Математический анализ данных и машинное обучение

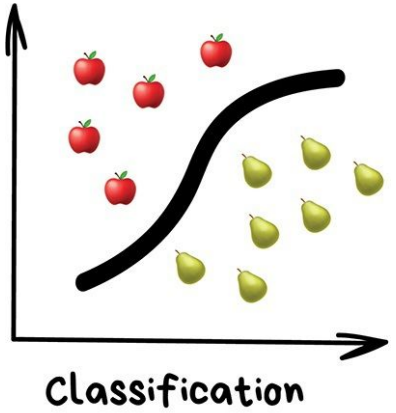
Лекция 2

Саркисян Вероника

План на сегодня

13:30 - 14:30	Лекция 1: Задача классификации
14:45 - 16:00	Лекция 2: Логистическая регрессия, KNN, SVM.
16:00 - 17:00	Перерыв
17:00 - 18:00	Семинар 1: Предобработка данных
18:15 - 19:30	Семинар 2: Методы классификации

Задача классификации



- **Бинарная классификация:** Y = {0,1}

  Пример: классификация спама (спам / не спам)

- **Многоклассовая классификация:** Y = {0,1, …, n}

  Пример: Определение языка текста (русский / украинский / белорусский / ...)

- **Классификация с пересекающимися классами:** Y = {(1,0,..,1,0), ...}

  Пример: определение жанра фильма. Фильм может относиться сразу к нескольким жанрам (триллер, драма, криминал, детектив и т.д.)

Метрики качества в задаче классификации

Самая простая и интуитивная метрика - accuracy (доля правильных ответов):

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

Чем она плоха? Рассмотрим задачу с сильно несбалансированными классами: распределение классов 90/10.

Accuracy у константного классификатора (a(x) = 0) = 90%.

Матрица ошибок (Confusion matrix)

Реальные метки классов

Прогноз модели

	y = 1	y = 0
a(x) = 1	**T**rue **P**ositive	**F**alse **P**ositive
a(x) = 0	**F**alse **N**egative	**T**rue **N**egative

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y_i	0	1	0	0	1	0	1
$a(x_i)$	0	0	1	1	0	1	1

TP = ?

TN = ?

FP = ?

FN = ?

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y_i	0	1	0	0	1	0	1
$a(x_i)$	0	0	1	1	0	1	1

TP = 1

TN = 1

FP = 3

FN = 2

Точность, полнота, F-мера

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

Доля верно классифицированных объектов

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

Доля положительных объектов относительно всех положительно определенных алгоритмом объектов

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

Доля всех найденных положительных объектов

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

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y_i	0	1	0	0	1	0	1
$a(x_i)$	0	0	1	1	0	1	1

Accuracy = 2 / 7

Precision = 1 / 4

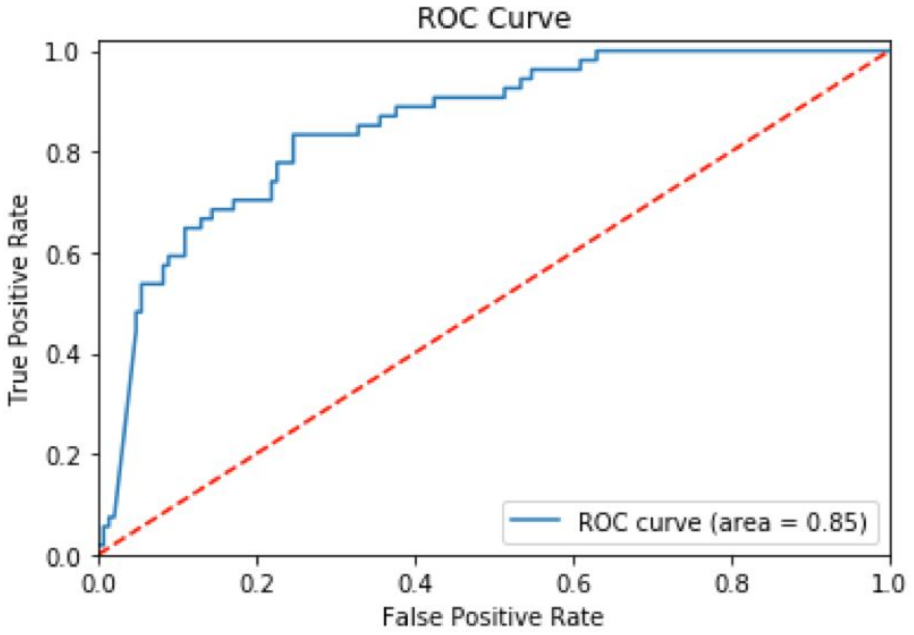
Recall = 1 / 3

AUC-ROC

Строим кривую в координатах False Positive Rate / True Positive Rate:

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

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



	x_1	x_2	x_3	x_4	x_5
y_i	0	1	0	1	1
$a(x_i)$	0.2	0.4	0.1	0.7	0.05

Многоклассовая классификация

Сводим задачу с серии бинарных задач:

- Один против всех (**one-vs-all**): обучаем K (K = число классов) бинарных классификаторов, каждый из которых отделяет свой класс ото всех прочих. Итоговый класс = ответ самого “уверенного” классификатора.
- Все против всех (**all-vs-all**): обучаем столько же классификаторов, сколько пар классов (C_K^2). Новый объект подаем на вход каждому из классификаторов, итоговый класс = тот, за который больше всего “голосов”.

Метрики качества многоклассовой классификации

Микро-усреднение

Усредняем микро-метрики по каждому классу:

$$\overline{\mathrm{TP}} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{TP}_k.$$

Вычисляем итоговые метрики:

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

Макро-усреднение

Вычисляем метрики для каждого из классов:

$$\mathrm{precision}_k(a, X) = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}.$$

Итоговые метрики = среднее по метрикам всех классов:

$$\mathrm{precision}(a, X) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{precision}_k(a, X)$$

Тест

1) Вычислите матрицу ошибок, accuracy, precision, recall:

	x_1	x_2	x_3	x_4	x_5	x_6
y_i	0	1	0	1	1	0
$a(x_i)$	1	1	0	0	1	1

2) Вычислите AUC-ROC:

	x_1	x_2	x_3	x_4	x_5
y_i	0	0	0	1	1
$a(x_i)$	0.1	0.3	0.6	0.7	0.4

Перерыв

Логистическая регрессия

Возьмем уже привычную нам линейную регрессию и обучим ее отличать положительный класс от отрицательного:

$$a(x) = \text{sign}\left(\langle w, x\rangle + w_0\right) = \text{sign}\left(\sum_{j=1}^{d} w_j x_j + w_0\right)$$

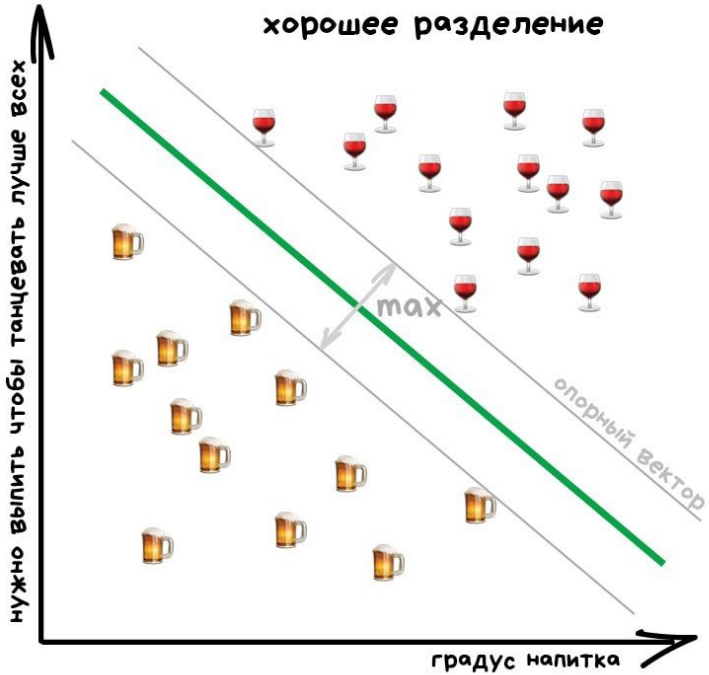
Для обучения модели в качестве метрики качества попробуем использовать долю правильных ответов:

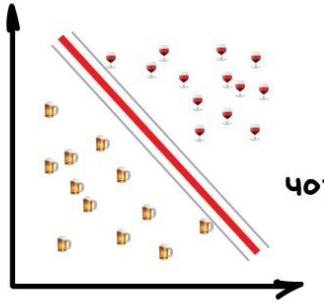
$$Q(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i) = y_i]$$

Оценки для функции потерь:

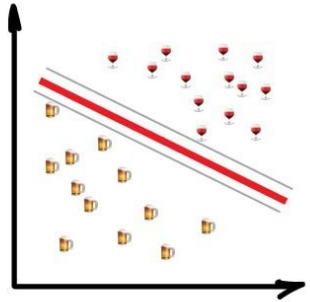
1. $\tilde{L}(M) = \log\left(1 + e^{-M}\right)$ — логистическая функция потерь
2. $\tilde{L}(M) = (1 - M)_+ = \max(0, 1 - M)$ — кусочно-линейная функция потерь (используется в методе опорных векторов)
3. $\tilde{L}(M) = (-M)_+ = \max(0, -M)$ — кусочно-линейная функция потерь (соответствует персептрону Розенблатта)
4. $\tilde{L}(M) = e^{-M}$ — экспоненциальная функция потерь
5. $\tilde{L}(M) = 2/(1 + e^{M})$ — сигмоидная функция потерь

Разделяем виды алкоголя





чот не очень =/



ты пьян,
иди домой

Метод Опорных Векторов

Метод опорных векторов: разделимый случай

Будем рассматривать классификаторы вида:

$$a(x) = \text{sign}(\langle w, x \rangle + b), \qquad w \in \mathbb{R}^d, b \in \mathbb{R}.$$

Расстояние от объекта до разделяющей гиперплоскости:

$$\rho(x_0, a) = \frac{|\langle w, x \rangle + b|}{\|w\|}.$$

Расстояние от гиперплоскости до ближайшего объекта выборки:

$$\min_{x \in X^\ell} \frac{|\langle w, x \rangle + b|}{\|w\|} = \frac{1}{\|w\|} \min_{x \in X} |\langle w, x \rangle + b| = \frac{1}{\|w\|}.$$

(здесь воспользовались тем, что можно одновременно умножать w и b на положительную константу)

Оптимизационная задача:

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,b} \\ y_i\left(\langle w, x_i \rangle + b\right) \geqslant 1, \quad i = 1, \ldots, \ell. \end{cases}$$

Метод опорных векторов: неразделимый случай

Введем штраф за попадание объектов внутрь разделяющей полосы:

$$y_i\left(\langle w, x_i\rangle + b\right) \geqslant 1 - \xi_i, \quad i = 1, \ldots, \ell.$$

Новая оптимизационная задача: параметр C отвечает за то, как сильно мы штрафуем за попадание внутрь полосы.

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,b,\xi} \\ y_i\left(\langle w, x_i\rangle + b\right) \geqslant 1 - \xi_i, \quad i = 1, \ldots, \ell, \\ \xi_i \geqslant 0, \quad i = 1, \ldots, \ell. \end{cases}$$

k-NN (метод k ближайших соседей)

1. Вычисляем расстояние до каждого из объектов обучающей выборки.
2. Выбираем k объектов обучающей выборки, расстояние до которых минимально.
3. Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей.