{\lambda}{2} \cdot \|w\|_2^2$$

$$loss(y_i, w_i) = \log(1 + \exp(-y_i \cdot x_i \cdot w_i))$$

Помогите Пятачку настроить веса

Решение

$$Q(w) = \underbrace{\sum_{i:x_i>3.5} Q(w_i)}_{Q_1} + \underbrace{\sum_{i:x_i>2.5 \bigcap dom(\bar{Q}_1)} Q(w_i)}_{Q_2} + \underbrace{\sum_{i:x_i \leq 2.5} Q(w_i)}_{Q_3} \to \min_{w_i}$$

Задача распадается на 3 отдельные подзадачи:

$$Q_j \to \min_{w_i}, \quad j = 1, 2, 3$$

$$loss(y_i, 0)' = \frac{y_i \cdot x_i}{1+\exp(y_i \cdot x_i \cdot 0)} = \frac{1}{2} y_i x_i$$

$$loss(y_i, 0)'' = -(y_i x_i)^2 \cdot (1 + \exp(y_i x_i \cdot 0))^{-2} \cdot \exp(y_i x_i \cdot 0) = -\frac{1}{2} y_i x_i^2$$

$$Q'_{1,w_{LR}} = \sum_i loss(y_i, 0)' + w_{LR} \cdot \sum_i loss(y_i, 0)'' + \lambda \cdot w_{LR} = 0$$

$$w_L^* = -\frac{\sum\limits_{i:x_i>3.5} loss(y_i,0)'}{\left(\sum\limits_{i:x_i>3.5} loss(y_i,0)''+\lambda\right)} = -\frac{\sum\limits_{i:x_i>3.5} \frac{1}{2} y_i x_i}{\sum\limits_{i:x_i>3.5} -\frac{1}{2} y_i x_i^2 + 1} = -\frac{2}{-8+1} = \frac{2}{7}$$

$$w_{RL}^* = -\frac{\sum\limits_{i:x_i>2.5} loss(y_i,0)'}{\left(\sum\limits_{i:x_i>2.5} loss(y_i,0)''+\lambda\right)} = -\frac{\sum\limits_{i:x_i>2.5} \frac{1}{2} y_i x_i}{\sum\limits_{i:x_i>2.5} -\frac{1}{2} y_i x_i^2 + 1} = -\frac{\frac{1}{2} \cdot 3 \cdot 1}{-\frac{1}{2} \cdot 9 \cdot 1 + 1} = \frac{3}{7}$$

$$w_{RR}^* = -\frac{\sum\limits_{i:x_i\le2.5} loss(y_i,0)'}{\left(\sum\limits_{i:x_i\le2.5} loss(y_i,0)''+\lambda\right)} = -\frac{\sum\limits_{i:x_i\le2.5} \frac{1}{2} y_i x_i}{\sum\limits_{i:x_i\le2.5} -\frac{1}{2} y_i x_i^2 + 1} = -\frac{1}{-2+1} = 1$$

0.8.4 Задача 4

Для предыдущей задачи найти веса методом максимального правдоподобия.

Решение

Функция правдоподобия в данном случае выглядит так:

$L(y, a) = \prod_{i=1}^{n} a(x_i)^{y_i} \cdot (1 - a(x_i))^{1-y_i}$ где $a(x_i)$ – решающее дерево.

Алгоритм классификации с помощью решающего дерева можно явно задать функцией: $a(x_i) = w_L^{[x_i>3.5]} \cdot w_{RL}^{[2.5<x_i\le3.5]} \cdot w_{RR}^{[x_i\le2.5]}$

Тогда функция правдоподобия принимает вид:

$$L(y,a) = \prod_{i=1}^{n} \left(w_L^{[x_i>3.5]} \cdot w_{RL}^{[2.5<x_i\le3.5]} \cdot w_{RR}^{[x_i\le2.5]}\right)^{y_i} \cdot \left(1 - w_L^{[x_i>3.5]} \cdot w_{RL}^{[2.5<x_i\le3.5]} \cdot w_{RR}^{[x_i\le2.5]}\right)^{1-y_i} = \prod_{i:x_i>3.5} w_L^{y_i} \cdot (1-w_L)^{1-y_i} \cdot \prod_{i:2.5<x_i\le3.5} w_{RL}^{y_i} \cdot$$

$$(1-w_{RL})^{1-y_i} \cdot \prod_{i:x_i \le 2.5} w_{RR}^{y_i} \cdot (1-w_{RR})^{1-y_i}$$

$$\log L = \sum_{i:x_i>3.5} y_i \cdot \log w_L + (1-y_i) \cdot \log(1-w_L) + \sum_{i:2.5<x_i\le 3.5} y_i \cdot \log w_{RL} + (1-y_i) \cdot \log(1-w_{RL}) + \sum_{i:x_i \le 2.5} y_i \cdot \log w_{RR} + (1-y_i) \cdot \log(1-w_{RR}) \to \max_{w_L, w_{RL}, w_{RR}}$$

$$\log L'_{w_L} = \frac{\sum\limits_{i:x_i>3.5} y_i}{w_L} - \frac{\sum\limits_{i:x_i>3.5} (1-y_i)}{1-w_L} = 0 \Rightarrow \boxed{w_L^* = \frac{\sum\limits_{i=1}^{n} y_i \cdot [x_i > 3.5]}{\sum\limits_{i=1}^{n} [x_i > 3.5]}}$$

Аналогично вычисляются оценки для w_{RL}, w_{RR}, Итак:

$$\boxed{w_{RL}^* = \frac{\sum\limits_{i=1}^{n} y_i \cdot [2.5 < x_i \le 3.5]}{\sum\limits_{i=1}^{n} [2.5 < x_i \le 3.5]}}, \quad \boxed{w_{RR}^* = \frac{\sum\limits_{i=1}^{n} y_i \cdot [x_i \le 2.5]}{\sum\limits_{i=1}^{n} [x_i \le 2.5]}}$$

$w_L^* = \frac{1}{1} = 1, \quad w_{RL}^* = \frac{1+0}{1+1} = \frac{1}{2}, \quad w_{RR}^* = \frac{0+1+0}{1+1+1} = \frac{1}{3}$

0.9 Градиентный бустинг

$a_N(x) = \sum_{j=1}^{N} \gamma_j \cdot b_j(x)$

Обучение происходит путём минимизации функционала:

$\sum_{i=1}^{l} L(y_i, a_{N-1} + \gamma_N \cdot b_N(x_i)) \to \min_{\gamma, b}$

$s_i = -\frac{\partial L}{\partial z}\big|_{z=a_{N-1}(x_i)}$

Новый алгоритм $b_j(x)$ находится путём минимизации:

$b_N = \text{argmin}_b \sum_{i=1}^{l} (b(x_i) - s_i)^2$

$\gamma_N = \text{argmin}_\gamma \sum_{i=1}^{l} L(y_i, a_{N-1} + \gamma \cdot b_N(x_i))$

Бустинг над деревьями:

$\sum_{i=1}^{l} L(y_i, a_{N-1} + \gamma_N \sum_{j=1}^{J_N} b_{Nj}[x_i \in R_j]) = \sum_{i=1}^{l} L(y_i, a_{N-1} + \sum_{x_i \in R_j} \gamma_{Nj}) \to \min_{\gamma_{Nj}}$

$\gamma_{Nj} = \text{argmin}_\gamma \sum_{i=1}^{l} L(y_i, a_{N-1} + \sum_{x_i \in R_j} \gamma)$

0.9.1 Задача 1

Найти сдвиги (s_i) в случае следующих функций потерь:

- $L(y, z) = (y - z)^2$
- $L(y, z) = |y - z|$
- $L(y, z) = \log(1 + \exp(-yz))$

Решение

1) $s_i = -\frac{\partial L(y_i, z)}{\partial z}|_{a_{N-1}(x_i)} = -(-2) \cdot (y_i - z)|_{a_{N-1}(x_i)} = 2 \cdot (y_i - a_{N-1}(x_i))$

2) $s_i = -\frac{\partial L(y_i, z)}{\partial z}|_{a_{N-1}(x_i)} = -(-1) \cdot sign(y_i - a_{N-1}(x_i)) = sign(y_i - a_{N-1}(x_i))$

3) $s_i = -\frac{\partial L(y_i,z)}{\partial z}|_{a_{N-1}(x_i)} = -\frac{\exp(-y_i \cdot a_{N-1}(x_i)) \cdot (-y_i)}{1+\exp(-y_i \cdot a_{N-1}(x_i))} = y_i \cdot \sigma(-y_i \cdot a_{N-1}(x_i))$ где $\sigma(x) = \frac{1}{1+\exp(-x)}$

0.9.2 Задача 2

Найти оптимальное значение весов произвольных алгоритмов (пункт 1) и оптимальное значение алгоритмов в случае бустинга над деревьями (пункты 2 – 4)

1. $L(y,z) = (y-z)^2$
2. $L(y,z) = (y-z)^2$
3. $L(y,z) = |y-z|$
4. $L(y,z) = e^{-yz},\ y_i \in \{-1,1\}$

Решение

1) $\gamma_N = \text{argmin}_\gamma \sum_{i=1}^{l}(y_i - a_{N-1} - \gamma \cdot b_N(x_i))^2$

$\sum_{i=1}^{l} -2 \cdot (y_i - a_{N-1}(x_i) - \gamma \cdot b_N(x_i)) \cdot b_N(x_i) = 0 \to \sum_{i=1}^{l} b_N(x_i) \cdot (y_i - a_{N-1}) = \gamma \cdot \sum_{i=1}^{l} b_N(x_i)^2$

$$\boxed{\hat{\gamma}_N = \frac{\sum_{i=1}^{l} b_N(x_i) \cdot (y_i - a_{N-1}(x_i))}{\sum_{i=1}^{l} b_N(x_i)^2}}$$

2) $\gamma_{Nj} = \text{argmin}_\gamma \sum_{x_i \in R_j}^{l} (y_i - a_{N-1} - \gamma)^2$

$\sum_{x_i \in R_j} (-2) \cdot (y_i - a_{N-1} - \gamma) = 0 \to \boxed{\hat{\gamma}_{Nj} = \frac{1}{|R_j|} \sum_{x_i \in R_j} (y_i - a_{N-1})}$

3) $\gamma_{Nj} = \text{argmin}_\gamma \sum_{x_i \in R_j}^{l} |y_i - a_{N-1} - \gamma| \to \boxed{\hat{\gamma}_{Nj} = median_{y_i \in R_j}\{y_i - a_{N-1}\}}$

4) $\gamma_{Nj} = \text{argmin}_\gamma \sum_{x_i \in R_j} e^{-y_i \cdot (a_{N-1}+\gamma)} = \text{argmin}_\gamma \sum_{y_i=-1} e^{a_{N-1}+\gamma} + \sum_{y_i=1} e^{-a_{N-1}-\gamma}$

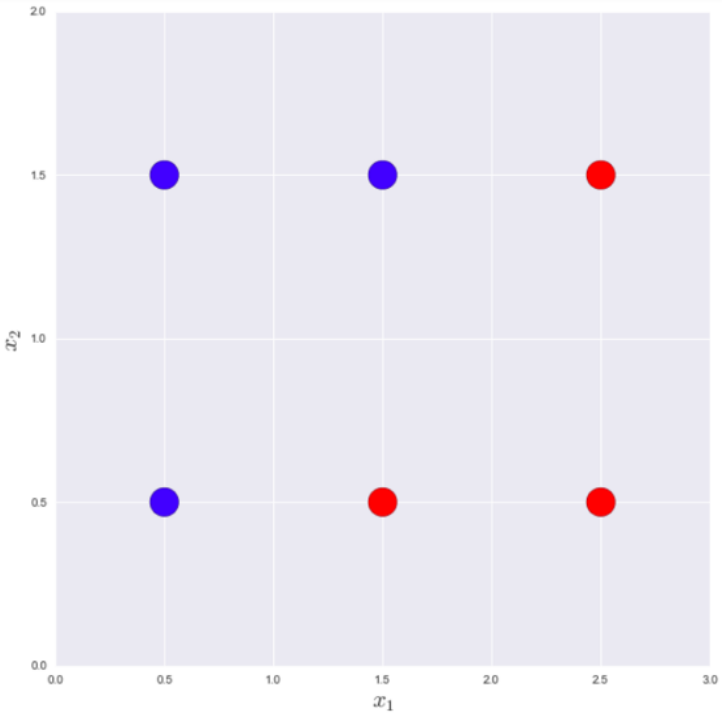
$$\sum_{y_i=-1} e^{a_{N-1}(x_i)} \cdot e^{\gamma} - \sum_{y_i=1} e^{-a_{N-1}(x_i)} \cdot e^{-\gamma} = 0 \Big| \cdot e^{\gamma}$$

$$e^{2\gamma} \cdot \sum_{y_i=-1} e^{a_{N-1}(x_i)} = \sum_{y_i=1} e^{-a_{N-1}(x_i)} \Rightarrow \boxed{\hat{\gamma}_{Nj} = \frac{1}{2} \log \left(\frac{\sum\limits_{y_i=1} e^{-a_{N-1}(x_i)}}{\sum\limits_{y_i=-1} e^{a_{N-1}(x_i)}} \right)}$$

0.9.3 Задача 3

Проделать 5 шагов градиентного бустинга над решающими пнями (деревья глубины 1) для задачи классификации данных, представленных ниже:

$\mathbf{x_i^1}$	$\mathbf{x_i^2}$	$\mathbf{y_i}$
1.5	0.5	0
2.5	0.5	0
2.5	1.5	0
0.5	0.5	1
0.5	1.5	1
1.5	1.5	1



В качестве нулевого алгоритма возьмём $b_0(x) = 0$ и все веса положим равными единице: $\gamma_i = 1$. Для обучения пней используем квадратичную функцию потерь $L(y_i, a_N) = \sum\limits_{i=1}^{6} (y_i - a_N(x_i))^2$ а для выбора порога разбиения изменение дисперсии при расщеплении:

$F(R_m, R_r, R_l) = H(R_m) - \frac{|R_l|}{|R_m|} \cdot H(R_l) - \frac{|R_r|}{|R_m|} \cdot H(R_r)$,

$H(R_k) = \frac{1}{|R_k|} \cdot \sum\limits_{x_i \in R_k} (y_i - \bar{y})^2 \Rightarrow H(R_k) = \bar{y^2} - (\bar{y})^2$

Решение

1 шаг:

Выберем порог для первого решающего пня, для этого надо выбрать такой порог $d : [x_i^k < d]$, при котором максимизируется функционал $F(R_m, R_r, R_l)$. Ясно, что порог, который оставляет всю выборку в одном листе обеспечивает качество функционала равное нулю: $F = 0$. Поэтому достаточно перебрать критерии: $[x_i^1 < 1], [x_i^1 < 2], [x_i^2 < 1], [x_i^2 < 2]$. ($R_l = R_d$ – левый лист или нижний лист.)

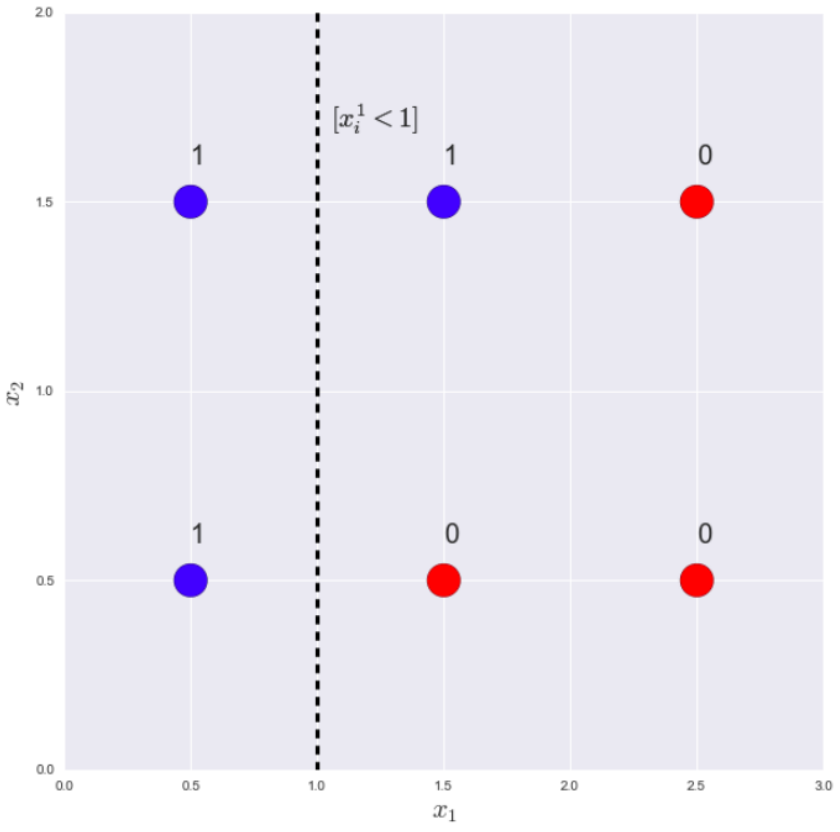
Критерий	$H(R_m)$	$H(R_l)$	$H(R_r)$	$F(R_m, R_r, R_l)$
$[x_i^1 < 1]$	$\frac{1}{2} - \frac{1}{4} = \frac{1}{4}$	0	$\frac{3}{4} - \frac{9}{16} = \frac{3}{16}$	$\frac{1}{4} - \frac{2}{6} \cdot 0 - \frac{4}{6} \cdot \frac{3}{16} = \frac{1}{8}$
$[x_i^1 < 2]$	$\frac{1}{4}$	$\frac{3}{4} - \frac{9}{16} = \frac{3}{16}$	0	$\frac{1}{4} - \frac{2}{6} \cdot 0 - \frac{4}{6} \cdot \frac{3}{16} = \frac{1}{8}$
$[x_i^2 < 1]$	$\frac{1}{4}$	$\frac{2}{3} - \frac{4}{9} = \frac{2}{9}$	$\frac{1}{3} - \frac{1}{9} = \frac{2}{9}$	$\frac{1}{4} - \frac{1}{2} \cdot \frac{2}{9} - \frac{1}{2} \cdot \frac{2}{9} = \frac{1}{36}$

Первый и второй критерии дают одинаковое наибольшее качество, будем считать, что из них был выбран первый ($[x_i^1 < 1]$). Итак:

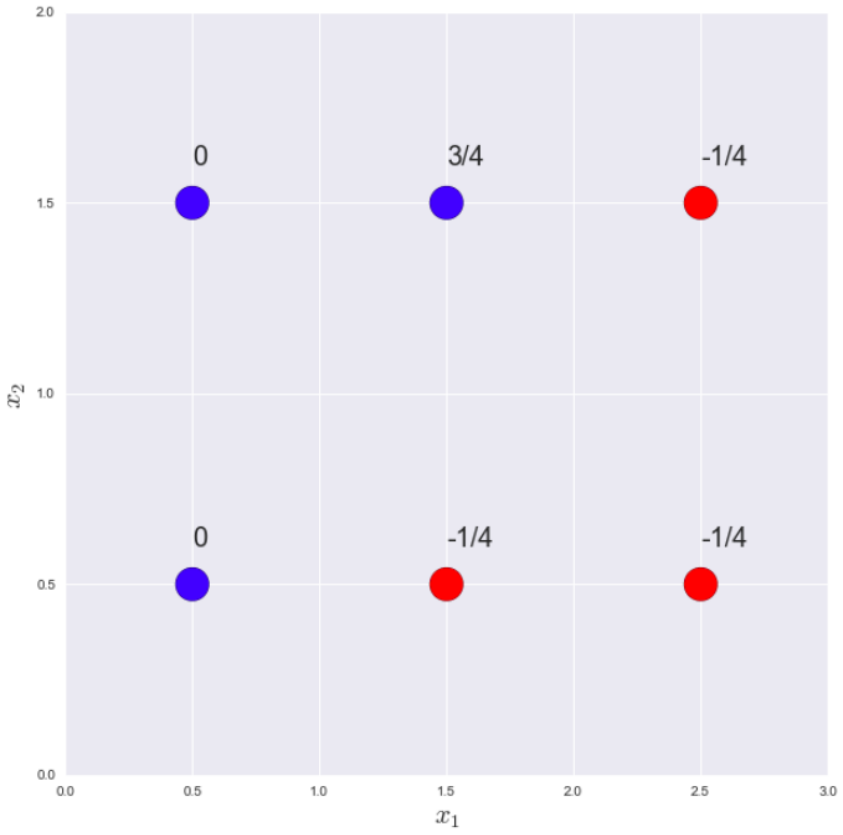
$$a_N = a_1 = b_0 + b_1 = [x_i^1 < 1].$$

В случае квадратичной функции потерь оптимальный прогноз – выдавать среднее значение в листе. Итак:

$$a_1(x) = 1 \cdot [x_i^1 < 1] + [x_i^1 \geq 1] \cdot \frac{1}{4}$$



2 шаг:
После получения алгоритма получим новую обучающую выборку:
$\epsilon_i = y_i - a_1(x_i)$

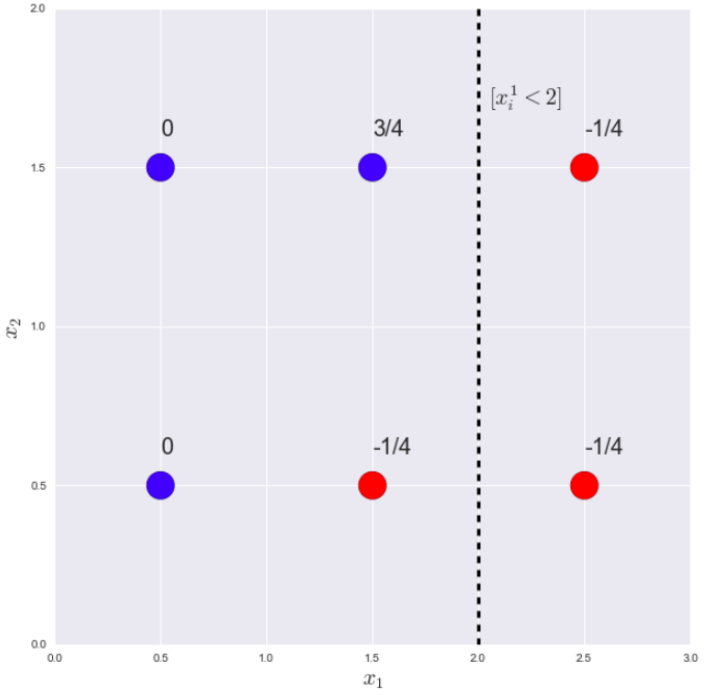


$\bar{y}_{all} = \frac{1}{6}[\frac{3}{4} + 3 \cdot \frac{-1}{4}] = 0$
$\bar{y^2}_{all} = \frac{1}{6}[\frac{9}{16} + \frac{3}{16}] = \frac{1}{8} = H(R_m)$

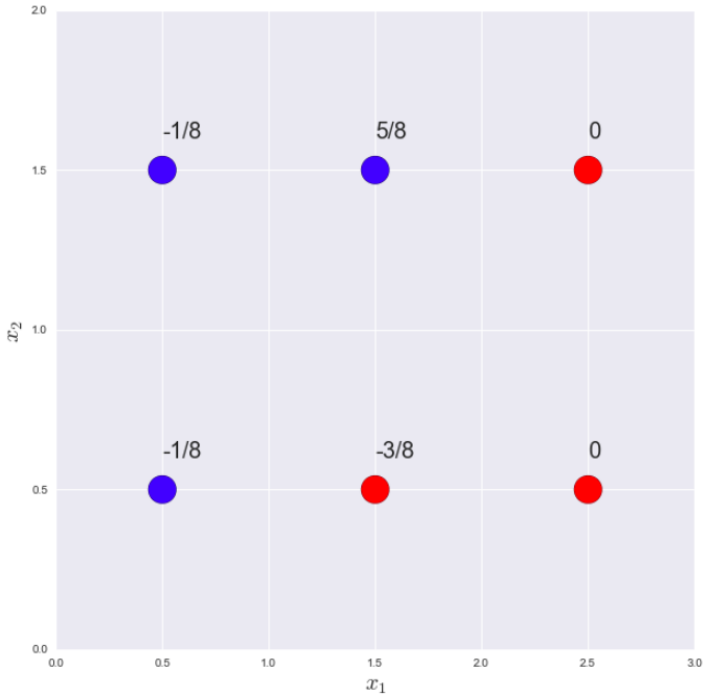
Критерий	$H(R_m)$	$H(R_l)$	$H(R_r)$	$F(R_m, R_r, R_l)$
$[x_i^1 < 2]$	$\frac{1}{8}$	$\bar{y}_l = \frac{1}{8}$ $\bar{y^2}_l = \frac{5}{32}$ $H(R_l) = \frac{10}{64} - \frac{1}{64} = \frac{9}{64}$	0	$\frac{1}{8} - \frac{4}{6} \cdot \frac{9}{64} = \frac{1}{32}$
$[x_i^2 < 1]$	$\frac{1}{8}$	$\bar{y}_l = -\frac{1}{6}$ $\bar{y^2}_l = \frac{1}{24}$ $H(R_l) = \frac{1}{24} - \frac{1}{36} = \frac{1}{72}$	$\bar{y}_r = \frac{1}{6}$ $\bar{y^2}_r = \frac{5}{24}$ $H(R_r) = \frac{5}{24} - \frac{1}{36} = \frac{13}{72}$	$\frac{1}{8} - \frac{1}{2} \cdot \frac{14}{72} = \frac{1}{36}$

Итак:
$b_2(x) = \frac{1}{8} \cdot [x_i^1 < 2] - \frac{1}{4} \cdot [x_i^1 \geq 2]$

$$a_2(x) = b_0(x) + b_1(x) + b_2(x) = 1 \cdot [x_i^1 < 1] + [x_i^1 \geq 1] \cdot \frac{1}{4} + \frac{1}{8} \cdot [x_i^1 < 2] - \frac{1}{4} \cdot [x_i^1 \geq 2]$$

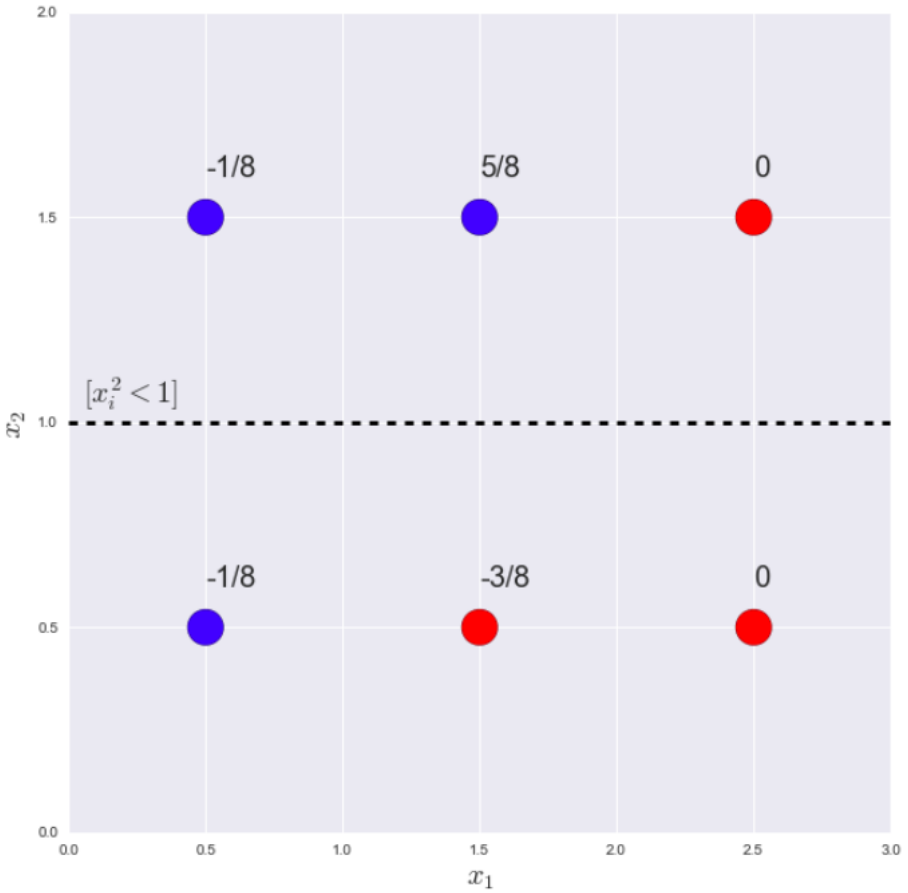


3 шаг:
Получаем новую выборку: $\epsilon_i = y_i - a_2(x_i)$

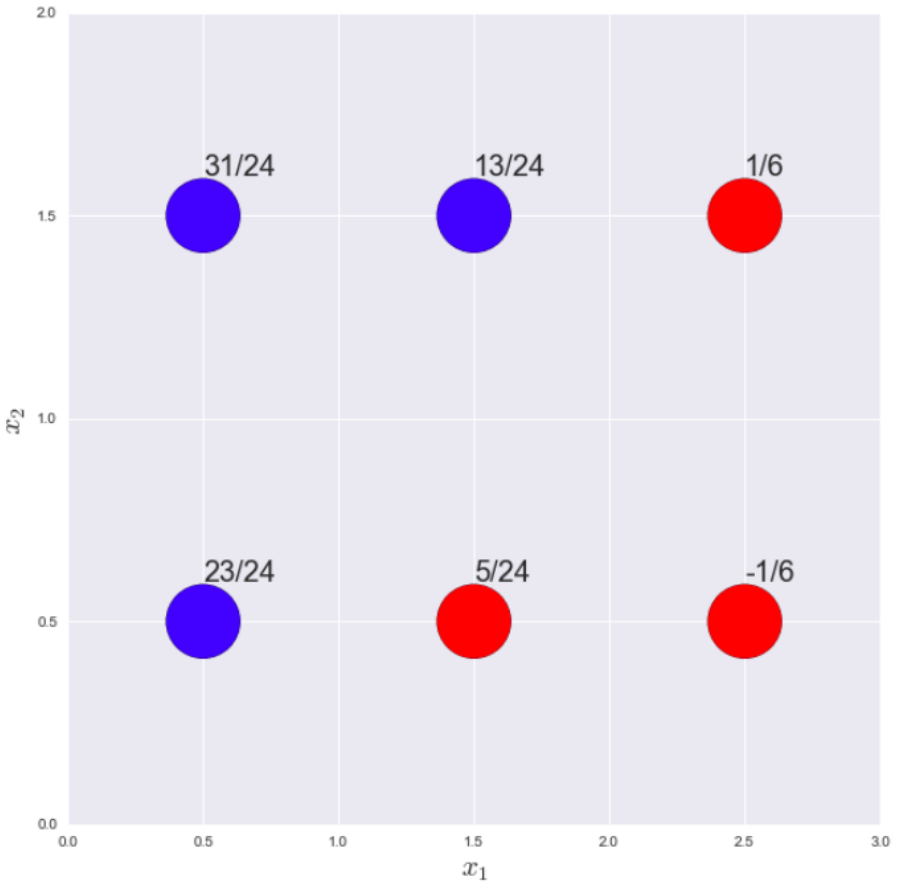


Критерий	$H(R_m)$	$H(R_l)$	$H(R_r)$	$F(R_m, R_r, R_l)$
$[x_i^1 < 1]$	$\frac{3}{32}$	0	$\bar{y_r} = \frac{1}{16}$ $\bar{y_r^2} = \frac{17}{128}$ $H(R_r) = \frac{33}{256}$	$\frac{3}{32} - \frac{2}{3} \cdot \frac{33}{256} = \frac{1}{128} \approx 0.0078$
$[x_i^2 < 1]$	$\frac{3}{32}$	$\bar{y_l} = -\frac{1}{6}$ $\bar{y_l^2} = \frac{10}{192}$ $H(R_l) = \frac{1}{48}$	$\bar{y_r} = \frac{1}{6}$ $\bar{y_r^2} = \frac{13}{96}$ $H(R_r) = \frac{13}{208}$	$\frac{3}{32} - \frac{1}{2} \cdot \frac{1}{48} - \frac{1}{2} \cdot \frac{31}{208} \approx 0.0088$

Итак: $b_3(x) = -\frac{1}{6} \cdot [x_i^2 < 1] + \frac{1}{6} \cdot [x_i^2 \geq 1]$
$a_3(x) = b_0(x) + b_1(x) + b_2(x) + b_3(x) = 1 \cdot [x_i^1 < 1] + [x_i^1 \geq 1] \cdot \frac{1}{4} + \frac{1}{8} \cdot [x_i^1 < 2] - \frac{1}{4} \cdot [x_i^1 \geq 2] - \frac{1}{6} \cdot [x_i^2 < 1] + \frac{1}{6} \cdot [x_i^2 \geq 1]$



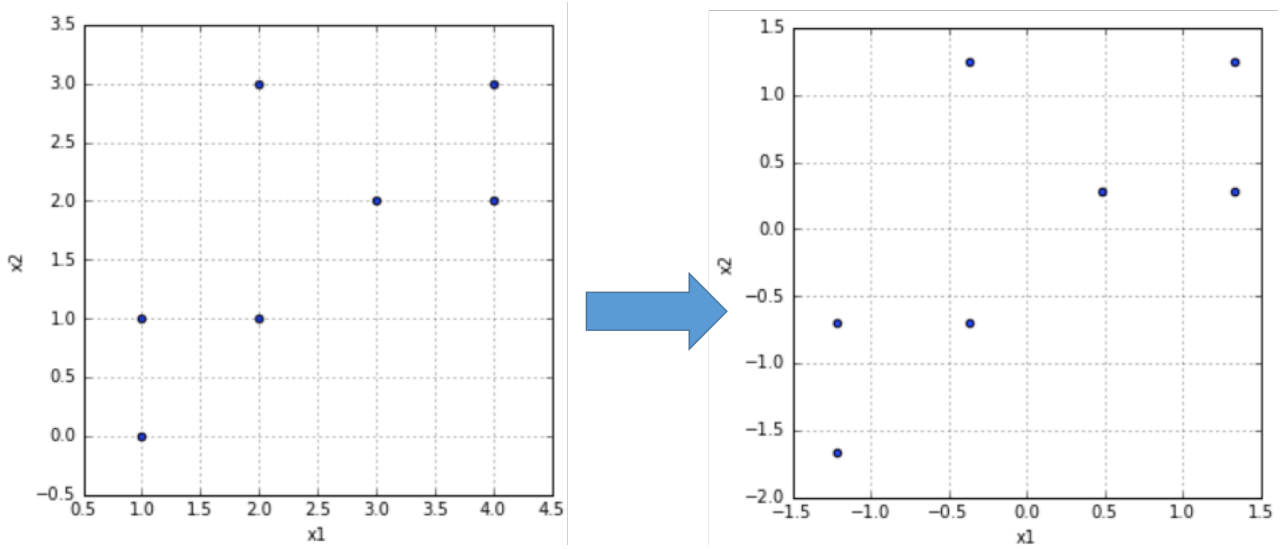
Результатирующий алгоритм можно представить как:



0.10 PCA

0.10.1 Задача 1

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

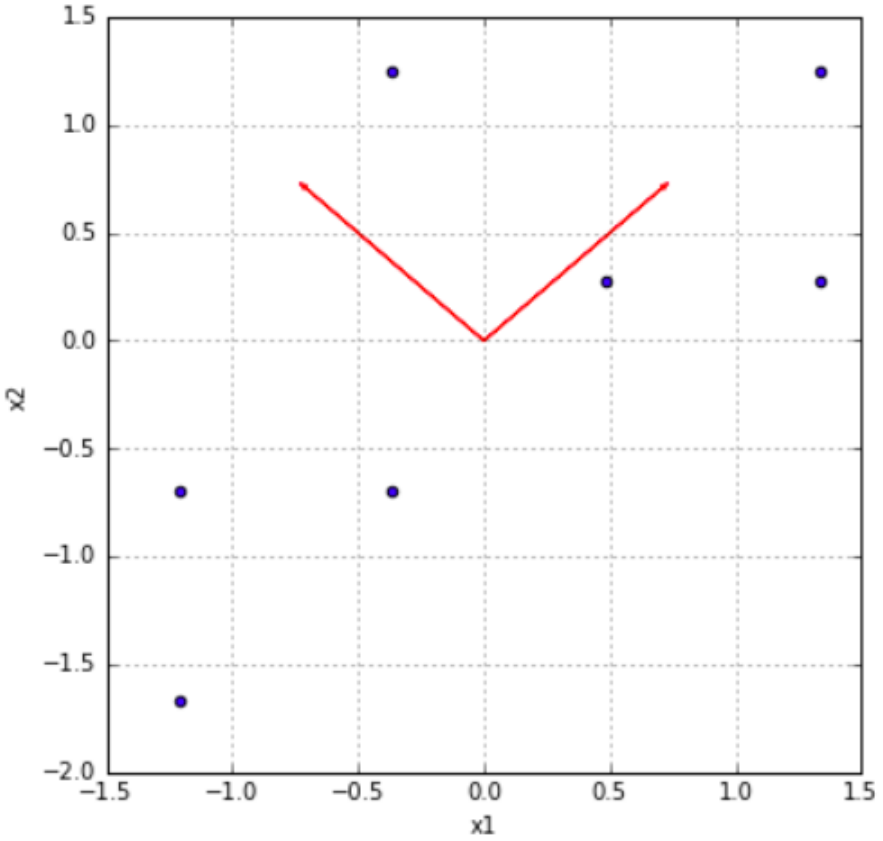


Решение
Перед непосредвственным решением задачи данные нужно нормировать и центрировать $X \to X'$.
$\mathbb{E}x^{(1)} = 2.43, \quad \sigma(x^{(1)}) = 1.18$
$\mathbb{E}x^{(2)} = 1.71, \quad \sigma(x^{(2)}) = 1$

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

Задача покомпонентного нахождения главных компонент ставится следующим образом:
$\begin{cases} \|Xa\|_2^2 \to \max\limits_a \\ \|a\|_2^2 = 1 \end{cases}$, где a – первая главная компонента
$L = a^T X^T X a + \lambda(a^T a - 1)$
$\nabla_a L = 2X^T X a + 2\lambda a = 0 \Rightarrow X^T X a = -\lambda a \Rightarrow$ решением этого уравнение 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^TX, следовательно нам нужно найти спектральное разложение матрицы X^TX.
Так как матрица X^TX – симметричная, то для неё существует спектральное разложение: $X^TX = UDU^T$
Итак: $X^TX = \begin{bmatrix} 7 & 4.83 \\ 4.83 & 7 \end{bmatrix} = \begin{bmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{bmatrix} \cdot \begin{bmatrix} 11.83 & 0 \\ 0 & 2.17 \end{bmatrix} \cdot \begin{bmatrix} 0.71 & 0.71 \\ -0.71 & 0.71 \end{bmatrix}$ (с точностью до округлений).
Тогда $\boxed{a_1 = (0.71, 0.71), \quad a_2 = (-0.71, 0.71)}$

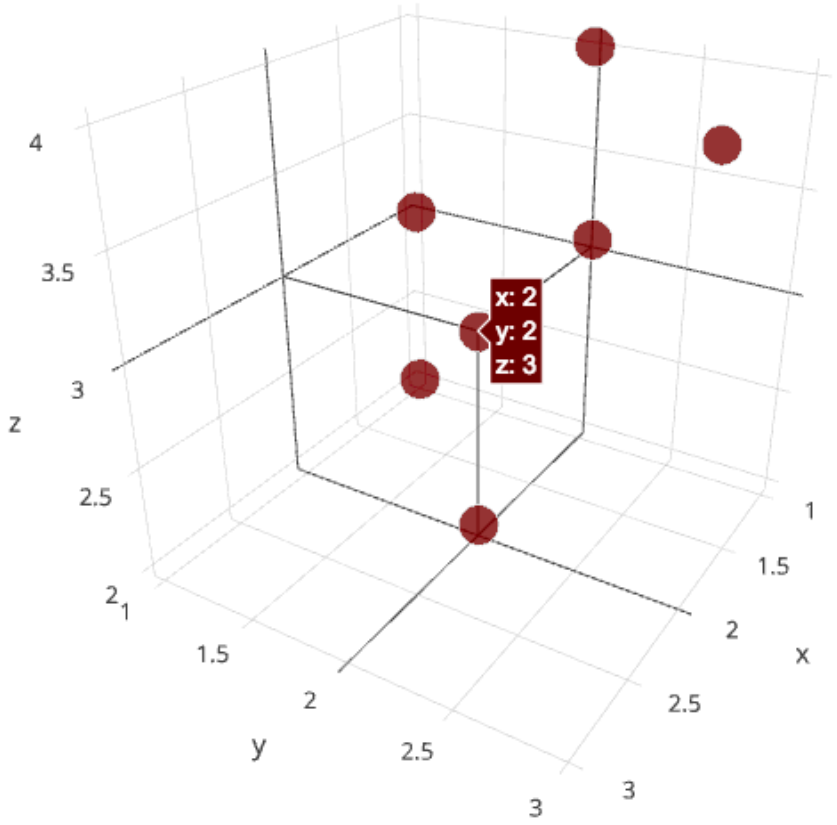


0.10.2 Задача 2

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: $X \to X'$, где X':

```
[-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. 2: X'

2) Найдём ковариационную матрицу для данного набора: $X'^T X'$ и представим её в виде собственного разложения (eigendecomposition):
$X'^T X' = S \cdot D \cdot S^T$

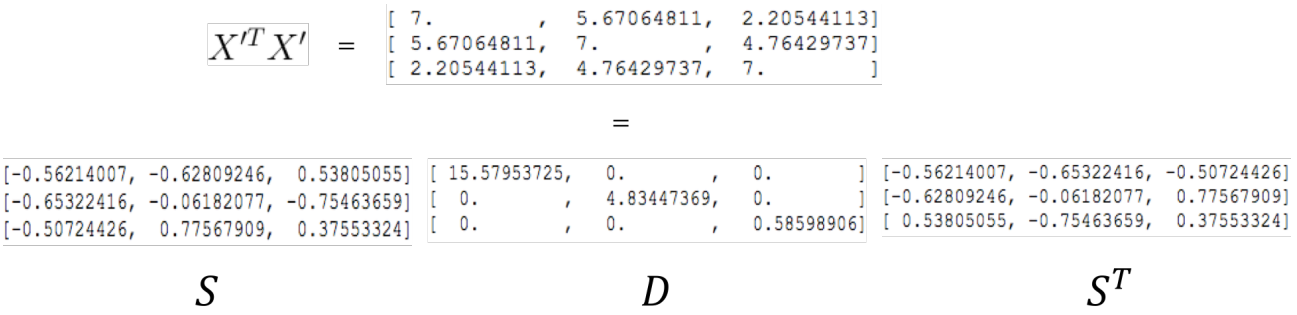
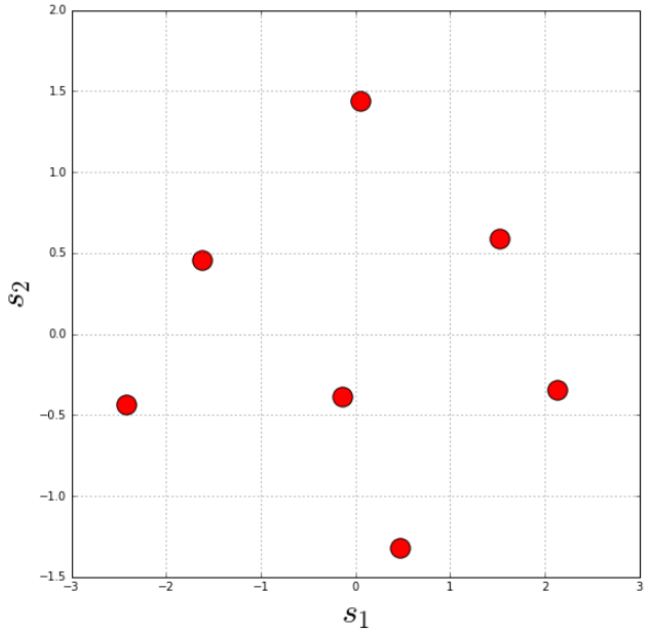


Рис. 3: $X'^T X' = S \cdot D \cdot S^T$

Первые 2 столбца матрицы S это и есть первые 2 главные компоненты.
Заметим, что если бы мы просто хотели представить данные в новом бзисе 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: $\|x - \alpha_1 \cdot s_1 - \alpha_2 \cdot s_2\|_2 \to \min\limits_{\alpha_i}$
Её можно представить в следующем эквивалентном виде:
$\|x - S \cdot \alpha\|_2^2 \to \min\limits_{\alpha}$, где S – матрица, по столбцам которой записаны собственные вектора, α – вектор, отвечающий за координаты вектора x в базисе s_j
$\|x - S \cdot \alpha\|_2^2 = (x^T - \alpha^T S^T)(x - S\alpha) = x^T x - 2x^T S\alpha + \alpha^T S^T S\alpha \to \min\limits_{\alpha}$
$\nabla_\alpha = -2S^T x + 2S^T S\alpha = 0 \Rightarrow \alpha = (S^T S)^{-1} S^T x$
Заметим, что матрица S – ортогональная, а значит: $S^T S = I$ и следовательно: $\alpha = S^T x$
Итого получаем: $X'_{projection} = S^T \cdot X'^T$:

```
[-0.56214007, -0.65322416, -0.50724426]
[-0.62809246, -0.06182077,  0.77567909]
```
$\cdot$
```
[-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 ]
```
=
```
[ 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]
```

Рис. 4:



0.10.3 Задача 3 (Вероятностная постановка PCA)

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

Решение

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

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

$\log L =$

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

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

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

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

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

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

0.10.4 Задача 4

Пусть $x_1, x_2 \in \mathbb{R}^d$ и $y_1, y_2 \in \mathbb{R}^k,\ k << d : \hat{x}_1 = c + Sy_1,\ \hat{x}_2 = c + Sy_2, S$ – матрица, у которой по столбцам записаны собственные векторы ковариационной матрицы нормированные (k штук). Доказать, что $\|\hat{x}_1 - \hat{x}_2\|_2^2 = \|y_1 - y_2\|_2^2$

Решение

$\|x_1 - x_2\|_2^2 = (x_1 - x_2)^T(x_1 - x_2) = (Sy_1 - Sy_2)^T(Sy_1 - Sy_2) =$
$(y_1 - y_2)^T \underbrace{S^T S}_{I}(y_1 - y_2) = (y_1 - y_2)^T(y_1 - y_2) = \|y_1 - y_2\|_2^2$

0.11 EM алгоритм

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

0.11.1 Задача 1

Пусть $x_1,\ldots,x_N$ – 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-алготима 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 $\beta^{[x_i=3]}\cdot(1-\beta)^{[x_i=2]}$

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

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

$$\mathsf{P}(z_i=1|x_i,\theta) = \frac{\mathsf{P}(z_i=1)\cdot\mathsf{P}(x_i|z_i=1,\theta)}{\sum_{k=0}^{1}\mathsf{P}(z_i=k)\cdot\mathsf{P}(x_i|z_i=k,\theta)} = \{\texttt{в силу формулы Байеса}\} =$$

$$\underbrace{\frac{\gamma\cdot\alpha^{[x_i=1]}\cdot(1-\alpha)^{[x_i=2]}}{\gamma\cdot\alpha^{[x_i=1]}\cdot(1-\alpha)^{[x_i=2]} + (1-\gamma)\cdot\beta^{[x_i=3]}\cdot(1-\beta)^{[x_i=2]}}}_{g_{i1}}$$

$$\mathsf{P}(z_i=0|x_i,\theta) = \underbrace{\frac{(1-\gamma)\cdot\beta^{[x_i=3]}\cdot(1-\beta)^{[x_i=2]}}{(1-\gamma)\cdot\beta^{[x_i=3]}\cdot(1-\beta)^{[x_i=2]} + \gamma\cdot\alpha^{[x_i=1]}\cdot(1-\alpha)^{[x_i=2]}}}_{g_{i0}}$$

$\mathbb{M}$-шаг:
$Q = \mathbb{E}_z \log\mathbb{P}(\mathbb{X},\mathbb{Z}|\Theta) \to \max\limits_{\gamma,\alpha,\beta}$

$$Q = \mathbb{E}_z\left(\sum_{i=1}^{N} z_i\cdot\{\log\gamma + [x_i=1]\log\alpha + [x_i=2]\log(1-\alpha)\}\right) +$$

$$\mathbb{E}_z\left(\sum_{i=1}^{N}(1-z_i)\cdot\{\log(1-\gamma) + [x_i=2]\log(1-\beta) + [x_i=3]\log\beta\}\right) =$$

$$\{\mathbb{E}_{z_i}(z_i) = \mathsf{P}(z_i=1|x_i,\theta) = g_{i1}, \mathbb{E}_{z_i}(1-z_i) = \mathsf{P}(z_i=0|x_i,\theta) = g_{i0}\} =$$

$\sum\limits_{i=1}^{N} g_{i1} \cdot \{\log\gamma + [x_i = 1]\log\alpha + [x_i = 2]\log(1-\alpha)\} + \sum\limits_{i=1}^{N} g_{i0} \cdot \{\log(1-\gamma) + [x_i = 2]\log(1-\beta) + [x_i = 3]\log\beta\} \to \max\limits_{\alpha,\beta,\gamma}$

Итак:

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

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

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

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

$\frac{\sum\limits_{i=1}^{N} g_{i1} \cdot [x_i=1]}{\sum\limits_{i=1}^{N} g_{i1} \cdot ([x_i=1]+[x_i=2])} \Rightarrow \boxed{\hat{\alpha} = \frac{\sum\limits_{x_i=1} g_{i1}}{\sum\limits_{x_i=1} g_{i1} + \sum\limits_{x_i=2} g_{i1}}}$

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

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

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

$\sum\limits_{i=1}^{N} g_{i1} = \underbrace{\sum\limits_{x_i=1} g_{i1}}_{a} + \underbrace{\sum\limits_{x_i=2} g_{i1}}_{b} + \underbrace{\sum\limits_{x_i=3} g_{i1}}_{c}$

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

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

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

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

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

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

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

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

$\sum\limits_{i=1}^{N} g_{i0} = \underbrace{\sum\limits_{x_i=1} g_{i0}}_{d} + \underbrace{\sum\limits_{x_i=2} g_{i0}}_{e} + \underbrace{\sum\limits_{x_i=3} g_{i0}}_{f}$

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

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

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

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

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 E-шага:

$\sum\limits_{i=1}^{N} g_{i1} = \underbrace{\sum\limits_{x_i=1} g_{i1}}_{a} + \underbrace{\sum\limits_{x_i=2} g_{i1}}_{b} + \underbrace{\sum\limits_{x_i=3} g_{i1}}_{c}$

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

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

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

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

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

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

$\sum\limits_{i=1}^{N} g_{i0} = \underbrace{\sum\limits_{x_i=1} g_{i0}}_{d} + \underbrace{\sum\limits_{x_i=2} g_{i0}}_{e} + \underbrace{\sum\limits_{x_i=3} g_{i0}}_{f}$

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

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

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

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

Пересчёт формул M-шага:

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

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

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

После второй итерации:

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