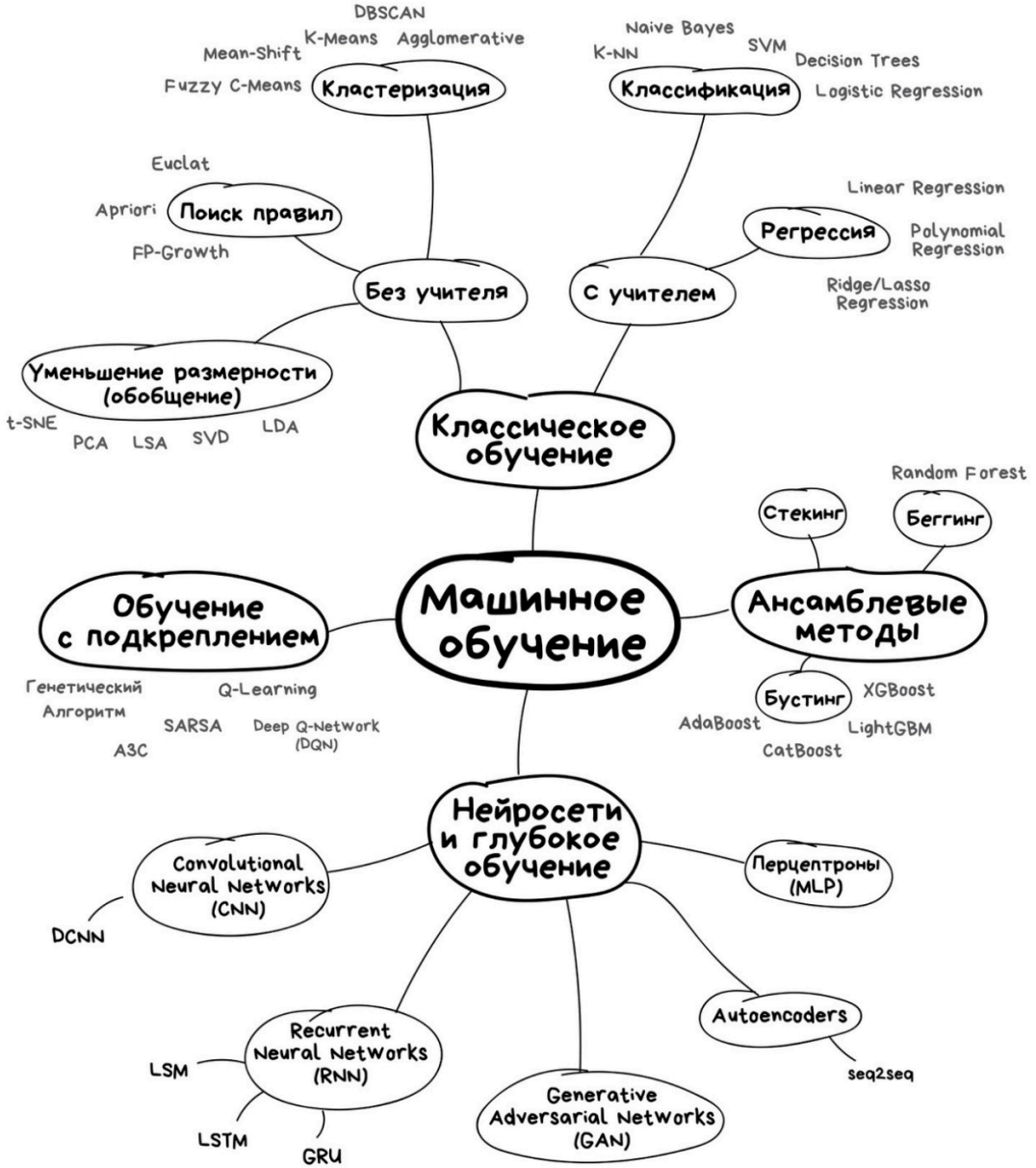
Классификация

Максимовская Анастасия

ОБО МНЕ:

- Data Scientist в Сбербанке
- Преподаватель и автор курсов по машинному обучению в Высшей школе экономики
- Учусь в магистратуре НИУ ВШЭ по финансовым технологиям и анализу данных



Источник: URL

План

1. Метрики классификации
2. Разбор простых задачек
3. Практическая часть

Метрики качества

- Точность (accuracy)

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

Метрики качества

- Матрица ошибок (confusion matrix):

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

Метрики качества

- Перепишем в контексте матрицы ошибок (confusion matrix):

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

Метрики качества

- Точность (precision) и полнота (recall):

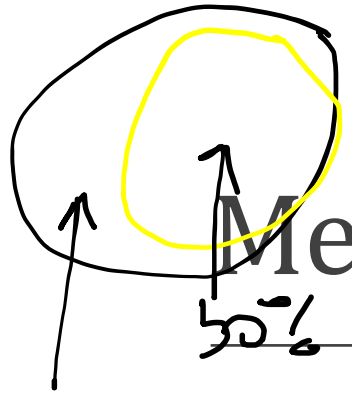
$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}};$$
$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

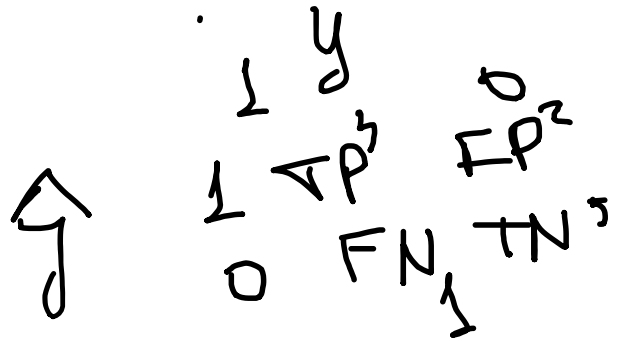
Метрики качества

- F-мера:

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

Метрики качества





- Для задач, связанных с выбором подмножества (выделение лояльных клиентов банка, например) можно использовать прирост концентрации (lift). Если при рассылке предложений о кредите клиентам из подмножества и всем клиентам будет получаться одна и та же доля откликнувшихся, то подмножество не будет представлять особой ценности.

$$\text{lift} = \frac{\text{precision}}{(\text{TP}+\text{FN})/\ell}$$

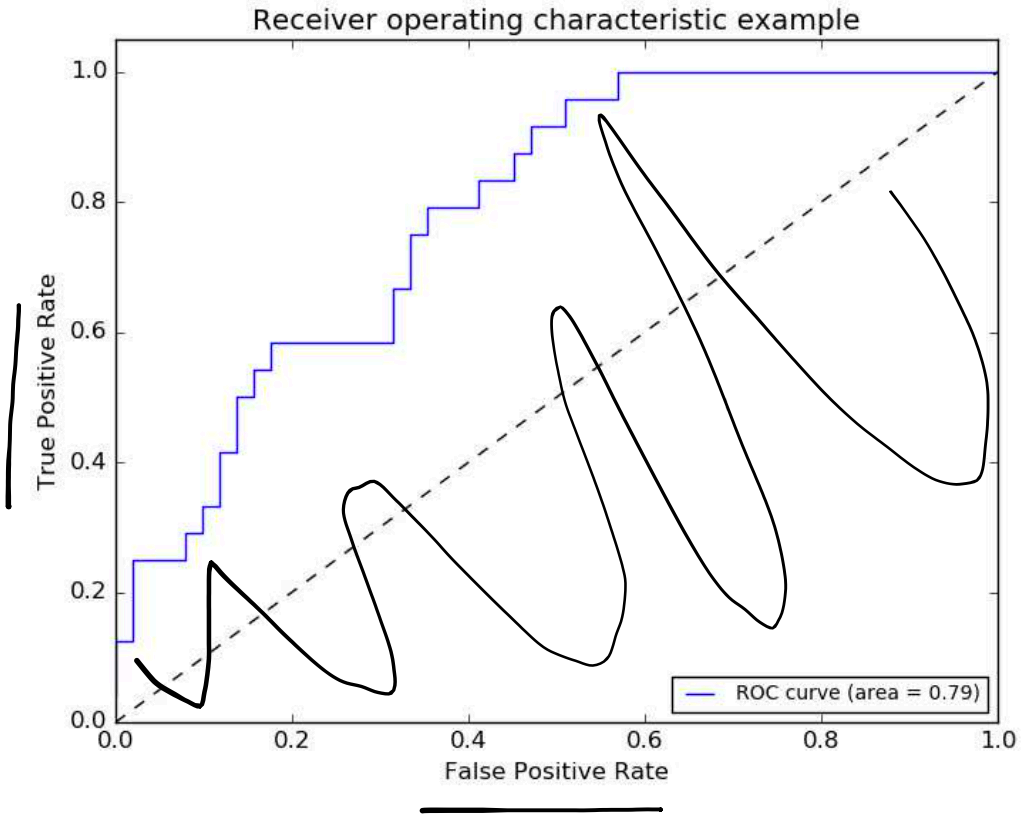
- Улучшение доли положительных объектов в данном подмножестве относительно доли в случайно выбранном подмножестве такого же размера.

Метрики качества

- ROC-AUC (Area under receiver operating characteristic):

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

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



Метрики качества

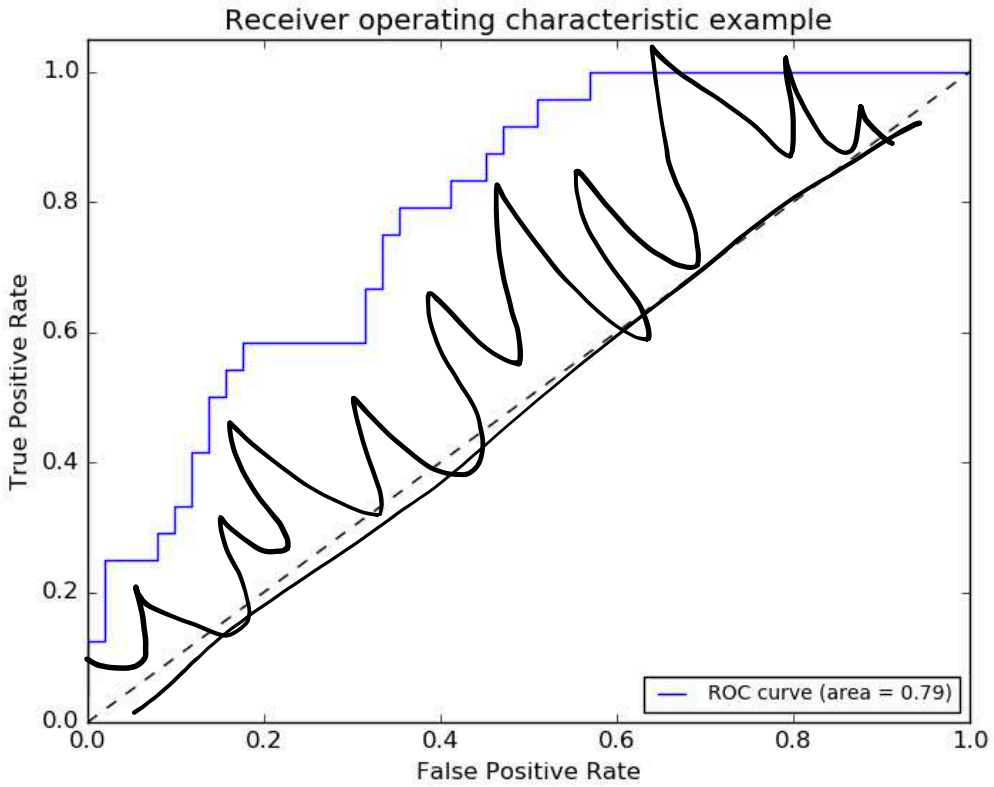
- Индекс Джини:

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

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

$$\mathrm{Gini} = 2\mathrm{AUC} - 1$$

- Это площадь между ROC-кривой и диагональю, соединяющей точки (0,0) и (1, 1).



Крутая статься про Gini: URL

Источники

1. Лекции по машинному обучению на ФКН (URL)
2. Курс по введению в Data Science (URL)
3. Для тех, кто хочет больше: лекция на эту тему на ФКН (URL)

ДОМАШНЕЕ ЗАДАНИЕ:

- Прочитать про решение задач по ссылке. Обратите внимание на упражнение 3!!! Если будут вопросы, то пишите, разберем на следующем вебинаре (URL)