ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

Лекция №3

Классификация

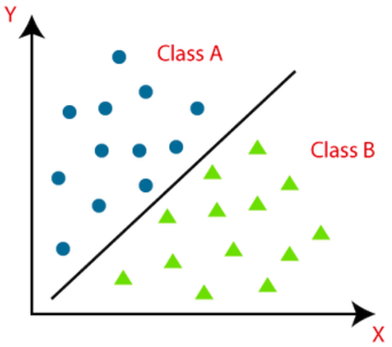
Вспомните, что такое классификация?
Возможно, вы знаете / сможете предложить как измерять качество классификации?

Классификация

Конечное число ответов

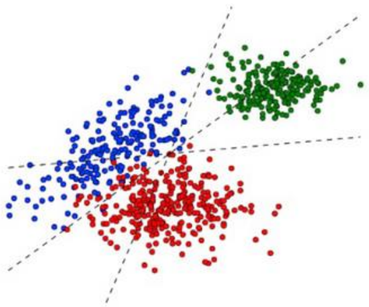
Бинарная:

- $\mathbb{Y}$ = {-1, +1}



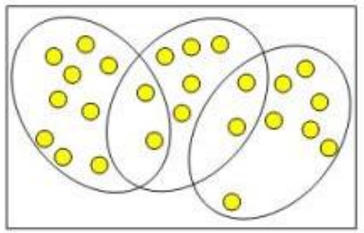
Многоклассовая:

- $\mathbb{Y}$ = {1, 2, …, K}



С пересекающимися классами:

- $\mathbb{Y} = \{0,1\}^{K}$
- Ответ - набор из K нулей и единиц
- i-ый элемент ответа принадлежит i-му классу



Классификация

$$X \in R^{n \times p}$$

$$Y \in C^n \qquad \text{e.g. } C = \{-1, 1\}$$

$$|C| < +\infty$$

$$c(X) = \hat{Y} \approx Y$$

Линейная классификация

The most simple linear classifier

$$c(x) = \begin{cases} 1, & \text{if } f(x) \geq 0 \\ -1, & \text{if } f(x) < 0 \end{cases}$$

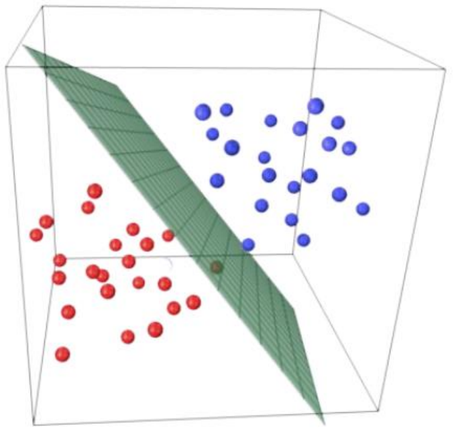
Why cutoff value is fixed?

(bias term is implied)

or equivalently

$$c(x) = \text{sign}(f(x)) = \text{sign}(x^T w)$$

Geometrical interpretation:
hyperplane dividing space into two subspaces



Margin

Let's define linear model's Margin as

$$M_i = y_i \cdot f(x_i) = y_i \cdot x_i^T w$$

main property:
negative margin reveals misclassification

$$M_i > 0 \Leftrightarrow y_i = c(x_i)$$

$$M_i \leq 0 \Leftrightarrow y_i \neq c(x_i)$$

Построение модели

Remembering old paradigm

$$\text{Empirical risk} = \sum_{\text{by objects}} \text{Loss on object} \rightarrow \min_{\text{model params}}$$

Essential loss is misclassification

$$L_{\text{mis}}(y_i^t, y_i^p) = [y_i^t \neq y_i^p] = [M_i \leq 0]$$

Iverson bracket $[P] = \begin{cases} 1, & \text{if } P \text{ is true} \\ 0, & \text{otherwise} \end{cases}$

Disadvantages

- Not differentiable
- Overlooks confidence

Solution:

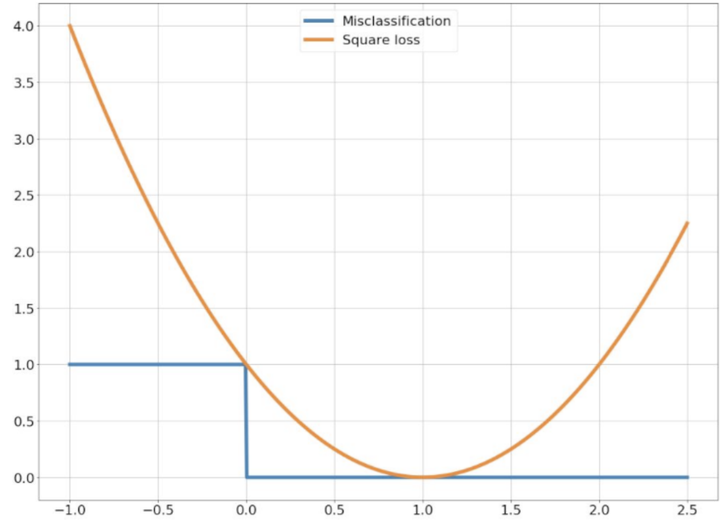
estimate it with a smooth function

Квадратичная функция ошибки

Let's treat classification problem as regression problem: $Y \in \{-1, 1\} \mapsto Y \in R$

thus we optimize MSE

$$L_{\text{MSE}} = (y_i - x_i^T w)^2 = \frac{(y_i^2 - y_i \cdot x_i^T w)^2}{y_i^2} =$$
$$= (1 - y_i \cdot x_i^T w)^2 = (1 - M_i)^2$$



Advantage: already solved

Disadvantage: penalizes for high confidence

01

Логистическая регрессия

Сигмоида

I. Let’s try to predict probability of an object to have positive class

$$p_+ = P(y = 1|x) \in [0, 1]$$

II. But all we can predict is a real number!

$$y = x^T w \in R$$

III. Time for some tricks

$$\frac{p_+}{1 - p_+} \in [0, +\infty)$$

$$\log \frac{p_+}{1 - p_+} \in R$$

Here is the match

IV. Reverse to closed form

$$\frac{p_+}{1 - p_+} = \exp(x^T w)$$

$$p_+ = \frac{1}{1 + exp(-x^T w)} = \sigma(x^T w)$$

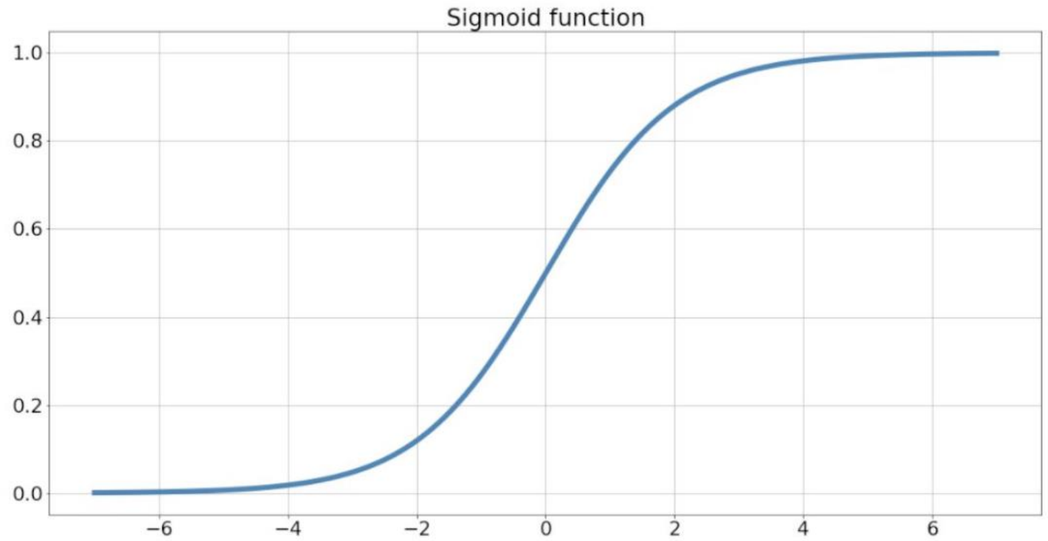
Сигмоида

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

Sigmoid is odd relative to (0, 0.5) point

Symmetric property:

$$1 - \sigma(x) = \sigma(-x)$$



Derivative: $\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$

Максимальное правдоподобие

Just to remind

$$\log L(w|X,Y) = \log P(X,Y|w) = \log \prod_{i=1}^{n} P(x_i, y_i|w)$$

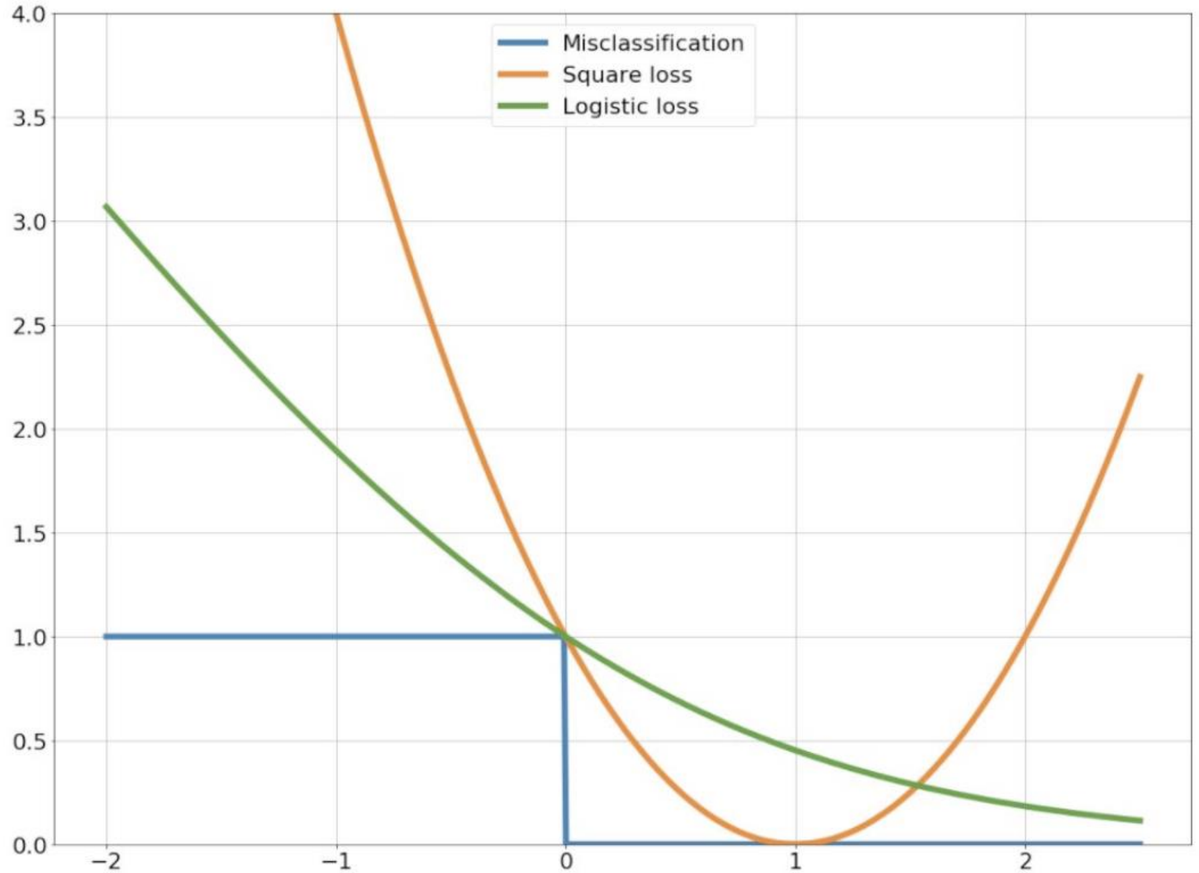
Calculating probabilities for objects

$$\text{if } y_i = 1: \quad P(x_i, 1|w) = \sigma_w(x_i) = \sigma_w(M_i)$$
$$\text{if } y_i = -1: \quad P(x_i, -1|w) = 1 - \sigma_w(x_i) = \sigma_w(-x_i) = \sigma_w(M_i)$$

$$\log L(w|X,Y) = \sum_{i=1}^{n} \log \sigma_w(M_i) = -\sum_{i=1}^{n} \log(1 + \exp(-M_i)) \to \min_w$$

Логистическая функция ошибки

$$L_{Logistic} = \log(1 + \exp(-M_i))$$



Логистическая регрессия

- Задача бинерной классификации
- Минимизация верхней оценки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log(1+\exp(-y_i\langle w,x_i\rangle))\to\min_w$$

Логистическая регрессия

- Задача бинерной классификации
- Минимизация верхней оценки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log(1+\exp(-y_i\langle w,x_i\rangle))\to\min_w$$

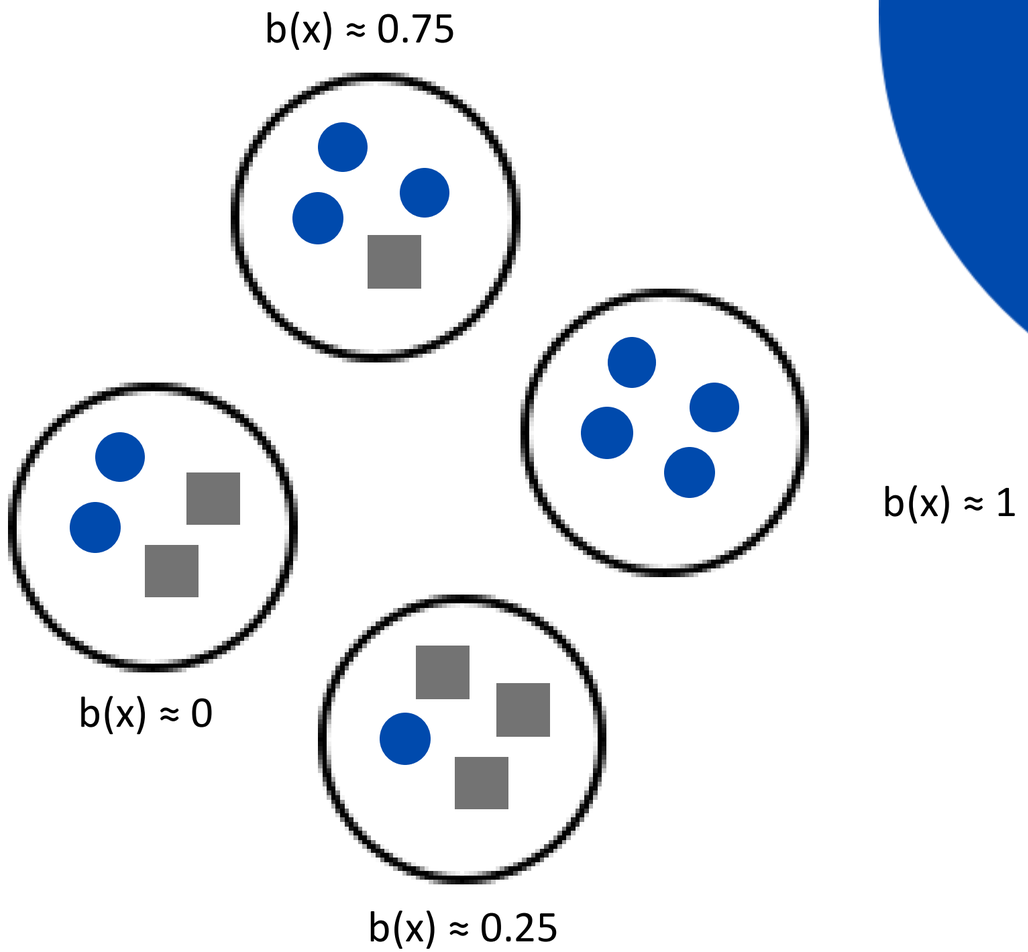
- Будем говорить, что модель b(x) предсказывает вероятности, если среди объектов с b(x) = p, доля положительных равна p

Логистическая регрессия

- Задача бинерной классификации
- Минимизация верхней оценки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log(1+\exp(-y_i\langle w,x_i\rangle))\to\min_w$$

- Будем говорить, что модель b(x) предсказывает вероятности, если среди объектов с b(x) = p, доля положительных равна p



Логистическая регрессия

- a(x) = sign ⟨w, x⟩
- обучим на логистическую функцию потерь:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log(1+\exp(-y_i\langle w, x_i\rangle)) \to \min_w$$

Логистическая регрессия

- a(x) = sign ⟨w, x⟩
- обучим на логистическую функцию потерь:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log(1+\exp(-y_i\langle w,x_i\rangle))\to\min_w$$

- Переведем выход модели на отрезок [0,1]
- Сигмоида:

$$\sigma(\langle w,x\rangle)=\frac{1}{1+\exp(-\langle w,x\rangle)}$$

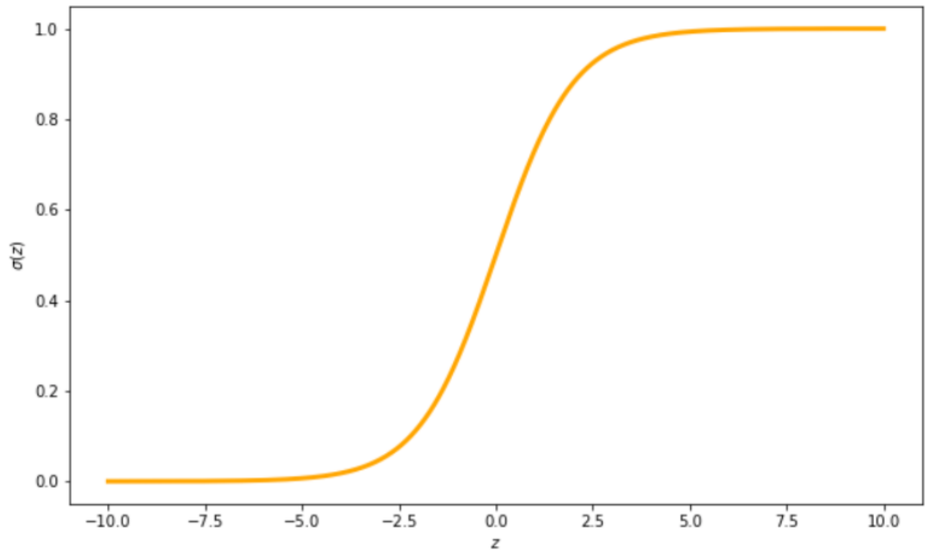
Логистическая регрессия

- a(x) = sign ⟨w, x⟩
- обучим на логистическую функцию потерь:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\log(1+\exp(-y_i\langle w,x_i\rangle)) \to \min_w$$

- Переведем выход модели на отрезок [0,1]
- Сигмоида:

$$\sigma(\langle w,x\rangle) = \frac{1}{1+\exp(-\langle w,x\rangle)}$$



Логистическая регрессия

- $b(x) = \sigma(\langle w, x \rangle)$
- Если $y_i = +1$, то $\sigma(\langle w, x \rangle) \to 1$, $\langle w, x \rangle \to +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x \rangle) \to 0$, $\langle w, x \rangle \to -\infty$

Логистическая регрессия

- $b(x) = \sigma(\langle w, x \rangle)$
- Если $y_i = +1$, то $\sigma(\langle w, x \rangle) \to 1$, $\langle w, x \rangle \to +\infty$
- Если $y_i = -1$, то $\sigma(\langle w, x \rangle) \to 0$, $\langle w, x \rangle \to -\infty$

- Задача - сделать отступы на всех объектах максимальными

Логистическая регрессия

- b(x) = σ(⟨w, x⟩)
- Если y_i = +1, то σ(⟨w, x⟩) -> 1, ⟨w, x⟩ -> +∞
- Если y_i = -1, то σ(⟨w, x⟩) -> 0, ⟨w, x⟩ -> -∞

- Задача - сделать отступы на всех объектах максимальными

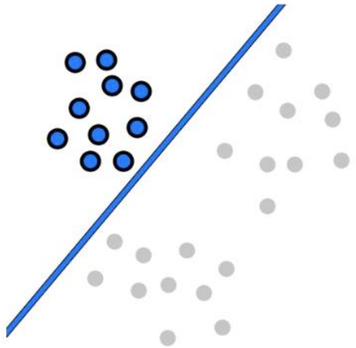
$$y_i\langle w, x_i\rangle \to \max_w$$

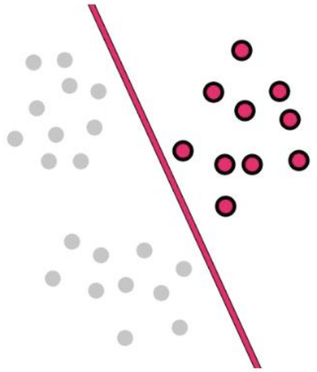
Логистическая регрессия

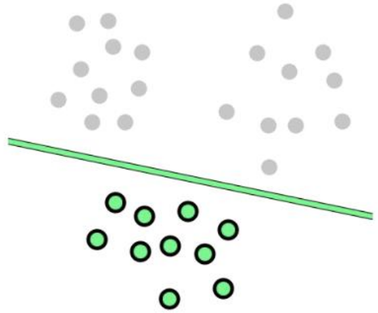
- b(x) = σ(⟨w, x⟩)
- Если $y_i = +1$, то σ(⟨w, x⟩) -> 1, ⟨w, x⟩ -> +∞
- Если $y_i = -1$, то σ(⟨w, x⟩) -> 0, ⟨w, x⟩ -> -∞

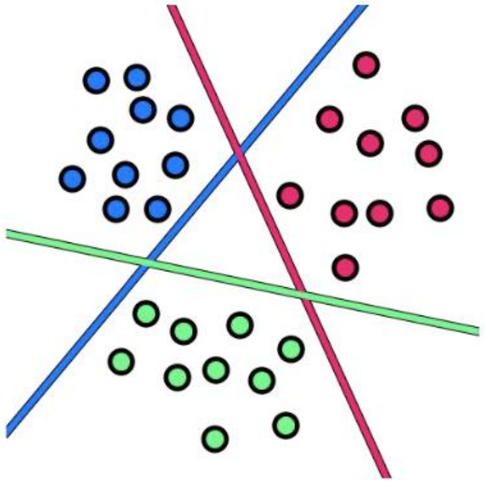
$$-\sum_{i=1}^{\ell}\{[y_i = 1]\sigma(\langle w, x_i\rangle) + [y_i = -1](1 - \sigma(\langle w, x_i\rangle))\} \to \min_w$$

Мультиклассовая задача One vs Rest

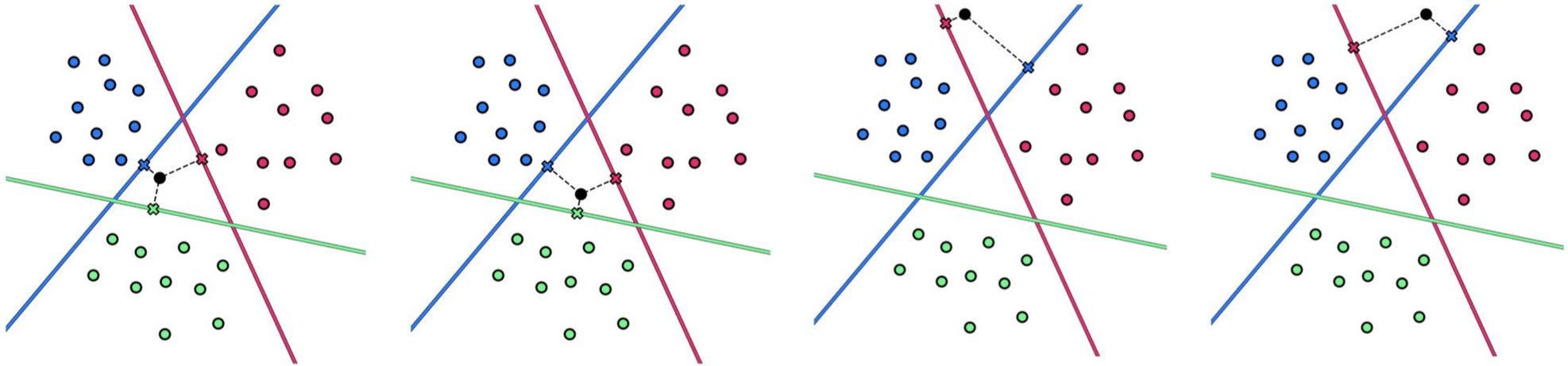




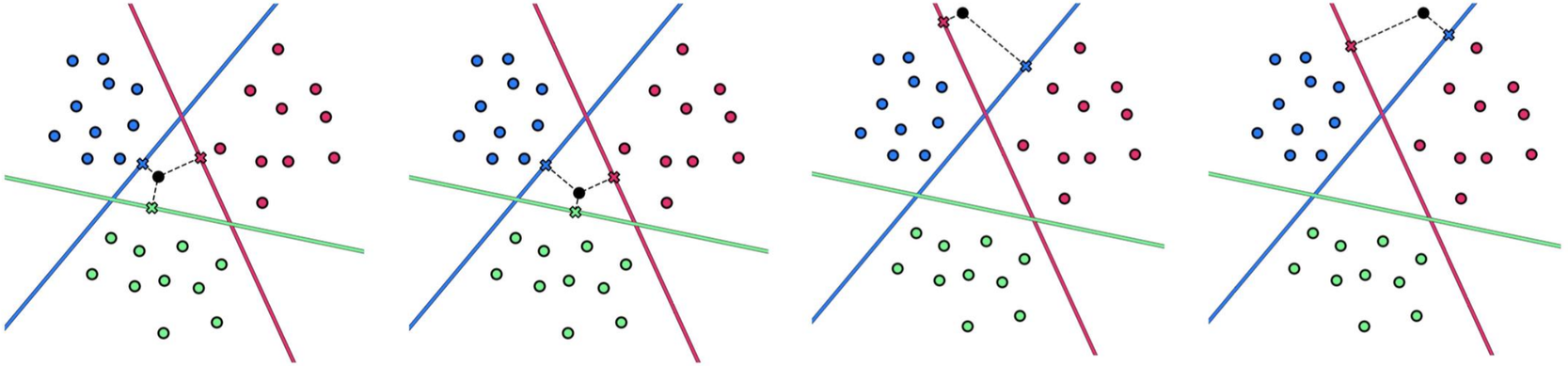


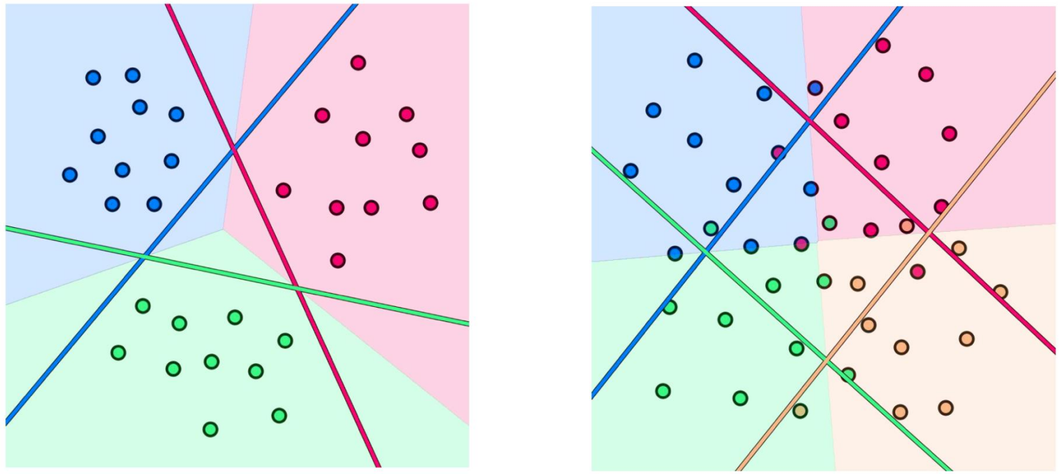


Мультиклассовая задача One vs Rest

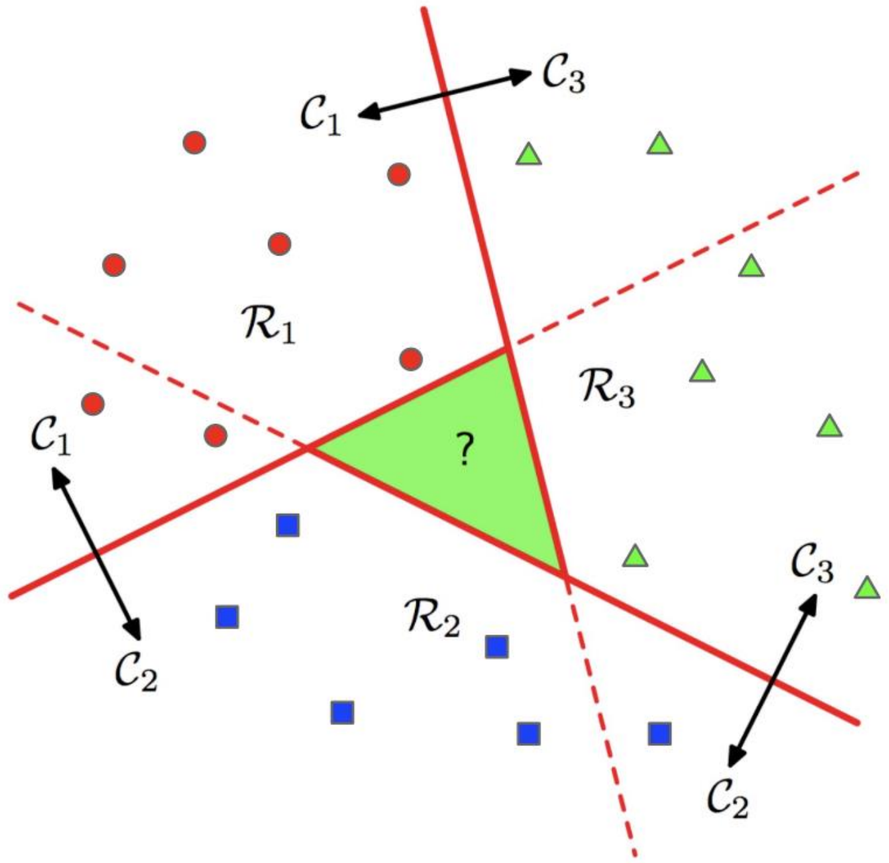


Мультиклассовая задача One vs Rest





Мультиклассовая задача One vs One



02

Метрики классификации

Accuracy

- Доля верных ответов или доля ошибок

$$Q(a,X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) = y_i] \qquad Q(a,X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) \neq y_i]$$

- Это нельзя называть точностью!

a(x)	y
-1	-1
+1	+1
+1	-1
+1	+1

- Доля ошибок: 0.2
- Доля верных ответов: 0.8

Accuracy

$$Q(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) \neq y_i]$$

Решаем задачу выявления редкого заболевания

- 950 здоровых (y = +1)
- 50 больных (y = -1)

Модель: a(x) = +1

Доля ошибок: 0.05

Accuracy

$$Q(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) \neq y_i]$$

Решаем задачу выявления редкого заболевания

- 950 здоровых (y = +1)
- 50 больных (y = -1)

Модель: a(x) = +1

Доля ошибок: 0.05

Вывод: баланс классов при использовании данной метрики очень важен

Несбалансированные выборки

- Несбалансированная выборка - объектов одного класса существенно больше
- Примеры: клики по рекламе, медицинская диагностика, отток клиентов и т.д.

- Accuracy не учитывает цены ошибок
- Пример: кредитный скоринг
- Модель 1:
  - 80 кредитов вернули
  - 20 кредитов не вернули
- Модель 2:
  - 48 кредитов вернули
  - 20 кредита не вернули
- Кто лучше?

Матрица ошибок

	y = 1	y = -1
a(x) = 1	True Positive (TP)	False Positive (FP)
a(x) = -1	False Negative (FN)	True Negative (TN)

Матрица ошибок

<True/False> <Positive/Negative>

Совпадение с ответом

Что говорит модель

НЕ было, но мы нашли

	y = 1	y = -1
a(x) = 1	True Positive (TP)	False Positive (FP)
a(x) = -1	False Negative (FN)	True Negative (TN)

Было, но мы НЕ нашли

Матрица ошибок

Модель 1

	y = 1	y = -1
a(x) = 1	80	20
a(x) = -1	20	80

Модель 2

	y = 1	y = -1
a(x) = 1	48	2
a(x) = -1	20	48

Матрица ошибок

Точность (precision)

- Можно ли доверять классификатору при a(x) = 1?

$$\text{precision}(a, X) = \frac{TP}{TP + FP}$$

Матрица ошибок

Точность (precision)

Модель 1

	y = 1	y = -1
a(x) = 1	80	20
a(x) = -1	20	80

precision = ?

Модель 2

	y = 1	y = -1
a(x) = 1	48	2
a(x) = -1	20	48

precision = ?

Матрица ошибок

Точность (precision)

Модель 1

	y = 1	y = -1
a(x) = 1	80	20
a(x) = -1	20	80

precision = 0.8

Модель 2

	y = 1	y = -1
a(x) = 1	48	2
a(x) = -1	20	48

precision = 0.96

Матрица ошибок

Полнота (recall)

- Как много положительных объектов находит классификатор?

$$\text{recall}(a, X) = \frac{TP}{TP + FN}$$

Матрица ошибок

Полнота (recall)

Модель 1

	y = 1	y = -1
a(x) = 1	80	20
a(x) = -1	20	80

recall = ?

Модель 2

	y = 1	y = -1
a(x) = 1	48	2
a(x) = -1	20	48

recall = ?

Матрица ошибок

Полнота (recall)

Модель 1

	y = 1	y = -1
a(x) = 1	80	20
a(x) = -1	20	80

recall = 0.8

Модель 2

	y = 1	y = -1
a(x) = 1	48	2
a(x) = -1	20	48

recall = 0.71

Точность и полнота

Что важнее?

- Антифрод, классифицируем транзакции,
  Высокая точность, низкая полнота: редко блокируем нормальные, пропускаем много мошеннических
  Низкая точность, высокая полнота: часто блокируем нормальные, редко пропускаем мошеннические, что лучше?
- Медицинская диагностика: надо найти не менее 80% больных, ограничение: recall > 0.8, что максимизируем?

03
Совмещение точности и полноты

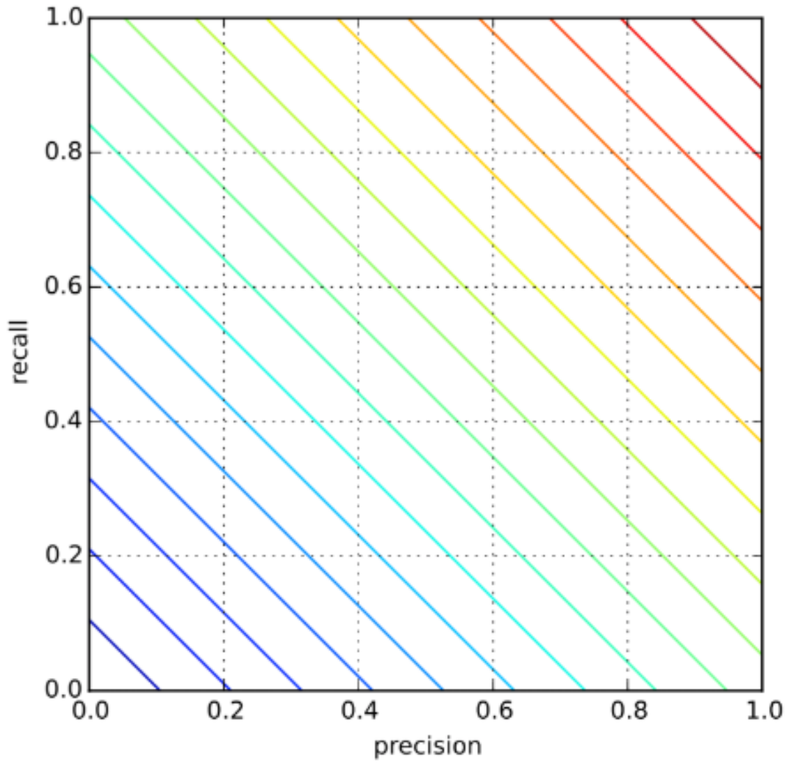
Точность и полнота

Две метрики важны, однако, как можно их совместить?

Точность и полнота

Арифметическое среднее

$$A = \frac{1}{2}(\text{precision} + \text{recall})$$

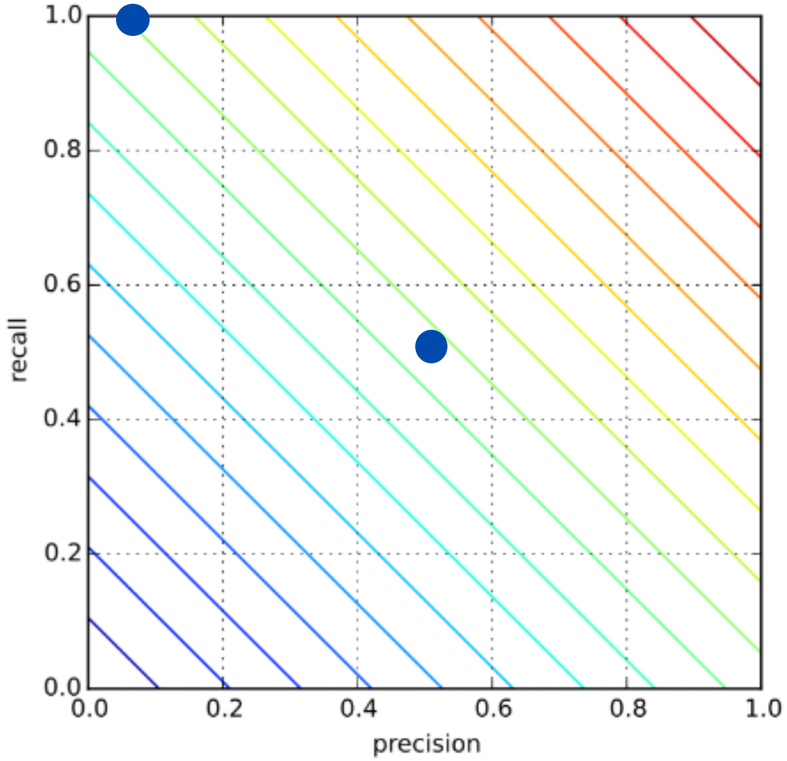


Точность и полнота

Арифметическое среднее

A = 0.5 * (precision + recall)

- pr = 0.1, rc = 1, A = 0.55, плохой алгоритм
- pr = 0.55, rc = 0.55, A = 0.55, алгоритм лучше, но качество такое же

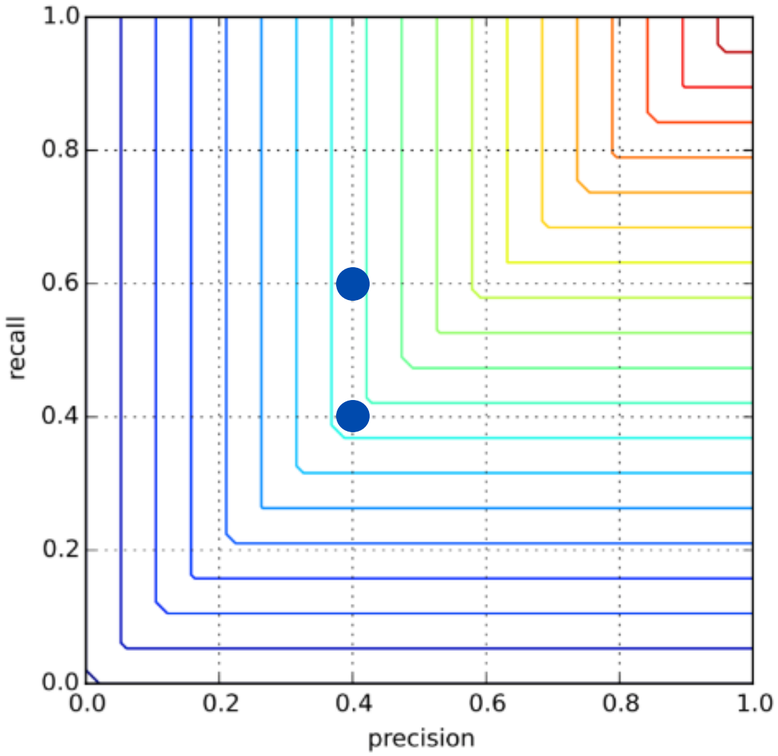


Точность и полнота

Минимум

M = min(precision, recall)

- pr = 0.4, rc = 0.4, M = 0.4, неплохо
- pr = 0.4, rc = 0.6, A = 0.4 алгоритм лучше, но качество такое же

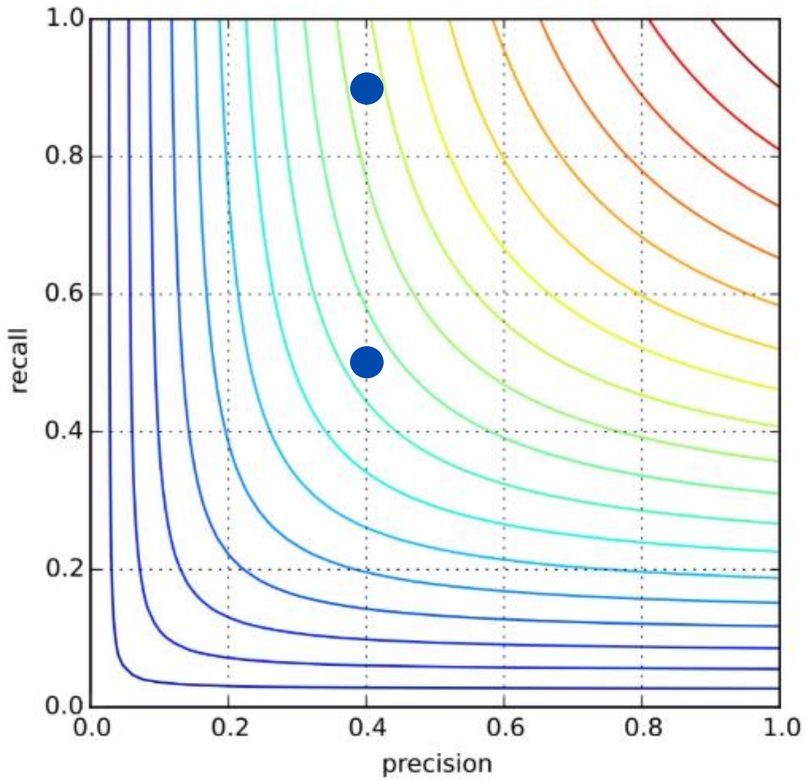


Точность и полнота

F-мера

$$F = \frac{2 * precision * recall}{precision + recall}$$

- pr = 0.4, rc = 0.5, F = 0.44
- pr = 0.4, rc = 0.9, F = 0.55

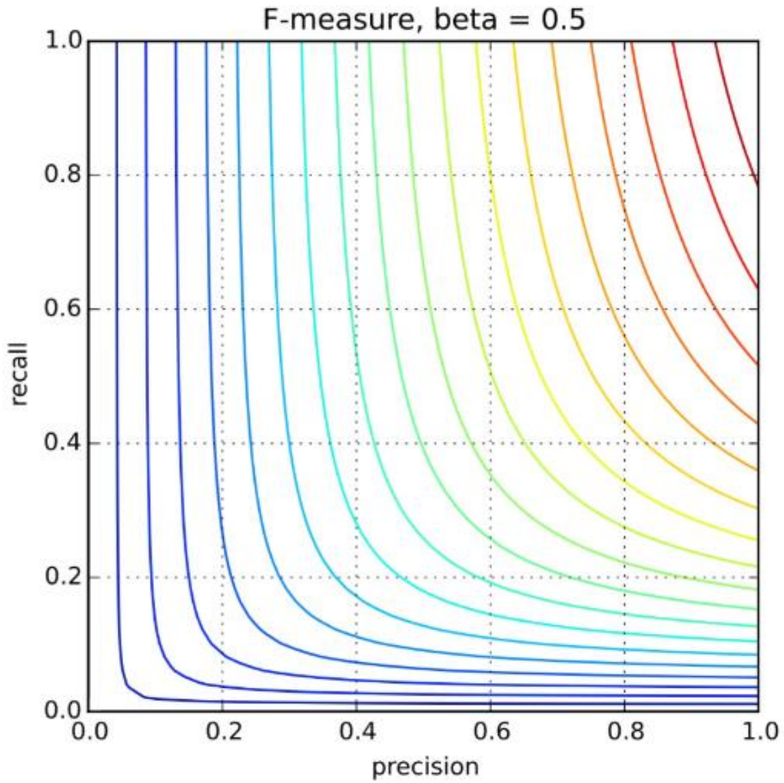


Точность и полнота

F-мера

$$F_\beta = (1 + \beta)^2 \frac{2 * \text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$$

- β = 0.5 - важнее точность



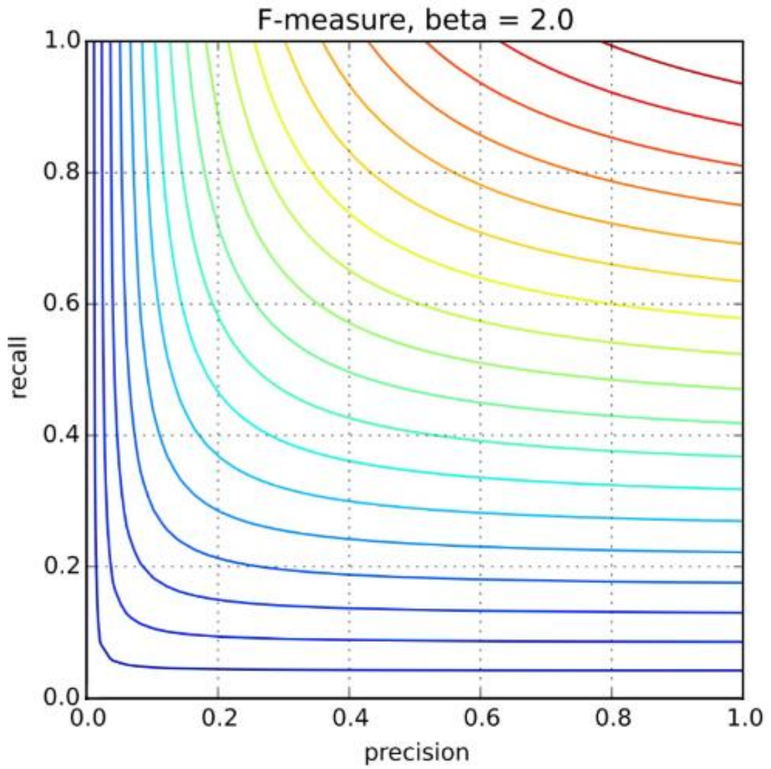
- β = 2 - важнее полнота

Точность и полнота

F-мера

$$F_\beta = (1 + \beta)^2 \frac{2 * \text{precision} * \text{recall}}{\beta^2 * \text{precision} + \text{recall}}$$

- β = 2 - важнее полнота



04

PR-кривая

PR-кривая

- Линейный классификатор:
  $a(x) = \text{sign}(\langle w, x\rangle - t) = 2[\langle w, x\rangle > t] - 1$

- $\langle w, x\rangle$ - оценка принадлежности классу +1, часто $t = 0$

- Высокий порог:
  - Мало объектов относим к +1
  - Точность выше
  - Полнота ниже
- Низкий порог:
  - Много объектов относим к +1
  - Точность ниже
  - Полнота выше

PR-кривая

- Линейный классификатор:
  $a(x) = \text{sign}(\langle w, x\rangle - t) = 2[\langle w, x\rangle > t] - 1$

- Оценка принадлежности

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

PR-кривая

- Линейный классификатор:
  $a(x) = \text{sign}(\langle w, x\rangle - t) = 2[\langle w, x\rangle > t] - 1$

- Оценка принадлежности

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

PR-кривая

- Линейный классификатор:
  $a(x) = \text{sign}(\langle w, x\rangle - t) = 2[\langle w, x\rangle > t] - 1$

- Оценка принадлежности

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

PR-кривая

- Линейный классификатор:
  $a(x) = \text{sign}(\langle w, x\rangle - t) = 2[\langle w, x\rangle > t] - 1$

- Оценка принадлежности

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

PR-кривая

- Как оценить качество b(x)?
- Порог выбирается позже
- Порог зависит от ограничения на точность или полноту

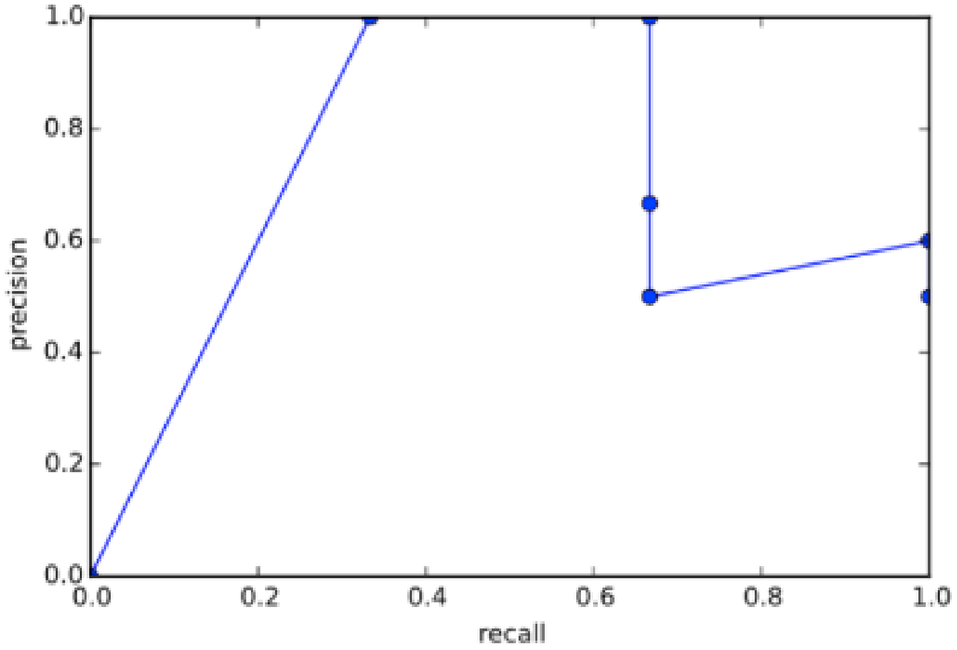
-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

PR-кривая

- Кредитный скоринг
- b(x) - оценка вероятности возврата кредита
- a(x) = [b(x) > 0.5]
- pr = 0.1, rc = 0.7
- дело в алгоритме или пороге?

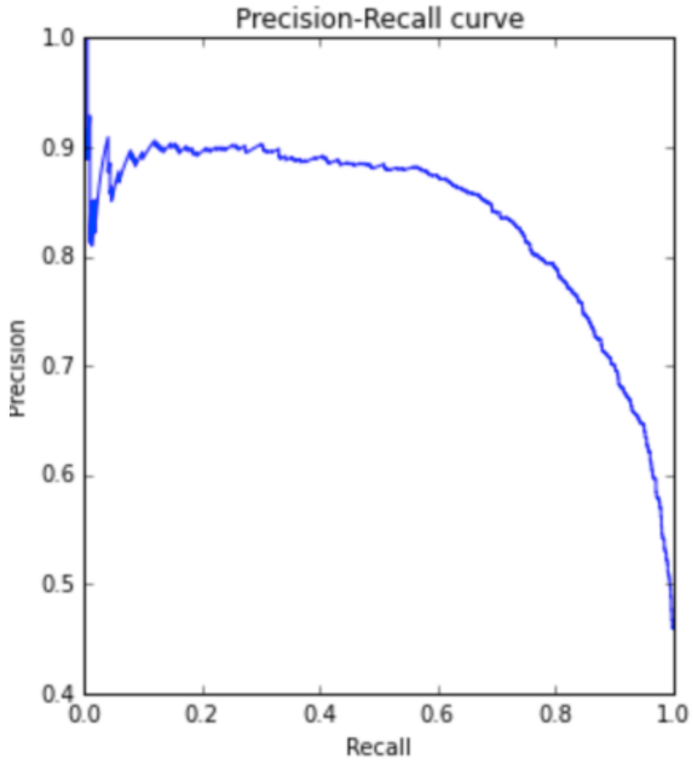
PR-кривая

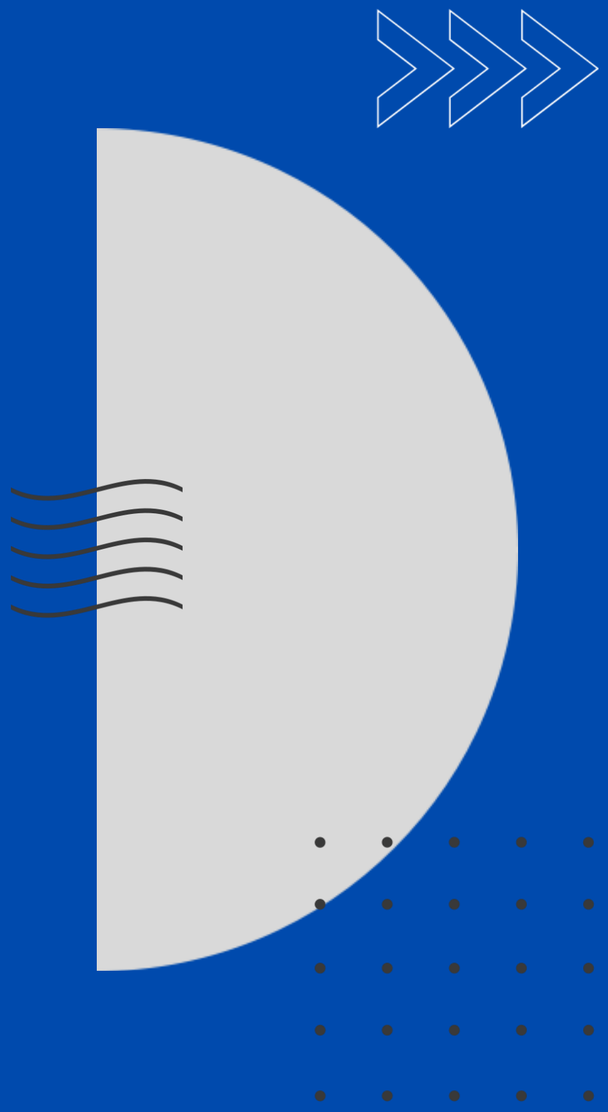
- Кривая точности-полноты
- Ось X - полнота
- Ось Y - точность
- Точки - значения при последовательных порогах



PR-кривая

- Левая точка: (0, 1)
- Правая точка: (1, r), r - доля положительных объектов
- Для идеального классификатора проходит через (1,1)
- AUC-PRC - площадь под PR-кривой





Место для ваших вопросов