Постановка задачи

Дано: задача классификации.

$X^\ell = \{x_1, \ldots, x_\ell\}$ — выборка;

$y_i = y(x_i) \in \{0, 1\}, \;\; i = 1, \ldots, \ell$ — известные бинарные ответы.

$a\colon X \to Y$ — алгоритм, решающая функция, приближающая y на всём множестве объектов X.

Вопрос:
Как измерить качество $a(x)$ на выборке X^ℓ?

Accuracy

Доля правильных ответов на выборке (accuracy):

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

- Соответствует интуитивным представлениям о качестве классификации
- Имеет проблемы с интерпретацией на несбалансированных выборках.

Accuracy

Пример (медицинская диагностика):

- 950 объектов класса 0,
- 50 объектов класса 1,
- $a(x) = 0$ для всех x.

Доля правильных ответов $a(x)$: 95%!

Решение: смотреть на базовую долю правильных ответов

$$\text{BaseRate} = \operatorname*{arg\,max}_{y_0 \in \{0,1\}} \frac{1}{\ell} \sum_{i=1}^{\ell} [y_0 = y_i]$$

В примере: $\text{BaseRate} = 95\%$.

Матрица ошибок

Ошибки бывают разные:

	$y = 1$	$y = 0$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = 0$	False Negative (FN)	True Negative (TN)

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}.$$

Матрица ошибок

Пример: задача медицинской диагностики ($y = 1$ — больные, $y = 0$ — здоровые).

	$y = 1$	$y = 0$
$a(x) = 1$	20	50
$a(x) = 0$	5	1000

Доля правильных ответов: 94.9%

	$y = 1$	$y = 0$
$a(x) = 1$	0	0
$a(x) = 0$	25	1050

Доля правильных ответов константного классификатора: 97.6%

У разных типов ошибки может быть разная *цена*.

Точность и полнота

Точность (precision) — насколько можно доверять классификатору:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

	$y = 1$	$y = 0$
$a(x) = 1$	20	50
$a(x) = 0$	5	1000

Точность классификатора: 28.6%

Точность константного классификатора: 0%

Точность и полнота

Полнота (recall) — как много объектов класса 1 находит классификатор:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

	$y = 1$	$y = 0$
$a(x) = 1$	20	50
$a(x) = 0$	5	1000

Полнота классификатора: 80%
Полнота константного классификатора: 0%

Точность и полнота

- Точность и полнота характеризуют разные стороны качества классификатора
- Чем выше точность, тем меньше ложных срабатываний
- Чем выше полнота, тем меньше ложных пропусков
- Приоритет в сторону точности или полноты выбирается в зависимости от задачи

Точность и полнота

Пример 1: определение мошеннических действий на банковских счетах.
Важнее **полнота**: лучше проверить лишний раз, чем пропустить вредоносные действия.

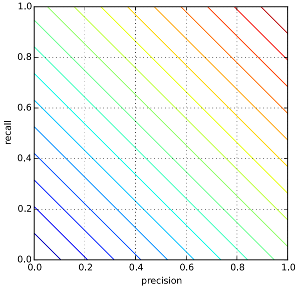
Пример 2: поиск вражеских самолетов для автоматического уничтожения ракетой
Важнее **точность**: нельзя допустить стрельбы по своему самолету.

Усреднение точности и полноты

Арифметическое среднее:

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

- Если precision $= 0.05$, recall $= 1$, то $A = 0.525$.
- Если precision $= 0.525$, recall $= 0.525$, то $A = 0.525$.
- Первый классификатор — константный, не имеет смысла.
- Второй классификатор показывает неплохое качество.



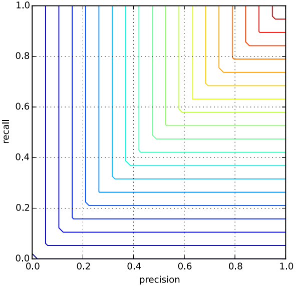
Усреднение точности и полноты

Минимум:

$$M = \min(\text{precision}, \text{recall})$$

- Если precision $= 0.05$, recall $= 1$, то $M = 0.05$.
- Если precision $= 0.525$, recall $= 0.525$, то $M = 0.525$.

- Если precision $= 0.2$, recall $= 1$, то $M = 0.2$.
- Если precision $= 0.2$, recall $= 0.3$, то $M = 0.2$.



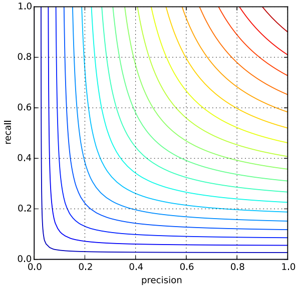
Усреднение точности и полноты

Гармоническое среднее, или F-мера:

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

- Если precision $= 0.05$, recall $= 1$, то $F = 0.1$.
- Если precision $= 0.525$, recall $= 0.525$, то $F = 0.525$.

- Если precision $= 0.2$, recall $= 1$, то $F = 0.33$.
- Если precision $= 0.2$, recall $= 0.3$, то $F = 0.24$.



Заключение

- Простая мера качества классификации — доля верных ответов
- Не учитывает цены ошибок
- Точность и полнота позволяют различать ложные срабатывания и ложные пропуски
- F-мера — способ усреднения точности и полноты

Постановка задачи

Дано: задача классификации.

$X^\ell = \{x_1, \ldots, x_\ell\}$ — выборка;

$y_i = y(x_i) \in \{0, 1\}, \;\; i = 1, \ldots, \ell$ — известные бинарные ответы.

$b\colon X \to \mathbb{R}$ — алгоритм, оценивающий принадлежность x к классу 1.

Вопрос:
Как измерить качество $b(x)$ на выборке X^ℓ?

Откуда берутся оценки принадлежности?

Как правило, классификатор имеет вид

$$a(x) = [b(x) > t].$$

- $b(x)$ — оценка принадлежности к классу 1
- t — порог классификации

Откуда берутся оценки принадлежности?

Линейный классификатор:

$$a(x) = [\langle w, x \rangle > 0].$$

[здесь идет картинка с разделяющей прямой; скалярное произведение оценивает расстояние; если объект близко к ней, то мы не уверены, если далеко, то уверены в ответы]

Откуда берутся оценки принадлежности?

Метод k ближайших соседей:

$$a(x) = \left[\sum_{i=1}^{k} [y^{(i)} = 1] > k/2 \right].$$

[здесь идет визуализация тоже: если среди соседей объекта все относятся к одному классу, то он уверен в классификации и выдает высокую оценку; если же среди соседей встречаются оба класса, то оценка понижается]

Зачем нужны оценки принадлежности?

Пример: кредитный скоринг.

- нужно предсказать, вернет ли клиент кредит;
- сортируем клиентов по оценке вероятности возврата $b(x)$;
- банк получает ранжированный список;
- порог выбирается в зависимости от стратегии банка;
- порог может многократно пересматриваться.

Зачем нужны оценки принадлежности?

Пример: задача определения самолетов противника.

- $b(x)$ оценивает вероятность того, что самолет принадлежит противнику;
- классификатор $a(x) = [b(x) > 0.9]$;
- precision $= 0.2$, recall $= 0.7$;
- как понять, в чем проблема — неправильном пороге или плохой функции $b(x)$?

PR-кривая

1. Отсортируем объекты по возрастанию оценки $b(x)$:

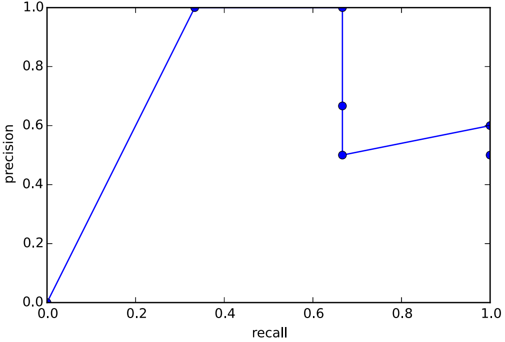
$$b\left(x_{(1)}\right) \leqslant \ldots \leqslant b\left(x_{(\ell)}\right).$$

2. Переберем все пороги классификации, начав с максимального:

$$\begin{aligned} t_\ell &= b\left(x_{(\ell)}\right), \\ &\ldots \\ t_1 &= b\left(x_{(1)}\right), \\ t_0 &= b\left(x_{(1)}\right) - \varepsilon. \end{aligned}$$

3. Для каждого порога посчитаем точность и полноту.
4. Нанесем соответствующую точку в осях «полнота-точность».
5. Соединим точки, получив Precision-Recall-кривую.

PR-кривая



$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

PR-кривая

Свойства:

- Левая точка: всегда $(0, 0)$ (все объекты относим к классу 0);
- Правая точка: $(1, \ell_+/\ell)$, ℓ_+ — число объектов класса 1 в выборке;
- Если выборка идеально разделима, то кривая пройдет через точку $(1, 1)$;
- Чем больше площадь под кривой, тем лучше.

AUC-PRC (Area Under Precision-Recall curve) — мера качества для $b(x)$.

ROC-кривая

ROC — «reciever operating characteristic».

- по оси X: False Positive Rate, доля ошибочных положительных классификаций:

  $$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

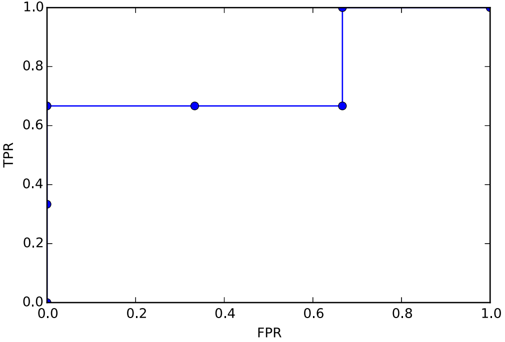
  $1 - \mathrm{FPR}$ называется *специфичностью* алгоритма.

- по оси Y: True Positive Rate, доля правильных положительных классификаций:

  $$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

  TPR называется *чувствительностью* алгоритма.

ROC-кривая



$b(x)$	0.14	0.23	0.39	0.52	0.73	0.90
y	0	1	0	0	1	1

Свойства:

- Левая точка: всегда $(0, 0)$ (все объекты относим к классу 0);
- Правая точка: всегда $(1, 1)$ (все объекты относим к классу 1);
- Если выборка идеально разделима, то кривая пройдет через точку $(1, 0)$;
- Площадь меняется от $1/2$ до 1;
- Чем больше площадь под кривой, тем лучше.

AUC-ROC (Area Under ROC-curve) — мера качества для $b(x)$.

ROC-кривая:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \qquad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

- Метрики качества нормируются на размеры классов, ROC-кривая не изменится при перемене соотношения классов.
- Интерпретация: AUC-ROC равен вероятности того, что случайно взятый объект класса 1 получит оценку выше, чем случайно взятый объект класса 0.
- Имеет проблемы при сильном дисбалансе классов.

Сравнение

PR-кривая:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- Точность нормируется на число положительных прогнозов, изменится при перемене соотношения классов.
- Максимально возможная площадь под PR-кривой зависит от соотношения классов.
- Хорошо подходит для измерения качества при сильном дисбалансе классов.

Сравнение

- 100 объектов класса 1;
- 1.000.000 объектов класса 0;
- Ранжирование: 50.000 объектов класса 0,
  затем 100 объектов класса 1,
  затем все остальные объекты класса 0;

Метрики качества:

- AUC-ROC: 0.95;
- AUC-PRC: 0.001.

Почему так получается?

- Выберем порог, при котором первые 50.095 объектов относятся к классу 1;
- $\text{TPR} = 0.95$, $\text{FPR} = 0.05$;
- $\text{precision} = 0.0019$, $\text{recall} = 0.95$.

Заключение

- Работать с оценками принадлежности может быть полезнее, чем с бинарными ответами
- Две основные метрики качества: AUC-PRC и AUC-ROC
- AUC-ROC не зависит от соотношения классов

Дано: $\{x_1, \dots, x_\ell\} \subset X$ — *выборка*;
$y_i = y(x_i) \in \{1, \dots, K\}$, $i = 1, \dots, \ell$ — известные ответы.

Найти: $a\colon X \to Y$ — алгоритм, решающую функцию, приближающую y на всём множестве объектов X.

Вопросы:

1. Как свести задачу к бинарной классификации?
2. Как измерить качество решения?

One-vs-all

Идея: построить K классификаторов, отделяющих каждый класс от остальных.

Получим K задач бинарной классификации:

- Объекты: $X^k = X^\ell$;
- Ответы: $y_i^k = [y_i = k]$;
- Оценка принадлежности: $b_k(x) \in \mathbb{R}$.

Итоговый алгоритм:

$$a(x) = \arg\max_{k=1,\dots,K} b_k(x).$$

All-vs-all

Идея: построить классификаторы для каждой пары классов.

Получим $K(K-1)$ задач бинарной классификации:

- Объекты: $X^{km} = \{x \in X^\ell \,|\, y(x) = k \text{ или } y(x) = m\}$;
- Ответы: $y_i^{km} = [y_i = k]$;
- Оценка принадлежности: $b_{km}(x) \in \mathbb{R}$;
- Симметрия: $b_{km}(x) = -b_{mk}(x)$.

Итоговый алгоритм:

$$a(x) = \operatorname*{arg\,max}_{k=1,\dots,K} \sum_{m=1}^{K} b_{km}(x).$$

Сравнение

One-vs-all:

- Линейное число классификаторов, но каждый обучается на полной выборке.
- Может возникнуть проблема с несбалансированными выборками.

All-vs-all:

- Квадратичное число классификаторов, но каждый обучается на небольшой подвыборке.

Метрики качества

Доля правильных ответов (accuracy):

$$\frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) = y_i].$$

Матрица ошибок:

	$y = 1$	$\dots$	$y = K$
$a(x) = 1$	q_{11}	$\dots$	q_{1K}
$\dots$	$\dots$	$\dots$	$\dots$
$a(x) = K$	q_{K1}	$\dots$	q_{KK}

где

$$q_{ij} = \sum_{m=1}^{\ell}[a(x_m) = i][y_m = j].$$

Метрики качества

Как обобщить точность, полноту, AUC?

Рассмотрим K задач отделения одного из классов от остальных.

- Микро-усреднение (micro-averaging):
  - Найдем TP, FP, FN, TN для каждой из задач;
  - Усредним их по всем задачам;
  - Вычислим итоговую метрику.

  Вклад каждого класса зависит от его размера.
- Макро-усреднение (macro-averaging):
  - Вычислим итоговую метрику для каждой из задач;
  - Усредним по всем классам.

  Все классы вносят равный вклад.

Метрики качества

	TP	FP	FN	TN
$y = 1$	900	120	100	930
$y = 2$	850	70	150	980
$y = 3$	10	100	40	1900

Чему равна точность (precision)?

Микро-усреднение:

$\overline{\text{TP}}$	$\overline{\text{FP}}$	$\overline{\text{FN}}$	$\overline{\text{TN}}$
586.7	96.7	96.7	1270

Точность: 86%

Макро-усреднение:

Класс 1	Класс 2	Класс 3
88%	92%	9%

Точность: 63%

Заключение

- Многоклассовую классификацию можно свести к серии бинарных задач
- Два подхода: one-vs-all и all-vs-all
- Вычисление качества также производится через сведение к бинарным задачам
- Микро-усреднение учитывает наиболее крупные классы
- Макро-усреднение учитывает все классы одинаково, без учета их размеров