

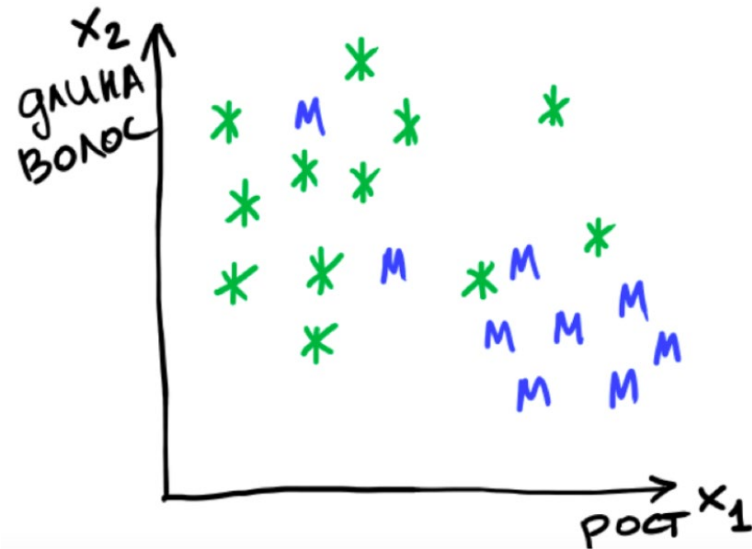


Алгоритмы классификации

ФИПЛ-2022

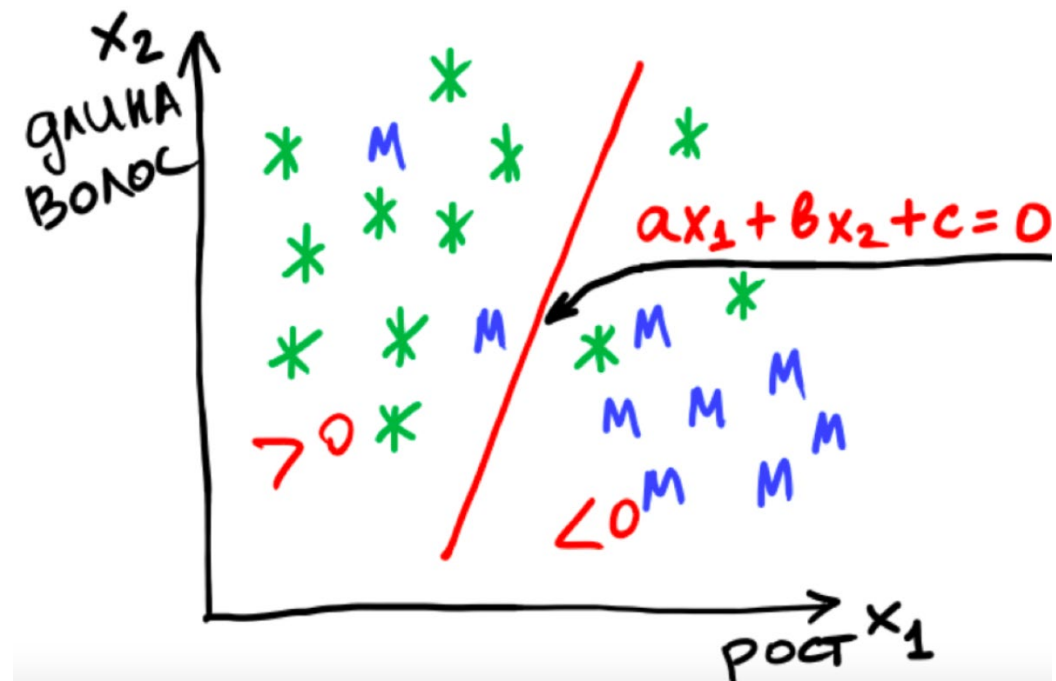
Задача классификации

Бинарный случай: у нас два класса, положительный и отрицательный



Задача классификации

Нужно провести линию так, чтобы разделить классы



Вспомним линейную регрессию

$$w_1x_1 + w_2x_2 + \dots + w_0 = y$$

Мы умеем подбирать веса (коэффициенты) таким образом, чтобы получалось какое-то чиселко.

Как сделать так, чтобы это чиселко определяло один из двух классов?

Классификация

$$w_1x_1 + w_2x_2 + \dots + w_0 = y$$

Мы умеем подбирать веса (коэффициенты) таким образом, чтобы получалось какое-то чиселко.

Как сделать так, чтобы это чиселко определяло один из двух классов?

Использовать знак числа: положительный – один класс, отрицательный – другой.

$$a(x, w) = \textit{sign}(\sum_{j=1}^l w_j x_j)$$

Классификация

$$a(x, w) = \textit{sign}(\sum_{j=1}^l w_j x_j)$$

- если $\sum_{j=1}^l w_j x_j > 0$, то у нас положительный класс;
- если $\sum_{j=1}^l w_j x_j < 0$, то у нас отрицательный класс;
- Получается, что $\sum_{j=1}^l w_j x_j = 0$ – уравнение разделяющей границы (плоскости) между классами.

Как обучить классификатор?

- Обучение – это минимизация функционала ошибки. Для линейной регрессии какую функцию минимизируем?

Как обучить классификатор?

- Обучение – это минимизация функционала ошибки. Для линейной регрессии какую функцию минимизируем?
- Обычно MSE: $\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2$
- Но с классификацией ошибки посчитать проще: у нас либо правильный класс, либо неправильный. Значит, формула упрощается:

$$Q(a, x) = \frac{1}{n} \sum_{i=1}^n [a(x_i) \neq y_i] \rightarrow \min,$$

где $[a(x_i) \neq y_i] = 1$, если предсказание неверное, и 0, если верное.

Как обучить классификатор?

$$Q(a, x) = \frac{1}{n} \sum_{i=1}^n [a(x_i) \neq y_i] \rightarrow \min,$$

где $[a(x_i) \neq y_i] = 1$, если предсказание неверное, и 0, если верное.

- Но эту функцию нельзя дифференцировать ☹
- Значит, нужно использовать какие-то другие функции
- Их напридумывали очень много! В зависимости от того, какая у нас функция потерь, такой и классификатор!
- Например, у логистической регрессии – логистическая функция потерь

Метрики качества

○Accuracy: $accuracy(a, x) = \frac{1}{n} \sum_{i=1}^n [a(x_i) \neq y_i]$





○Precision: $precision(a, x) = \frac{TP}{TP+FP}$

○Recall: $recall(a, x) = \frac{TP}{TP+FN}$

○F-score: $\frac{2 \times precision \times recall}{precision + recall}$

Матрица ошибок

Confusion matrix

		PREDICTIVE VALUES	
		POSITIVE (CAT)	NEGATIVE (DOG)
ACTUAL VALUES	POSITIVE (CAT)	<p>TRUE POSITIVE</p>  <p>3</p> <p>YOU ARE A CAT</p>	<p>FALSE NEGATIVE</p>  <p>1</p> <p>YOU ARE A DOG</p> <p>TYPE II ERROR</p>
	NEGATIVE (DOG)	<p>FALSE POSITIVE</p>  <p>2</p> <p>YOU ARE A CAT</p> <p>TYPE I ERROR</p>	<p>TRUE NEGATIVE</p>  <p>4</p> <p>YOU ARE NOT A CAT</p>

Матрица ошибок

Пример:

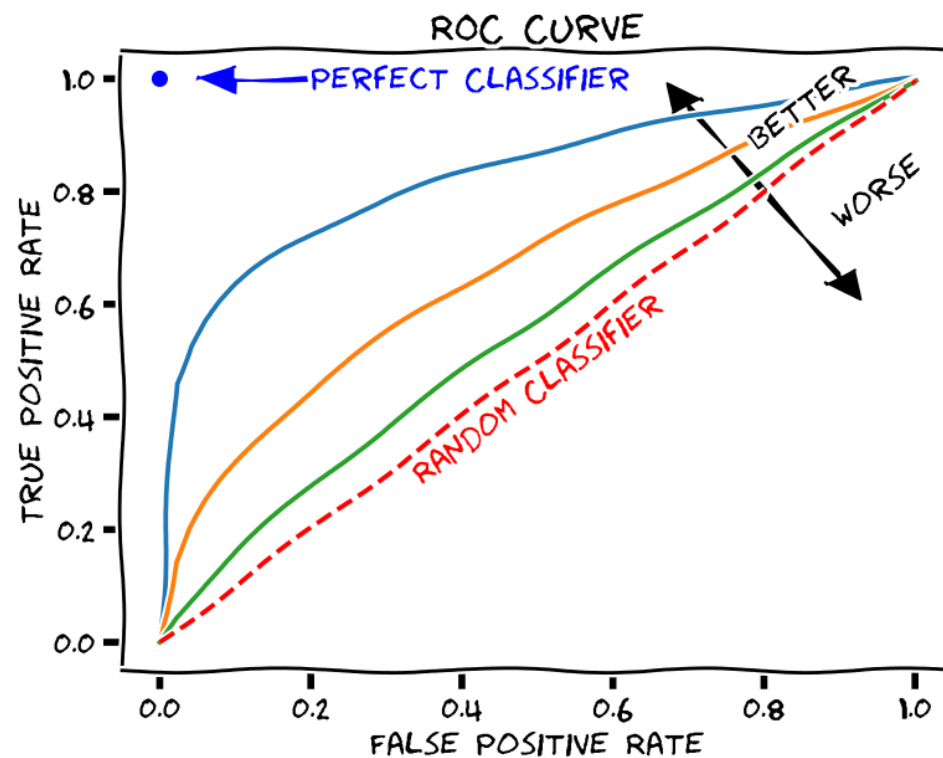
Посчитайте метрики качества. Какой расклад лучше?

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$a(x) = 1$ Получили кредит	80	20
$a(x) = -1$ Не получили кредит	20	80

	$y = 1$ Могут вернуть	$y = -1$ Не могут вернуть
$a(x) = 1$ Получили кредит	48	2
$a(x) = -1$ Не получили кредит	52	98

Метрики качества

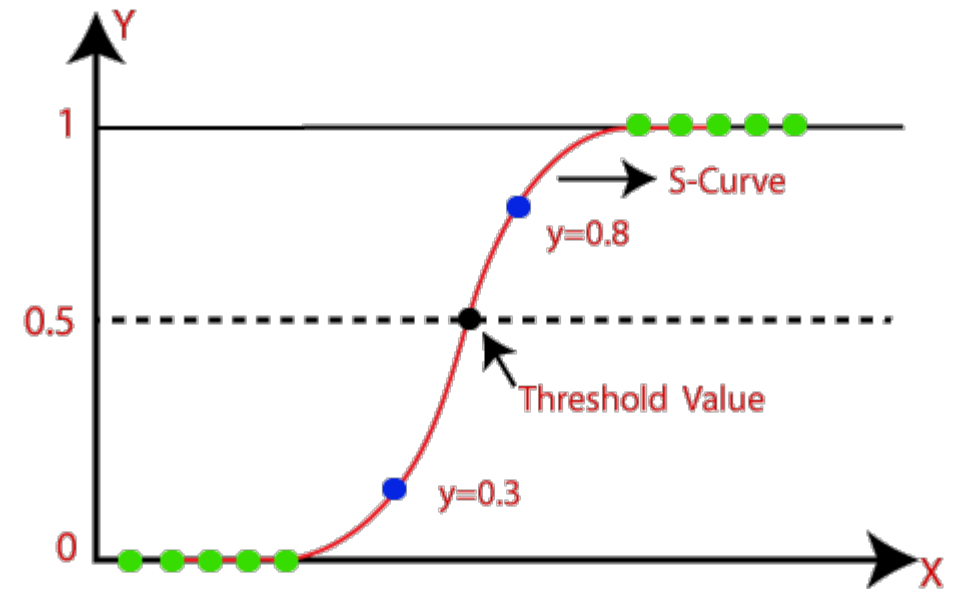
Есть еще т.н. ROC-AUC кривая:



Алгоритмы классификации

Логистическая регрессия

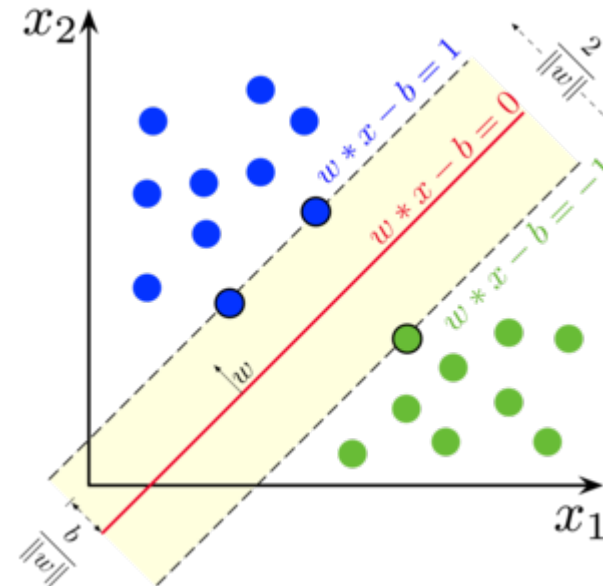
- Берем обычную линейную регрессию, но прогоняем ее ответы через специальную функцию, которая их загонит в интервал от 0 до 1.
- Получится вероятность принадлежности к классу.
- Функция – сигмоида: $\frac{1}{1+e^{-x}}$



Алгоритмы классификации

Метод опорных векторов

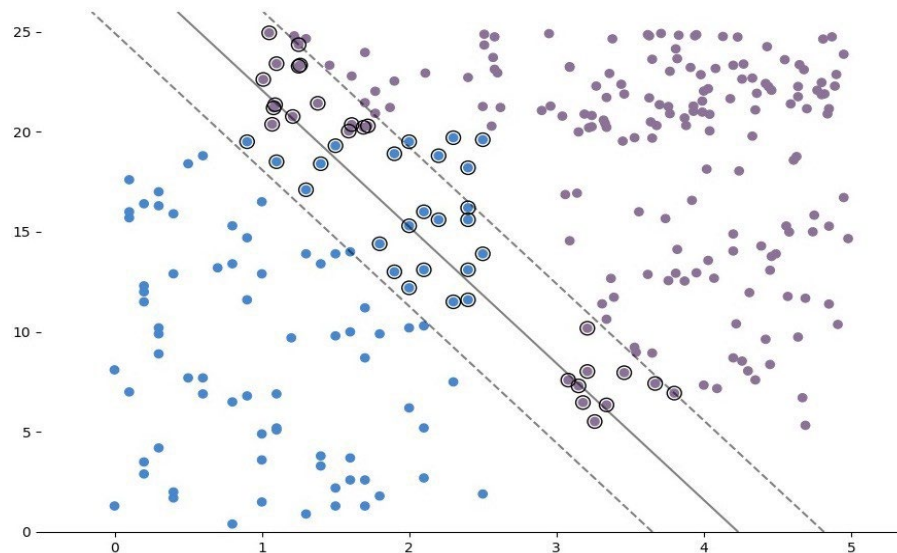
- ...тут я было хотела понаписать страшных формул, но пожалела вас.
- Максимизируем расстояние между точками и нашей разделяющей прямой: то есть, добиваемся того, чтобы желтая полоса на рисунке была как можно шире.



Алгоритмы классификации

Метод опорных векторов

- Но что делать, если наши объекты слишком перемешаны?



- Можно ввести штрафы за попавшие на проезжую часть объекты: теперь максимизируем ширину полосы, но еще и минимизируем штрафы.
- У SVM есть гиперпараметр C : чиселко, которое позволяет находить баланс между шириной полосы и количеством штрафов.

Алгоритмы классификации

Наивный Байес

○ Теорема Байеса:

$$P(c|x) = \frac{P(x|c) \cdot P(c)}{P(x)}$$

- $P(c|x)$ – вероятность того, что объект со значением признака x принадлежит классу c
- $P(c)$ – априорная вероятность класса c
- $P(x|c)$ – вероятность того, что значение признака равно x , при условии, что объект принадлежит классу c
- $P(x)$ – априорная вероятность значения признака x

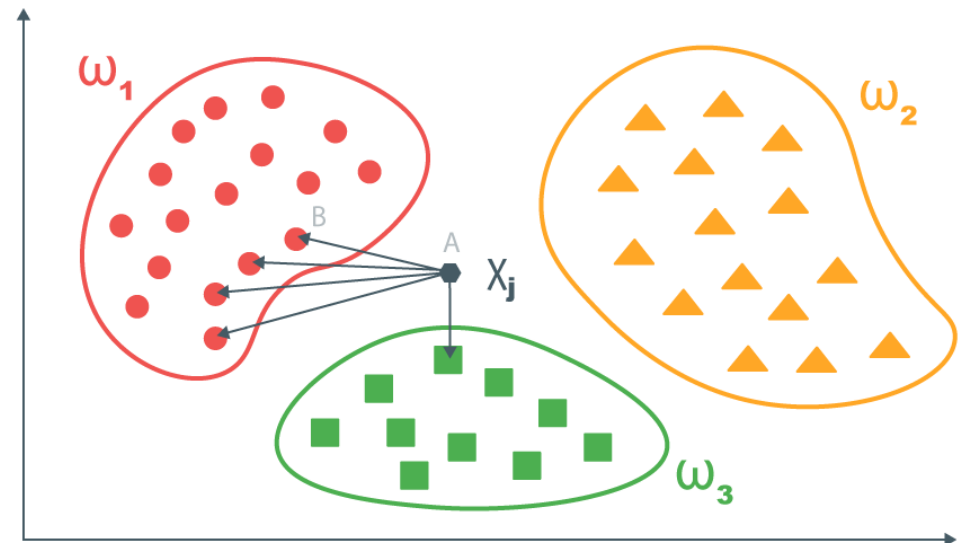
Алгоритмы классификации

Метод К ближайших соседей

Идея: схожие объекты находятся близко друг к другу в пространстве признаков.

Как классифицировать новый объект?

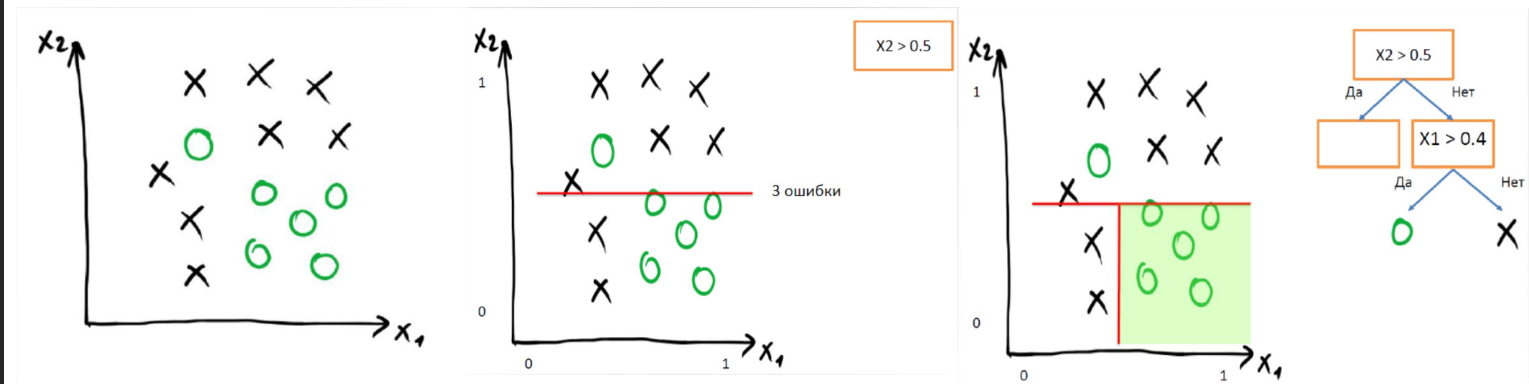
- Вычислить расстояние до каждого из объектов обучающей выборки
- Выбрать k объектов обучающей выборки, расстояние до которых минимально
- Соседи решают класс голосованием! Соседей какого класса больше, тот и наш.



Алгоритмы классификации

Деревья решений

Пробуем делить выборку таким образом, чтобы было как можно меньше ошибок. Выбираем всегда только один признак! Делим, пока не получим идеальный результат (или нет...)



GridSearchCV и гиперпараметры

- *Параметры модели* – величины, настраивающиеся по обучающей выборке (например, веса в линейной регрессии)
- *Гиперпараметры модели* – величины, контролирующие процесс обучения (например, η – learning rate, C в SVM или K у соседей). Они не могут быть настроены в процессе обучения.

Как подбирать гиперпараметры?

По кросс-валидации: главное – не использовать тестовую выборку!

Для этого есть функция GridSearchCV.

