Гробов А.В.

Кафедра физики элементарных частиц

МИФИ

Машинное обучение

Лекция 4

- Метрики качества
- Деревья решений
- Практика

Метрики качества

- Отличное описание матрицы ошибок можно найти здесь - https://en.wikipedia.org/wiki/Sensitivity_and_specificity
- Метрики качества служат для контроля за процессом и результатом обучения алгоритма.
- Для задач классификации и регрессии используются разные метрики.

Классификация

- В задаче классификации в качестве меры качества удобно использовать долю верных\неверных ответов – accuracy.

$$accuracy = \frac{1}{l} \cdot \sum_{i=1}^{l} [\widehat{y_i} == y_i]$$

- Проблема возникает в случае несбалансированных наборов данных.

Пример: 50 объектов класса 1 и 50 объектов класса 0, accuracy =0.9 будет означать, что 90 объектов правильно классифицированы.

10 объектов класса 1 и 90 объектов класса 0 – accuracy = 0.9 также означает, что 90 объектов правильно классифицированы, но это могут быть 1 объект класса 1, и 89 объектов класса 0. Модель ошибается на большинстве объектов класса 1.

А как быть в ситуации когда у нас 10 объектов класса 1 и 1000 объектов класса 0?

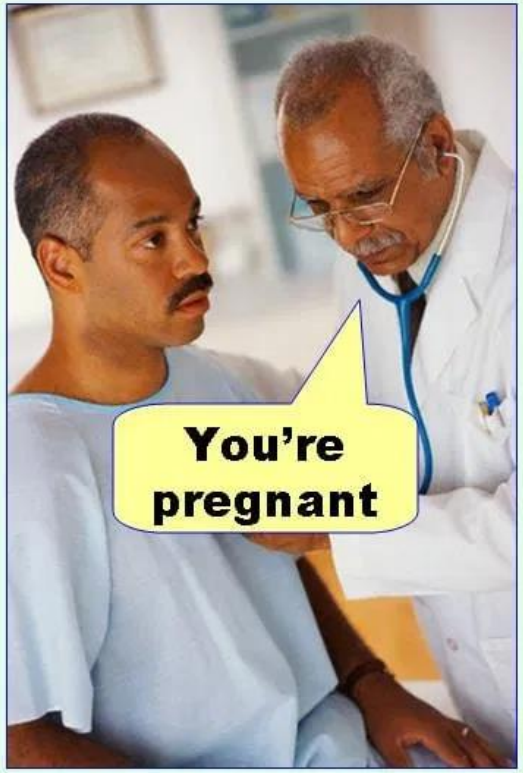
Матрица ошибок

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

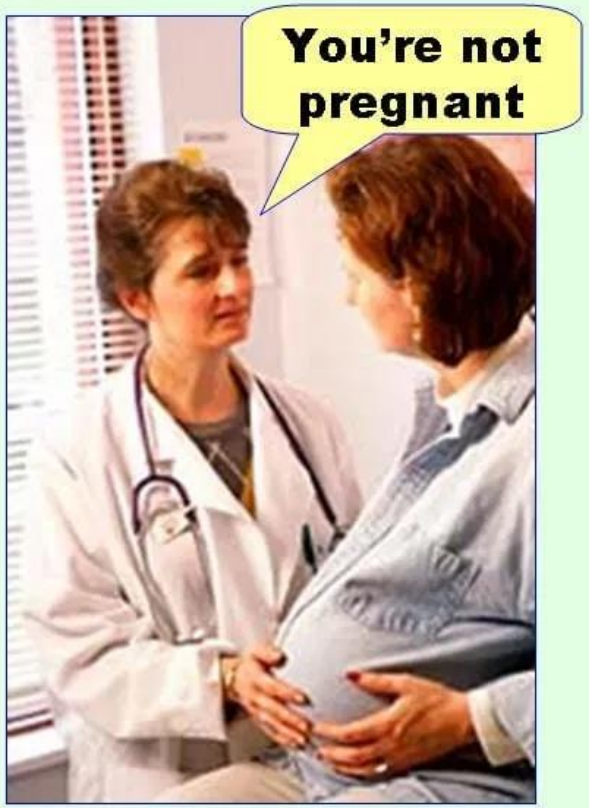
Error matrix or confusion matrix

Матрица ошибок

Type I error
(false positive)



Type II error
(false negative)



Точность и полнота

- Вводят дополнительные метрики, которые называются точность (Precision, PPV) и полнота (Recall, Sensitivity, TPR, acceptance)
- Полнота – показывает долю правильных ответов на сигнальных объектах, другими словами какая доля объектов класса 1 определена правильно.

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

- Точность – показывает насколько мы можем доверять модели: это отношение числа верно определенных сигнальных объектов ко всем объектам, которые модель определила как сигнальные.

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

Пример: медицинская диагностика

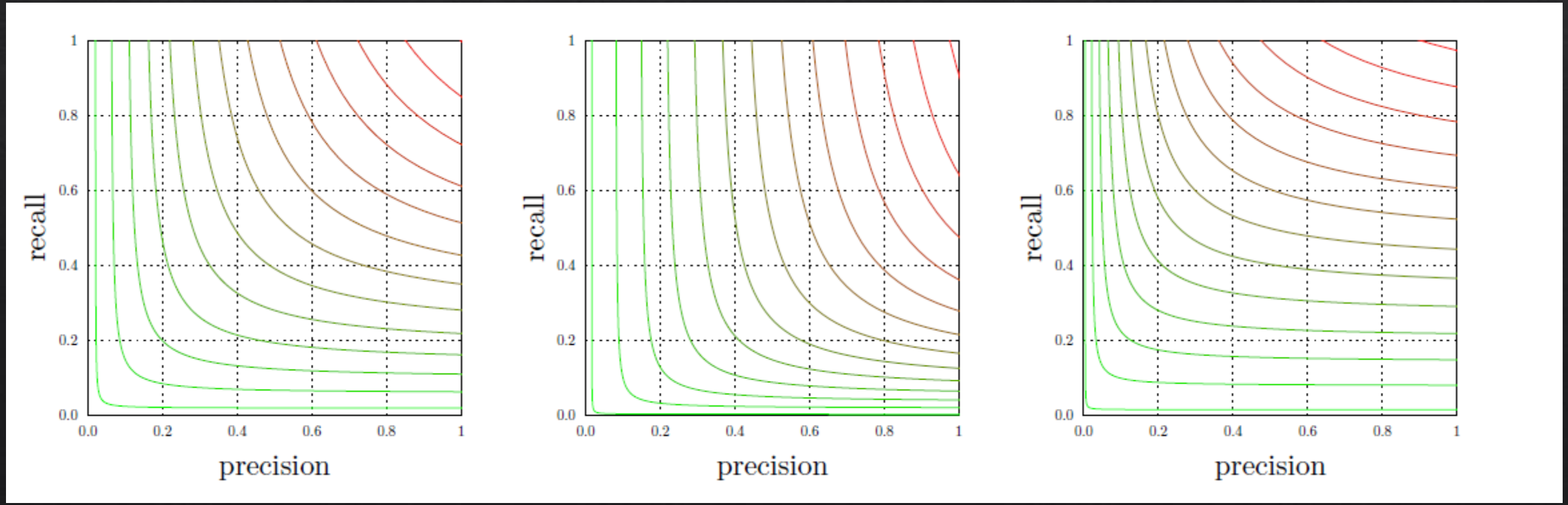
	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	20	200
$y = 0$	180	10000

Какое условие нужно поставить, если необходимо определять болезнь в 90% случаев?

F-мера

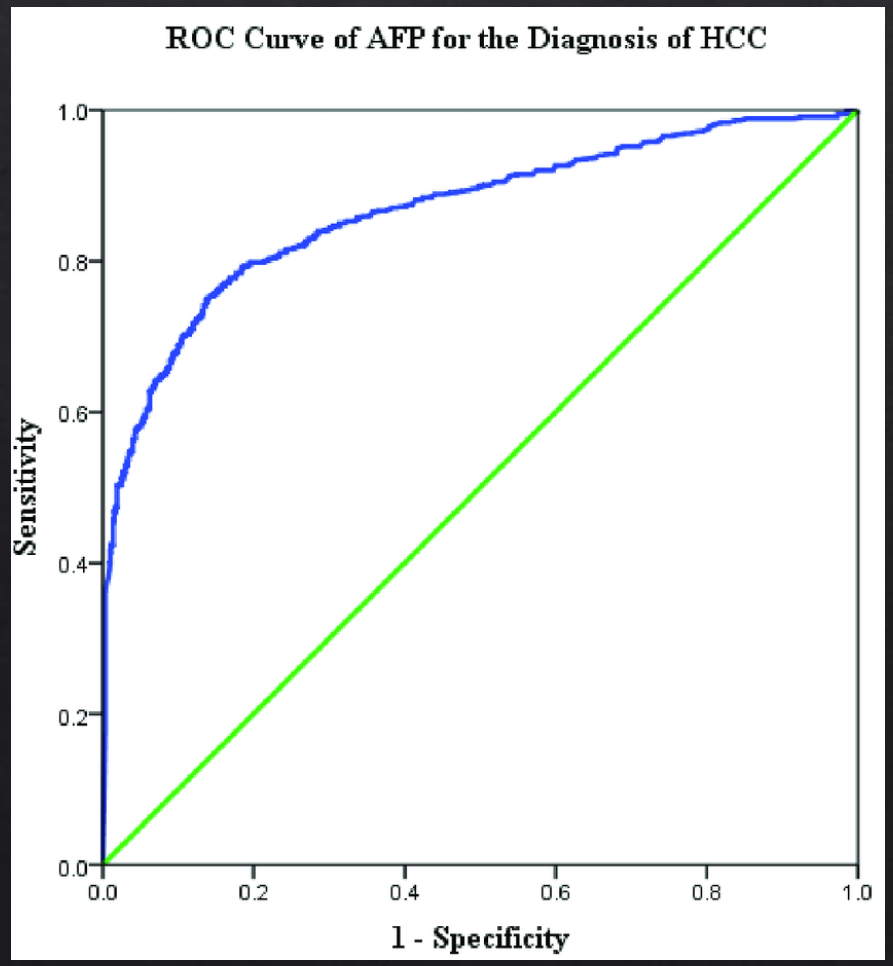
- F-score

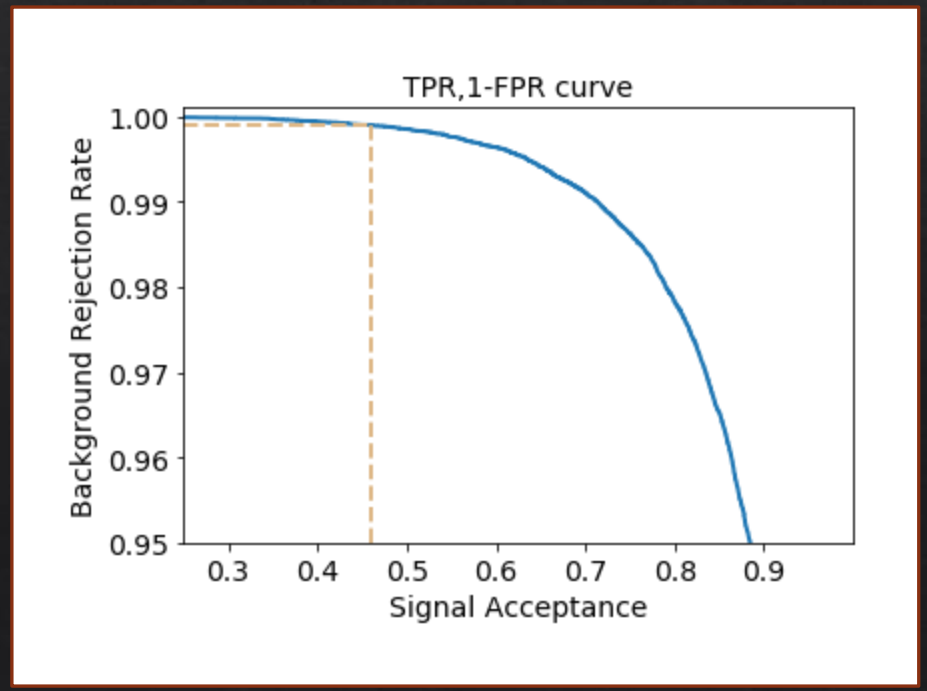
$$F = \frac{1 + \beta^2}{\beta^2} \cdot \frac{precision \cdot recall}{precision + recall}$$



ROC - кривая

- https://en.wikipedia.org/wiki/Receiver_operating_characteristic





Перерыв

Домашнее задание

- Использовать метрики на наборе данных Ирис. Оптимизировать их для выделения класса virginica. Остальные классы считать фоновыми.
- Построить ROC кривую и выбрать на ее основе критерий отбора для аксептанса (полнота) класса virginica 80%.
- Что такое решающее дерево?


Вопросы?