

Лекция 7: оценка результатов классификатора

Евгений Борисов

оценка результатов классификатора

разделение данных на части (классы)

Учебный набор: [объект, ответ]

Задача: классификатор
объект → вектор-признак → результат

Обучение: минимизация ошибки
ошибка = результат - правильный ответ

Критерий остановки:
достигнут порог значения ошибки,
и/или порог количества циклов

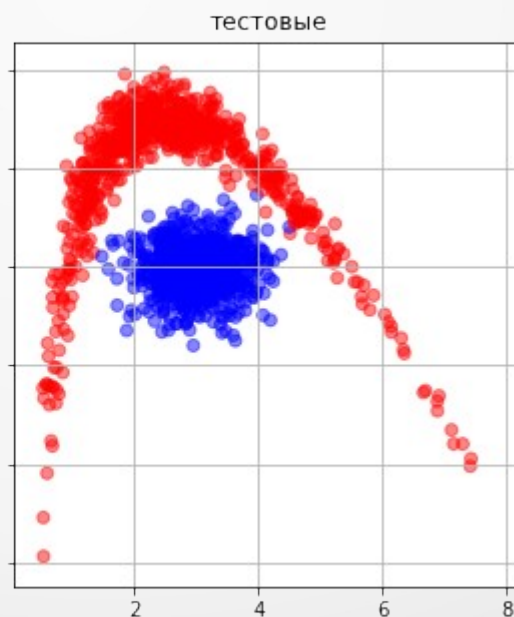
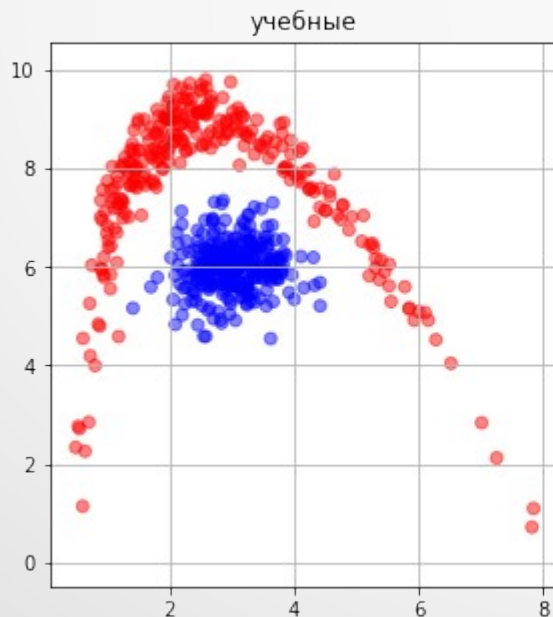
оценка результатов классификатора

разделяем набор данных

- учебный
- тестовый

недообучение (underfitting)
большая ошибка на учебном наборе

переобучение (overfitting)
малая ошибка на учебном наборе
большая ошибка на тестовом наборе



оценка результатов классификатора

классификация - задача разделения объектов на классы

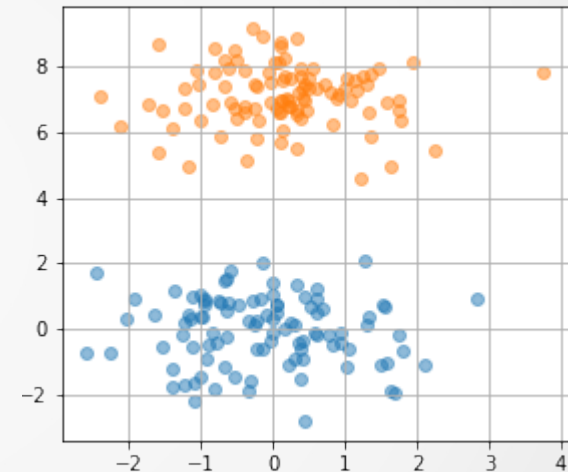
$X \subset \mathbb{R}^n$ - объекты

$Y \in \{0, 1\}$ - метки классов

$p \in [0, 1]$ - оценка

$a: X \rightarrow p$ - считаем оценку

$y = \begin{cases} 0, & p < b \\ 1, & p \geq b \end{cases}$ - если оценка выше порога
то объект принадлежит «первому» классу



оценка результатов классификатора

метрики качества на тестовом наборе

- погрешность (accuracy)
- матрица ошибок (confusion matrix)
- точность (precision)
- полнота (recall)
- F-мера
- ROC/AUC

оценка результатов классификатора

погрешность (accuracy)

правильные ответы / всего примеров

оценка для сбалансированного набора,
т.е. количество примеров в классах почти одинаковое

оценка результатов классификатора

Пример:

*тест на болезнь «зеленуху» имеет вероятность ошибки 0.1
(как позитивной, так и негативной)*

зеленухой болеет 10% населения.

*Какая вероятность того,
что человек болен зеленухой,
если у него позитивный результат теста?*

оценка результатов классификатора

Подсказка: формула Байеса

$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

оценка результатов классификатора

Подсказка: формула Байеса

$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

Ответ: $(0.9*0.1) / (0.9*0.1 + 0.1*0.9) = 0.5$

оценка результатов классификатора

Подсказка: формула Байеса

$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

Ответ: $(0.9*0.1) / (0.9*0.1 + 0.1*0.9) = 0.5$

Замечание о корректности этого результата:

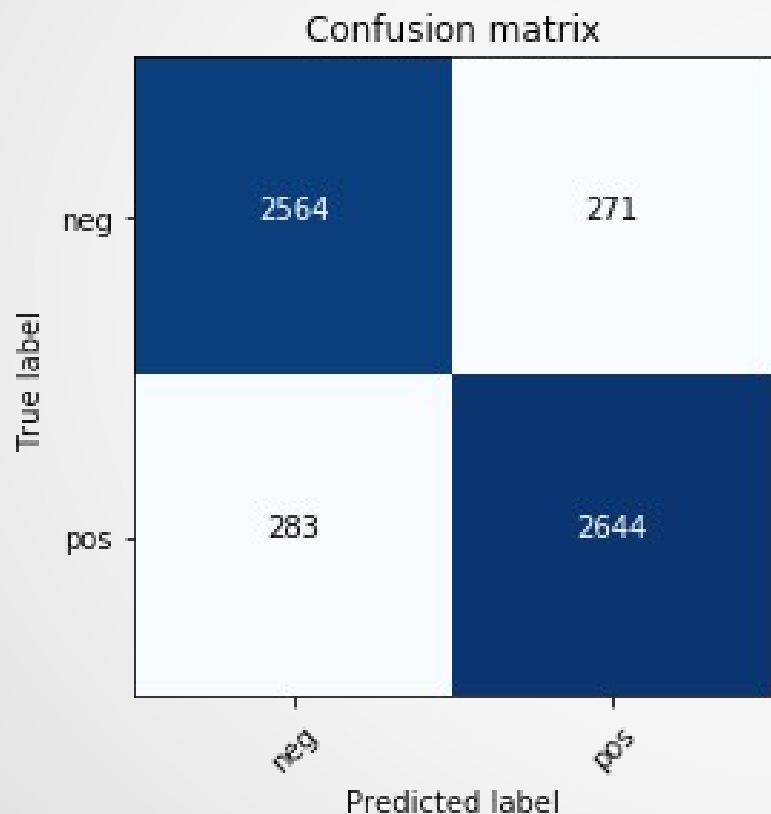
анализы проводят, когда есть подозрение на какую-то болезнь,

поэтому вероятность болезни надо вычислять по этой «подозрительной» группе.

это понижает требования к точности прибора...

оценка результатов классификатора

матрица ошибок (confusion matrix)



два класса — четыре группы

- TP истинно положительные
- TN истинно отрицательные
- FP ложно положительные
- FN ложно отрицательные

оценка результатов классификатора

точность (precision)

$$TP / (TP + FP)$$

(метрики для отдельного класса)

доля объектов действительно принадлежащих данному классу относительно всех объектов, которые классификатор отнес к этому классу

полнота (recall)

$$TP / (TP + FN)$$

доля объектов, найденных классификатором, относительно всех объектов этого класса

F-мера

$$(precision * recall) / (precision + recall)$$

усреднение точности и полноты

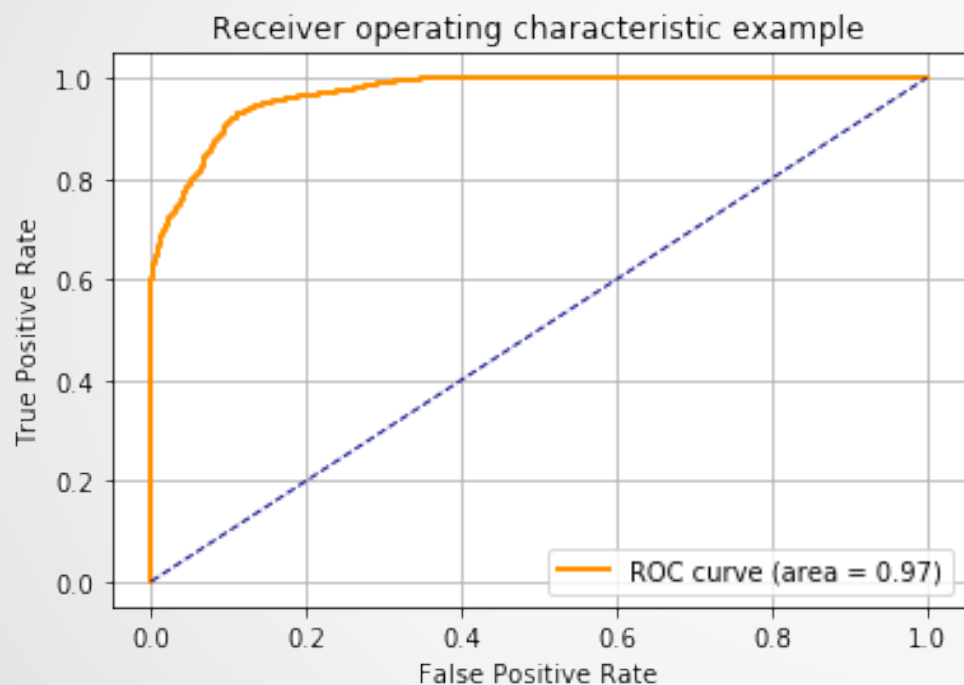
оценка результатов классификатора

Пример *classification_report*

	precision	recall	f1-score	support
0	0.90	0.90	0.90	2835
1	0.91	0.90	0.91	2927
avg / total	0.90	0.90	0.90	5762

оценка результатов классификатора

*ROC - receiver operating characteristic,
рабочая характеристика приёмника*



*AUC - area under ROC curve,
площадь под ROC-кривой
характеристика качества
классификации*

$$TPR = TP / (TP + FN)$$

полнота (recall), доля
объектов, найденных
классификатором,
относительно всех объектов
этого класса

$$FPR = FP / (FP + TN)$$

доля объектов negative класса
алгоритм предсказал неверно

ROC - зависимость полноты **TPR**
от доли ложно-негативных **FPR**
при изменении порога сора

оценка результатов классификатора

git clone https://github.com/mechanoid5/ml_lectorium.git

Александр Дьяконов AUC ROC (площадь под кривой ошибок)

Кривая ошибок <http://www.machinelearning.ru>

оценка результатов классификатора



Вопросы ?

оценка результатов классификатора

источники данных для экспериментов



sklearn.datasets

UCI Repository

Kaggle



- выбрать датасеты из UCI Repository
- обработать их с помощью Pandas
- применить kNN и DTree на этих наборах данных
- оценить результаты классификации
- <https://github.com/rougier/numpy-100>
- <https://github.com/ajcr/100-pandas-puzzles>