Классификация текстов

Задача классификации

D – коллекция текстов

C – множество тематических рубрик

Задача: найти f, такую, что

$$f: D \times C \to \{0, 1\}$$

или

$$f: D \times C \to [0, 1]$$

Области применения

- Почтовые сообщения: спам – не спам
- Отзывы покупателей на товар, услугу: положительный, нейтральный, отрицательный
- Отзывы на фильм, музыкальное произведение, книгу и т.д.

Представление текстов коллекции в виде числовых векторов

- Коллекция текстов:
  - Текст № 1 посвящен роли воды в жизни растений
  - Текст № 2 посвящен понятию влажности, документ
  - Текст № 3 посвящен описанию почвы
  - Текст № 4 посвящен экологическим проблемам
  - Тексты № 5-8 содержат советы по поливу растений

Представление текстов коллекции в виде числовых векторов

w_{ti}	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8
1-Гибель	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$	$w_{1,5}$	$w_{1,6}$	$w_{1,7}$	$w_{1,8}$
2-Горизонт	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	$w_{2,4}$	$w_{2,5}$	$w_{2,6}$	$w_{2,7}$	$w_{2,8}$
3-Грунт	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$	$w_{3,4}$	$w_{3,5}$	$w_{3,6}$	$w_{3,7}$	$w_{3,8}$
4-Деревья	$w_{4,1}$	$w_{4,2}$	$w_{4,3}$	$w_{4,4}$	$w_{4,5}$	$w_{4,6}$	$w_{4,7}$	$w_{4,8}$
5-Загнивание	$w_{5,1}$	$w_{5,2}$	$w_{5,3}$	$w_{5,4}$	$w_{5,5}$	$w_{5,6}$	$w_{5,7}$	$w_{5,8}$
6-Климат	$w_{6,1}$	$w_{6,2}$	$w_{6,3}$	$w_{6,4}$	$w_{6,5}$	$w_{6,6}$	$w_{6,7}$	$w_{6,8}$
7-Обмен	$w_{7,1}$	$w_{7,2}$	$w_{7,3}$	$w_{7,4}$	$w_{7,5}$	$w_{7,6}$	$w_{7,7}$	$w_{7,8}$
8-Организм	$w_{8,1}$	$w_{8,2}$	$w_{8,3}$	$w_{8,4}$	$w_{8,5}$	$w_{8,6}$	$w_{8,7}$	$w_{8,8}$
9-Поглощение	$w_{9,1}$	$w_{9,2}$	$w_{9,3}$	$w_{9,4}$	$w_{9,5}$	$w_{9,6}$	$w_{9,7}$	$w_{9,8}$
10-Погода	$w_{10,1}$	$w_{10,2}$	$w_{10,3}$	$w_{10,4}$	$w_{10,5}$	$w_{10,6}$	$w_{10,7}$	$w_{10,8}$
11-Полив	$w_{11,1}$	$w_{11,2}$	$w_{11,3}$	$w_{11,4}$	$w_{11,5}$	$w_{11,6}$	$w_{11,7}$	$w_{11,8}$
12-Поры	$w_{12,1}$	$w_{12,2}$	$w_{12,3}$	$w_{12,4}$	$w_{12,5}$	$w_{12,6}$	$w_{12,7}$	$w_{12,8}$
13-Рыхление	$w_{13,1}$	$w_{13,2}$	$w_{13,3}$	$w_{13,4}$	$w_{13,5}$	$w_{13,6}$	$w_{13,7}$	$w_{13,8}$

Представление текстов коллекции в виде числовых векторов – число вхождений слова

n_{ti}	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8
Гибель	2	0	0	2	4	0	2	0
Горизонт	0	10	1	0	0	0	0	0
Грунт	0	8	0	0	3	1	0	0
Деревья	0	1	4	22	0	0	16	13
Загнивание	0	0	0	0	5	0	0	0
Климат	2	1	3	0	1	0	0	0
Обмен	6	0	0	1	0	1	0	0
Организм	17	0	0	1	0	0	0	0
Поглощение	0	0	23	3	0	1	1	1
Погода	0	0	0	0	0	1	1	1
Полив	0	1	0	0	6	12	15	19
Поры	0	6	0	0	0	0	0	1
Рыхление	0	0	0	0	0	4	2	1

Представление текстов коллекции в виде числовых векторов

- Методы построения векторов:
  - Бинарные векторы: $w_{ti} = \{0, 1\}$
  - Схема TF-IDF:

$$tf(t, d_i) = \frac{n_{ti}}{|d_i|}$$

$$idf(t, D) = log \frac{|D|}{|\{d_k \in D: \ t \in d_k\}|}$$

$$w_{ti} = tf(t, d_i) \times idf(t, D)$$

Представление текстов коллекции в виде числовых векторов – бинарные векторы

	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8
Гибель	1	0	0	1	1	0	1	0
Горизонт	0	1	1	0	0	0	0	0
Грунт	0	1	0	0	1	1	0	0
Деревья	0	1	1	1	0	0	1	1
Загнивание	0	0	0	0	1	0	0	0
Климат	1	1	1	0	1	0	0	0
Обмен	1	0	0	1	0	1	0	0
Организм	1	0	0	1	0	0	0	0
Поглощение	0	0	1	1	0	1	1	1
Погода	0	0	0	0	0	1	1	1
Полив	0	1	0	0	1	1	1	1
Поры	0	1	0	0	0	0	0	1
Рыхление	0	0	0	0	0	1	1	1

Представление текстов коллекции в виде числовых векторов – схема TF-IDF

	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8
Гибель	0,035	0	0	0,027	0,201	0	0,038	0
Горизонт	0	0,602	0,026	0	0	0	0	0
Грунт	0	0,341	0	0	0,213	0,035	0	0
Деревья	0	0,020	0,035	0,204	0	0	0,204	0,140
Загнивание	0	0	0	0	0,753	0	0	0
Климат	0,035	0,030	0,039	0	0,050	0	0	0
Обмен	0,150	0	0	0,019	0	0,035	0	0
Организм	0,602	0	0	0,027	0	0	0	0
Поглощение	0	0	0,204	0,028	0	0,017	0,013	0,011
Погода	0	0	0	0	0	0,035	0,027	0,022
Полив	0	0,020	0	0	0,204	0,204	0,191	0,204
Поры	0	0,361	0	0	0	0	0	0,032
Рыхление	0	0	0	0	0	0,142	0,053	0,022

Наивный Байесовский подход

$$c(d_i) = \underset{c_j}{\operatorname{argmax}}\{\log P(c_j|d_i)\} =$$

$$= \underset{c_j}{\operatorname{argmax}}\left\{\log P(c_j) + \sum_{t=1}^{|T|} w_{ti}\left(\log\left(\frac{p_{tj}}{1-p_{tj}}\right)\right) + \sum_{k=1}^{|T|}\log(1-p_{tj})\right\}$$

$P(c_j)$ – оценка вероятности выбрать рубрику c_j

p_{tj} – оценка вероятности встретить слово t в рубрике c_j

Наивный Байесовский подход

Обучающая выборка:

$$P(c_j) = \frac{\left|\{d_k \in c_j\}\right|}{|D|}$$

$$p_{tj} = \frac{\left|\{d_k \in c_j : \ t \in d_k\}\right|}{\left|\{d_k \in c_j\}\right|}$$

$$p_{tj} = \frac{1 + \left|\{d_k \in c_j : \ t \in d_k\}\right|}{2 + \left|\{d_k \in c_j\}\right|}$$

Пример

Обучающая коллекция:

$c_1 = \{d_1, d_2, d_3\}$

$c_2 = \{d_4\}$

$c_3 = \{d_5, d_6, d_7, d_8\}$

Классифицируемый документ: d_9

n_{ti}	№ 9
Гибель	1
Горизонт	0
Грунт	0
Деревья	0
Загнивание	1
Климат	0
Обмен	0
Организм	0
Поглощение	0
Погода	1
Полив	22
Поры	0
Рыхление	0

Пример

$$\underset{c_j}{\operatorname{argmax}}\{\log P(c_j|d_i)\}$$

$$\log P(c_1|\, d_9) \approx -5{,}95$$
$$\log P(c_2|\, d_9) \approx -6{,}06$$
$$\log P(c_3|\, d_9) \approx -4{,}68$$

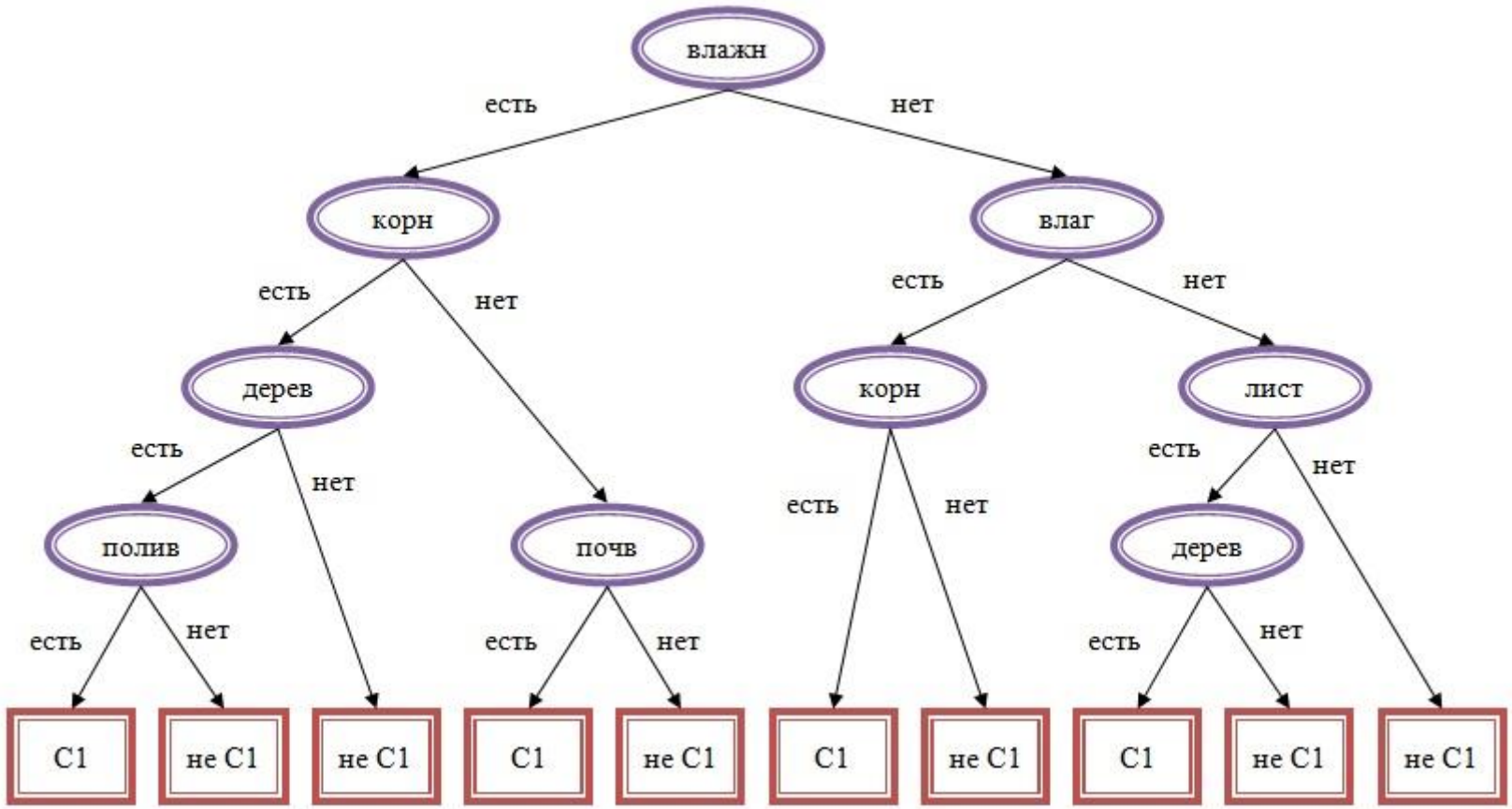
Пример

$$\underset{c_j}{\operatorname{argmax}}\{\log P(c_j|d_i)\}$$

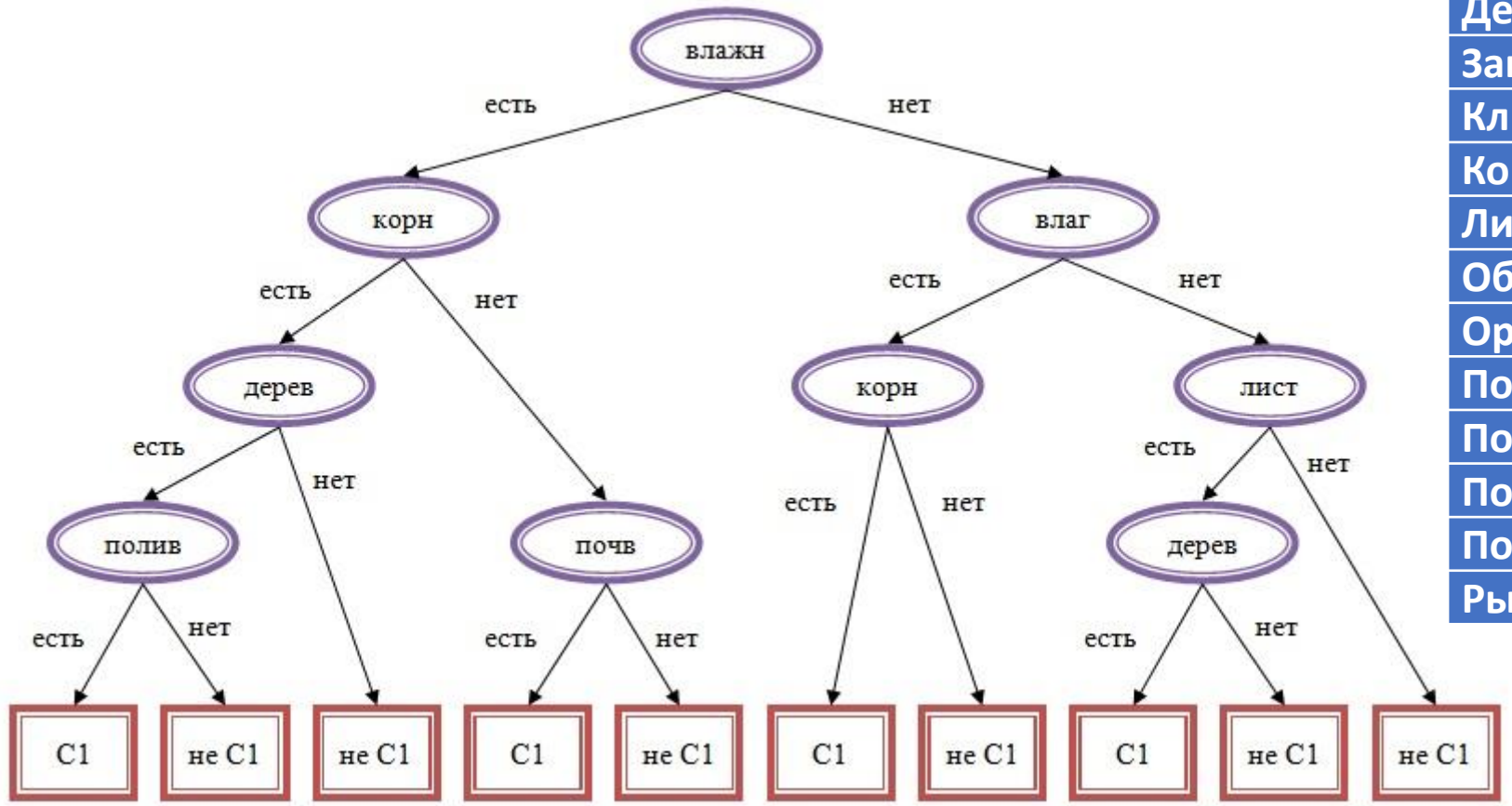
$$\log P(c_1|\, d_9) \approx -5{,}95$$
$$\log P(c_2|\, d_9) \approx -6{,}06$$
$$\log P(c_3|\, d_9) \approx -4{,}68$$

Текст d_9 относится к рубрике c_3

Дерево решений

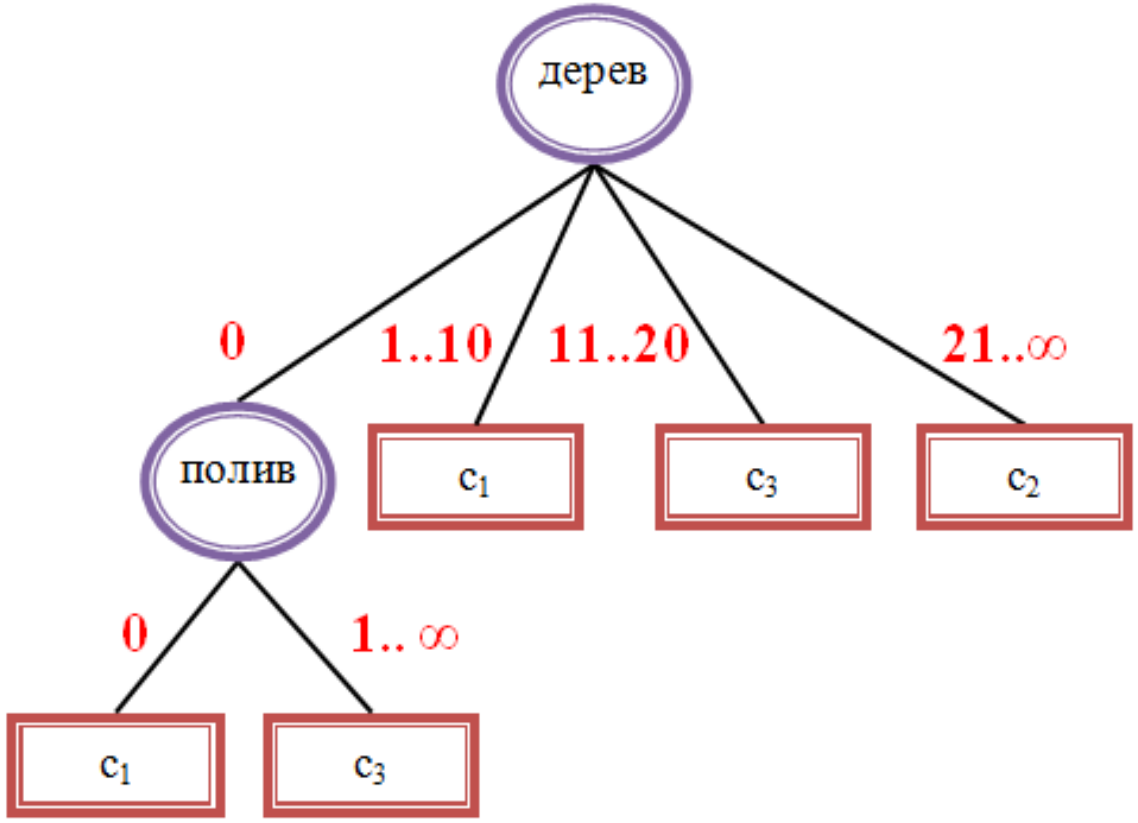


Дерево решений



n_{ij}	№ 9
Влага	3
Влажность	3
Гибель	1
Горизонт	0
Грунт	0
Деревья	0
Загнивание	1
Климат	0
Корень	6
Лист	10
Обмен	0
Организм	0
Поглощение	0
Погода	1
Полив	22
Почва	0
Рыхление	0

Дерево решений



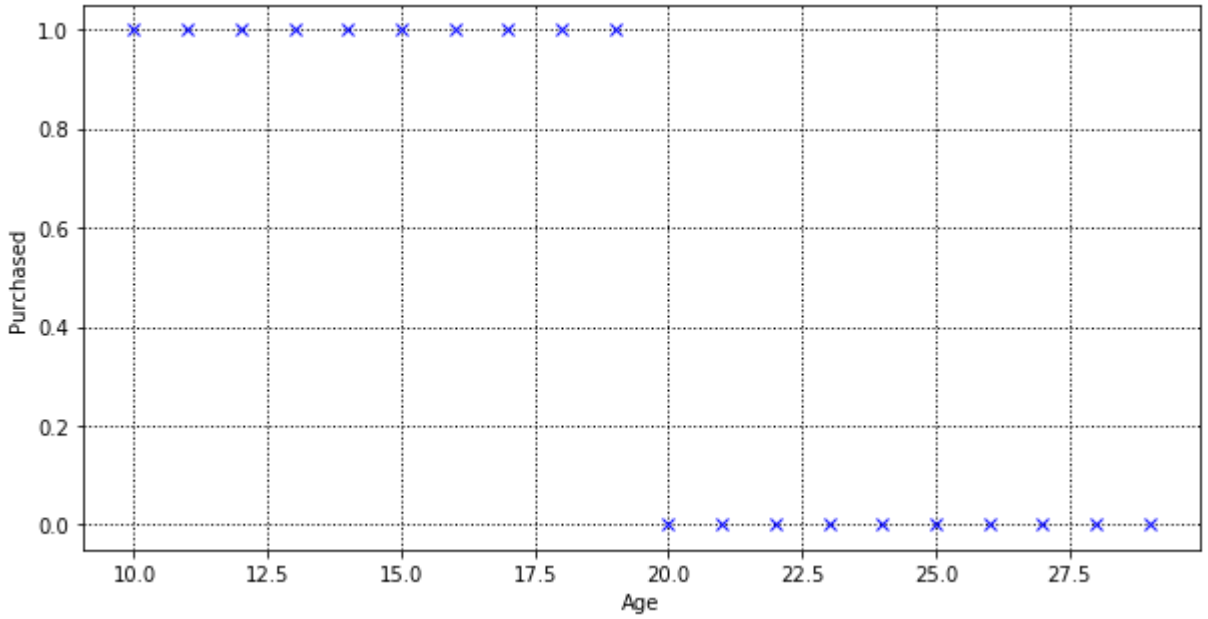
n_{tj}	№ 9
Влага	3
Влажность	3
Гибель	1
Горизонт	0
Грунт	0
Деревья	0
Загнивание	1
Климат	0
Корень	6
Лист	10
Обмен	0
Организм	0
Поглощение	0
Погода	1
Полив	22
Почва	0
Рыхление	0

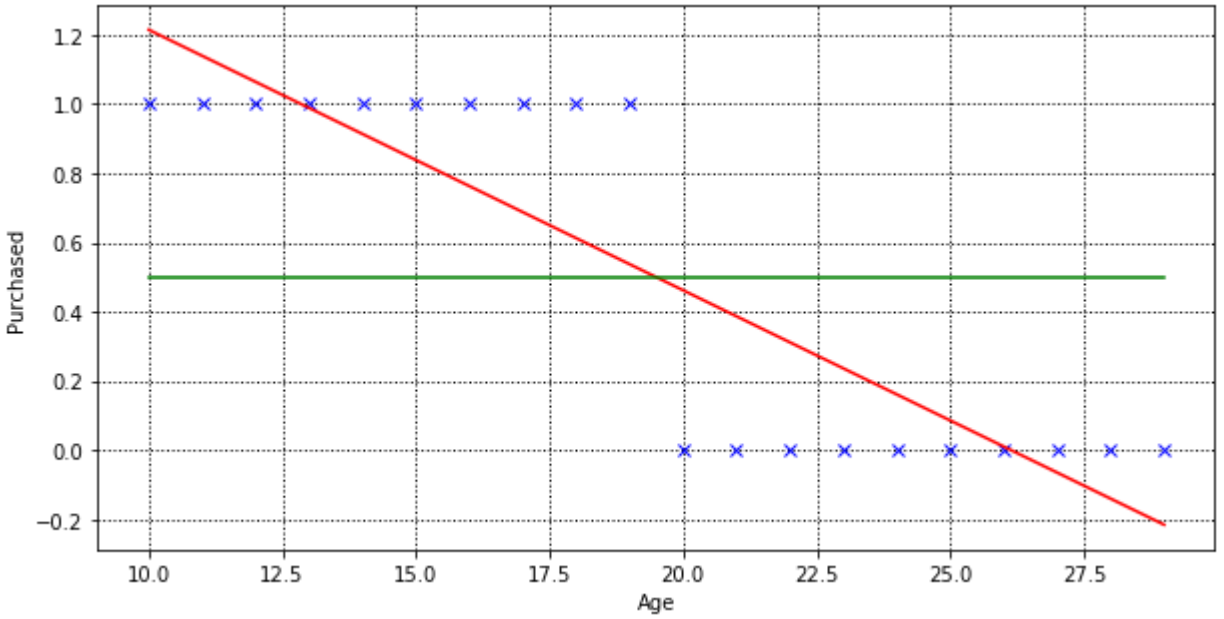
Линейная регрессия

$$y = a_0 + \sum_{j=1}^{n} a_j x_j$$

Линейная регрессия

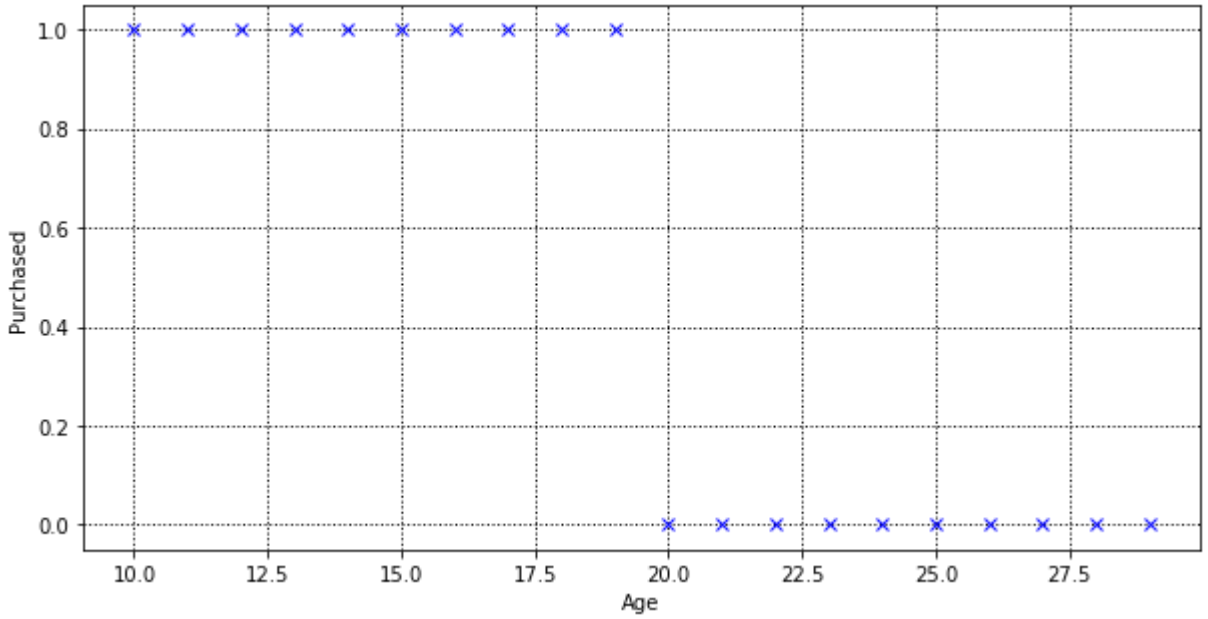
$$Purchased = a_0 + a_1 \cdot Age$$

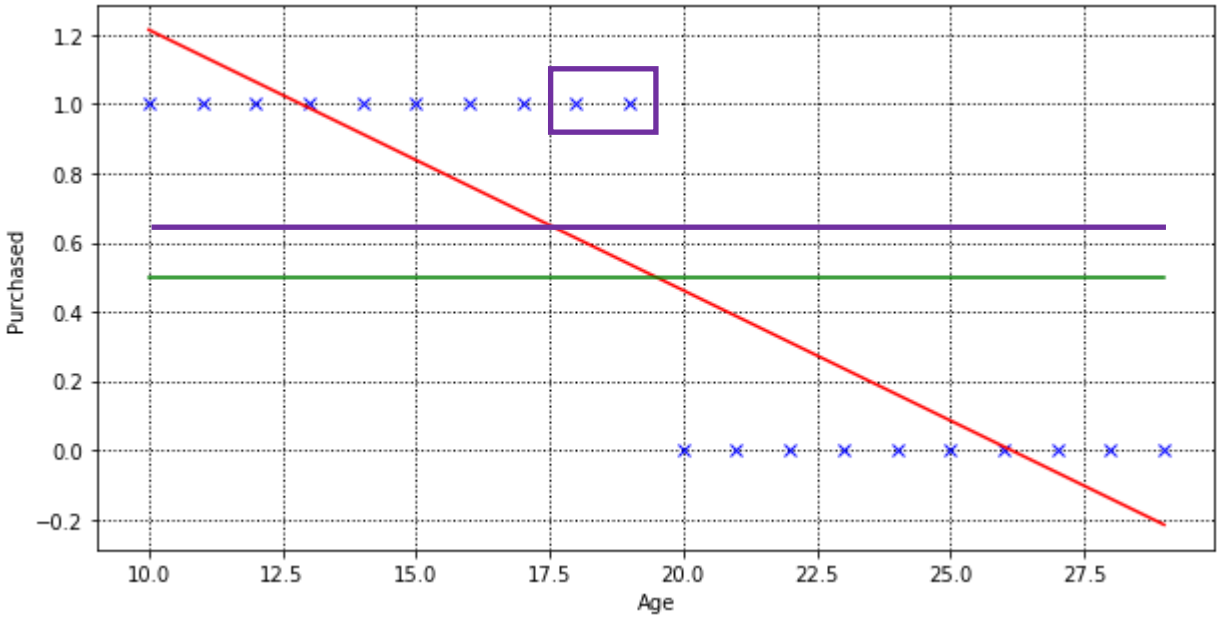




Линейная регрессия

$$Purchased = a_0 + a_1 \cdot Age$$



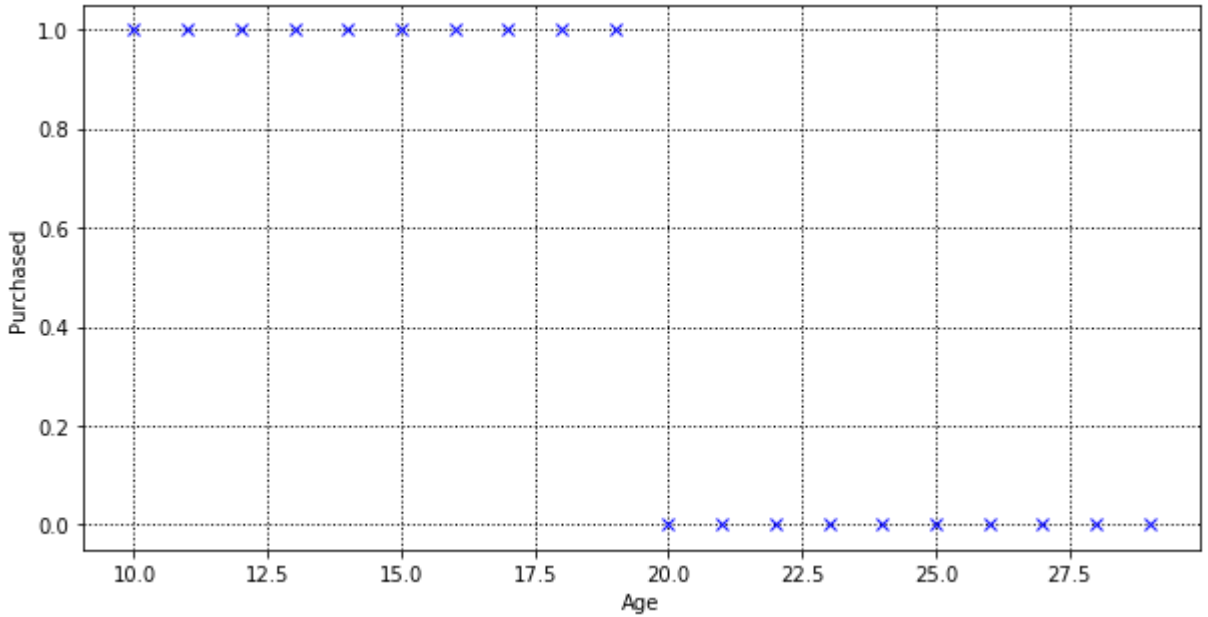


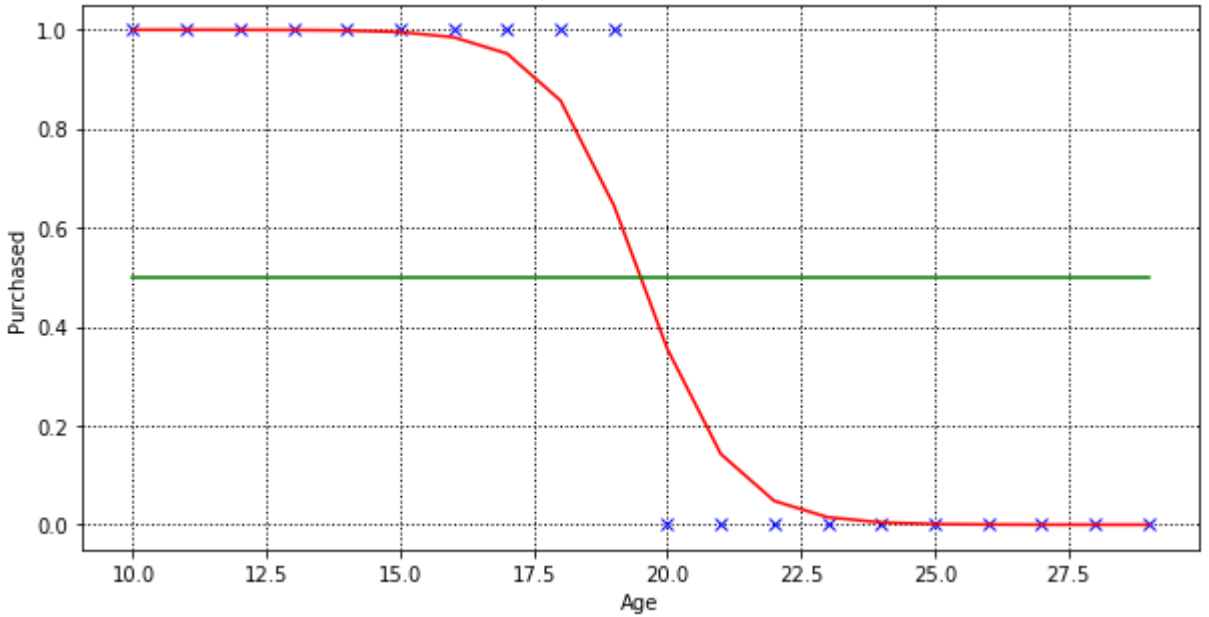
Логистическая регрессия

$$y = \frac{1}{1 + e^{-a_0 - \sum a_j x_j}}$$

Логистическая регрессия

$$Purchased = \frac{1}{1 + e^{-a_0 - a_1 \cdot Age}}$$





Логистическая регрессия

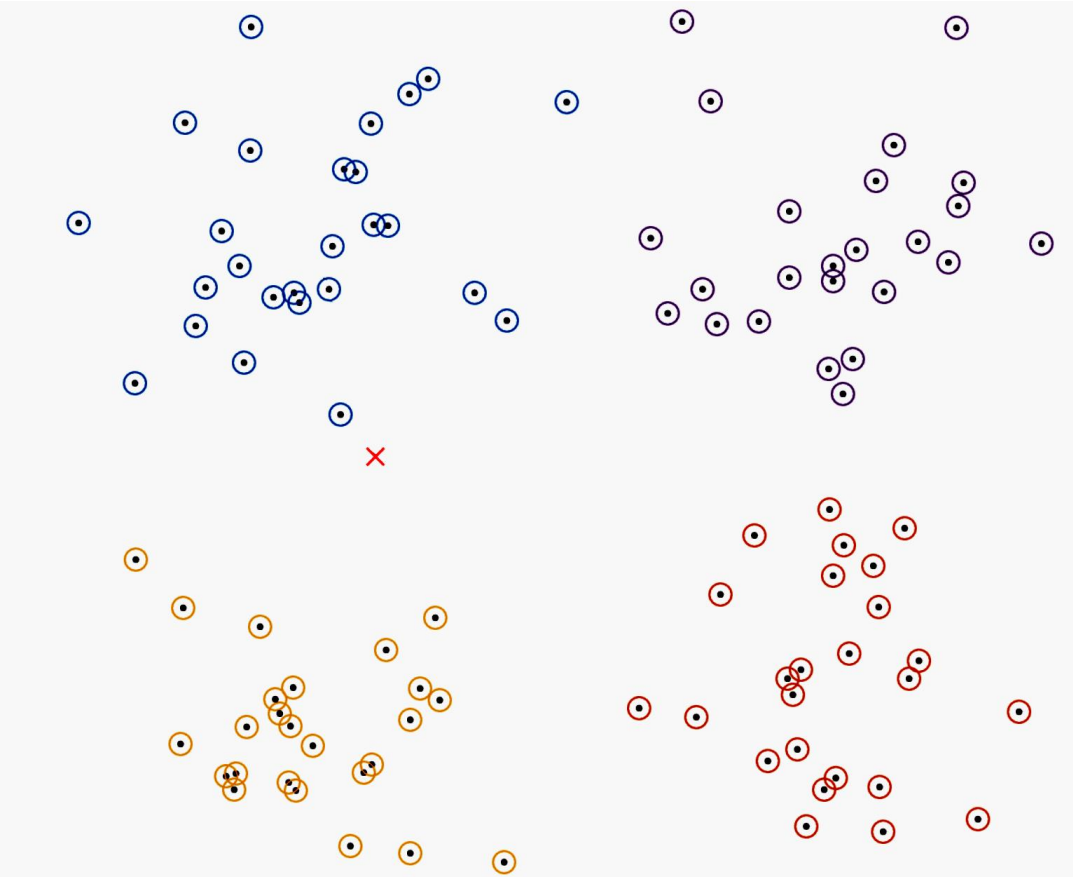
$$c_j(d_i) = \frac{1}{1 + e^{-a_{j0} - \sum_t a_{jt} w_{ti}}}$$

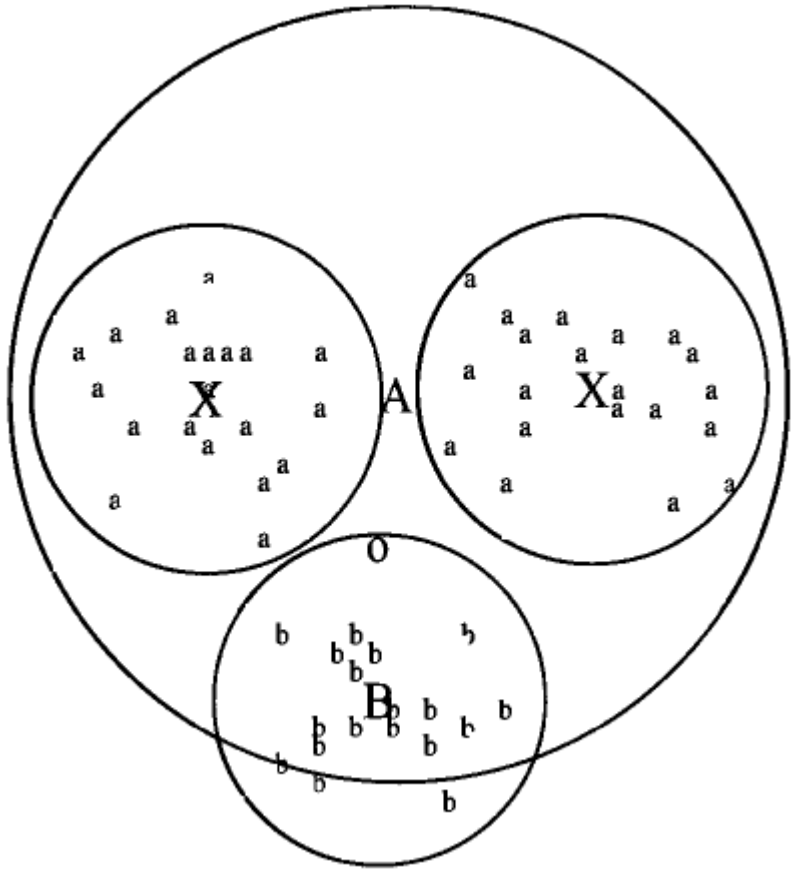
a_{j0}, a_{jt} – оцениваются на основе обучающей коллекции текстов отдельно для каждой тематической рубрики c_j

Метод Rocchio

$$w_{tj} = \beta \cdot \sum_{d_i \epsilon c_j} \frac{w_{ti}}{|\{d_i \epsilon c_j\}|} - \gamma \cdot \sum_{d_i \notin c_j} \frac{w_{ti}}{|\{d_i \notin c_j\}|}$$

$$c(d_i) = \underset{c_j}{\operatorname{argmax}}\{\cos(c_j, d_i)\}$$



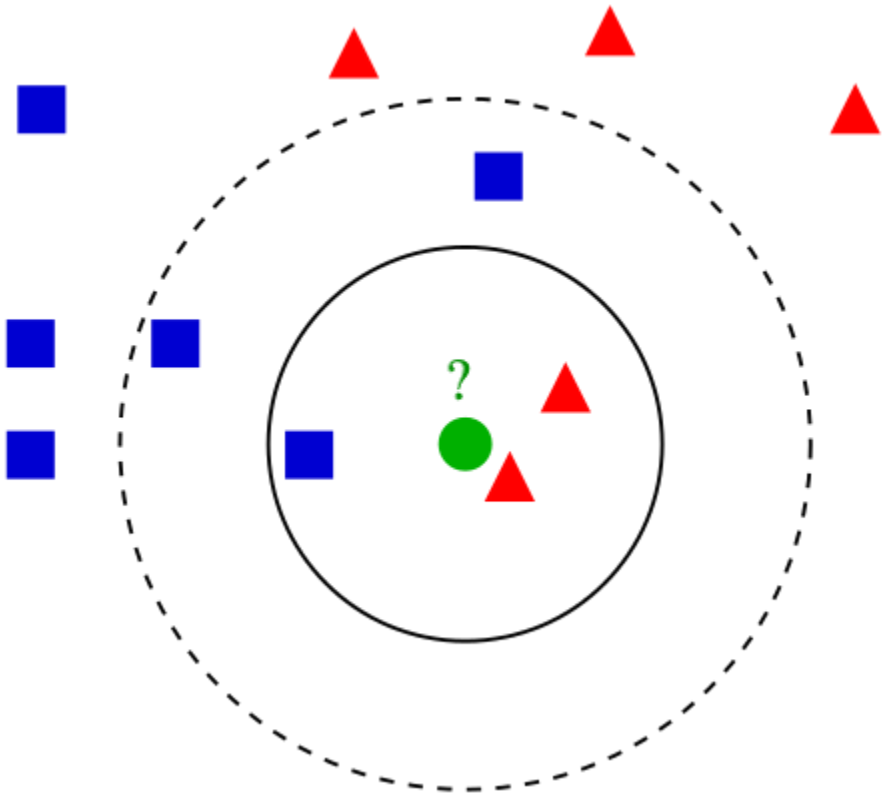


Метод Rocchio

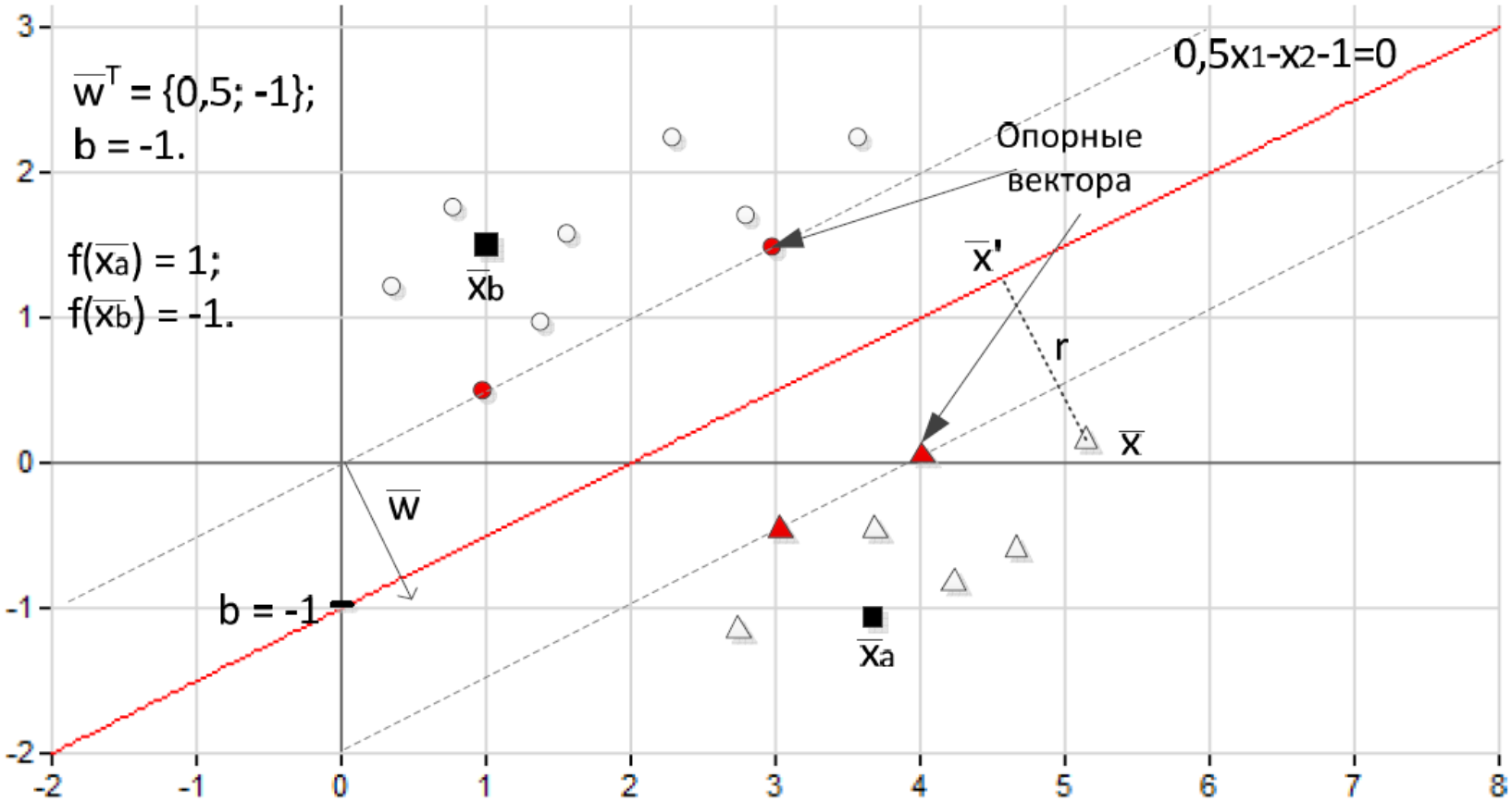
	c_1	c_2	c_3		№ 9
Влага	16,3	0	8		3
Влажность	12,3	10	2,25		3
Гибель	0,67	2	1,5		1
Горизонт	3,67	0	0		0
Грунт	2,67	0	1		0
Деревья	1,67	22	7,25		0
Загнивание	0	0	1,25		1
Климат	2	0	0,25		0
Корень	8	14	6,75		6
Лист	1	8	3,75		10
Обмен	2	1	0,25		0
Организм	5,67	1	0		0
Поглощение	7,67	3	0,75		0
Погода	0	0	0,75		1
Полив	0,33	0	13		22
Почва	29	30	13,5		0
Рыхление	0	0	1,75		0

Категория	$\cos(c_j, d_i)$
c_1	0,157606
c_2	0,184584
c_3	0,675892

Метод k-ближайших соседей



Метод опорных векторов



Оценка классификатора

$$Accuracy = \frac{P}{N} - \text{точность}$$

- P – количество правильных решений
- N – размер выборки

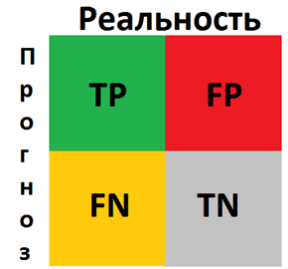
Оценка классификатора

$$Accuracy = \frac{P}{N} - \text{точность}$$

- P – количество правильных решений
- N – размер выборки

Точность (precision) и полнота (recall)

		Исходные данные	
		Относится	Не относится
Решение классификатора	Относится	TP	FP
	Не относится	FN	TN



TP — истинно-положительное решение
TN — истинно-отрицательное решение
FP — ложно-положительное решение
FN — ложно-отрицательное решение

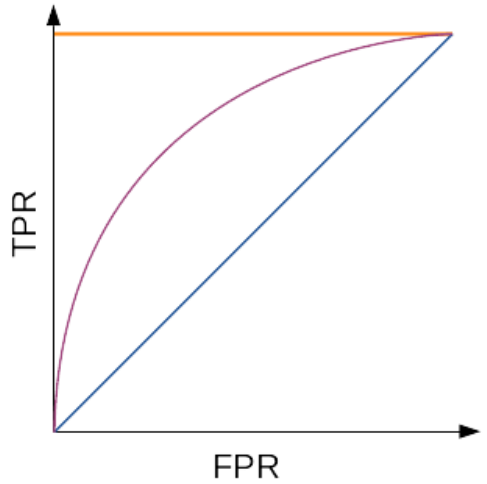
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Площадь под кривой ошибок (AUC ROC)

- ROC-кривая (ROC – receiver operating characteristic – кривая ошибок)
- Качество оценивают как площадь под этой кривой (AUC – area under the curve)



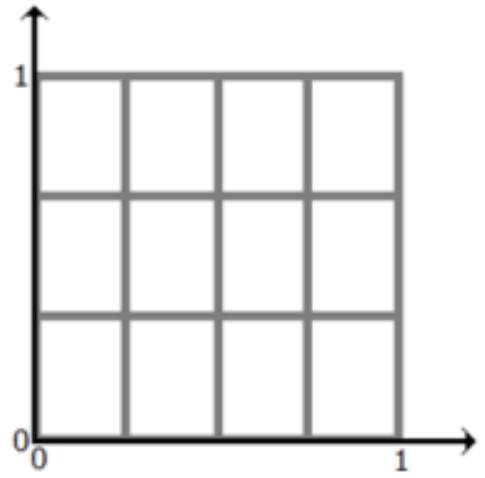
Алгоритм построения ROC

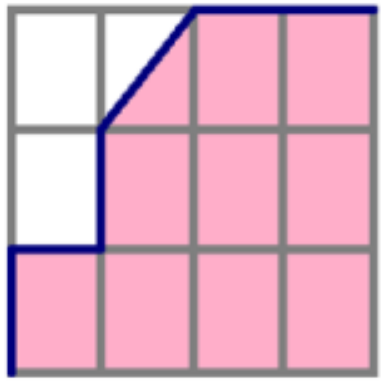
- Единичный квадрат разбиваем сеткой $m \times n$ ячеек
  - m – количество элементов, отнесенных к классу
  - n – количество элементов, не отнесенных к классу
- Для каждого элемента выборки получаем вероятность отнесения его к классу и правильное значение
- Отсортируем элементы по убыванию вероятности
- Начиная с первой строки таблицы и точки (0, 0):
  - Если элемент относится к классу, то делаем шаг вверх
  - Если элемент не относится к классу, то делаем шаг вправо
  - Если несколько элементов имеют одинаковую вероятность, то делаем шаг по диагонали

Площадь под кривой ошибок (AUC ROC)

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0





AUC ROC равен доле пар объектов вида (объект класса 1, объект класса 0), которые алгоритм верно упорядочил, т.е. первый объект идёт в упорядоченном списке раньше

Порог отнесения к классу

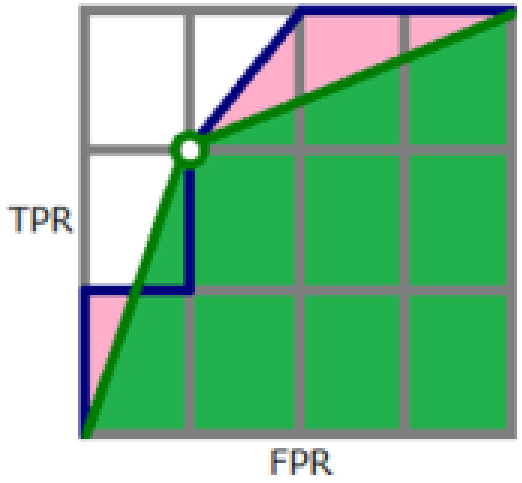
Выбору порога соответствует выбор точки на ROC-кривой (FPR, TPR)

TPR (True Positive Rate) – процент элементов, относящихся к классу, которые верно классифицированы

FPR (False Positive Rate) – процент элементов, не относящихся к классу, которые неверно классифицированы

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

id	> 0.25	класс
4	1	1
1	1	0
6	1	1
3	0	0
5	0	1
2	0	0
7	0	0



Матрица ошибок (Confusion Matrix)

- Матрица ошибок – это матрица размера $N \times N$, где N — это количество классов. Столбцы этой матрицы соответствуют исходным данным, а строки – решениям классификатора
- На пересечении строки класса, который вернул классификатор, и столбца класса, к которому действительно относится документ, находится соответствующее количество элементов выборки

Матрица ошибок (Confusion Matrix)

Правильные значения класса

	0.91	0.96	0.94	0.75	1.00	0.83	0.85	0.97	1.00	0.86	1.00	0.79	1.00	0.75	1.00	1.00	0.96	0.90	0.81	0.89	0.94	0.98	0.86	0.89	0.94	0.92	0.96
0.80		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0.95	1	94	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1.00	2	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.29	3	0	0	6	0	0	3	2	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	3	0	2	0
1.00	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	1
0.92	6	1	0	0	0	0	152	0	0	1	0	0	0	0	0	0	0	1	4	2	3	0	0	0	0	2	0
0.97	7	1	0	1	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0
0.33	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
0.97	9	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0.82	10	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0.87	11	0	0	0	0	0	0	0	0	0	0	34	0	4	0	0	0	0	0	0	0	0	0	1	0	0	0
1.00	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.57	13	0	0	0	0	0	0	0	0	0	0	9	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
0.63	14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	3	0	0	0	0	0	0	0	0	0
0.50	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0	0
0.77	16	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	47	0	1	3	4	0	0	2	0	1	0
0.87	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	69	1	2	5	0	0	0	0	0	0
0.97	18	0	0	0	0	1	4	0	0	1	0	0	0	0	0	0	0	0	197	1	0	0	0	0	0	0	0
0.78	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	35	183	13	0	0	2	0	1	0
0.97	20	0	0	0	0	0	10	3	0	1	0	0	0	0	0	0	0	0	0	4	702	0	0	0	0	6	0
0.93	21	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	2	0	0	0
0.29	22	0	0	1	0	0	2	0	0	6	0	0	0	0	0	0	0	0	1	1	1	0	6	2	0	1	0
0.91	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	6	0	0	115	0	0	0
1.00	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
0.93	25	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	4	5	0	0	0	1	196	0
0.98	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	78