

Семинар 3-4: привлечение клиентов и классификация

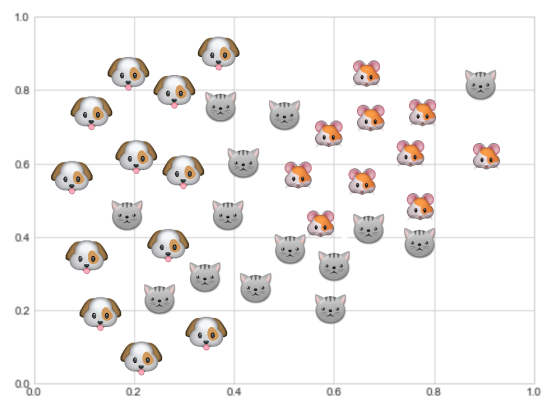
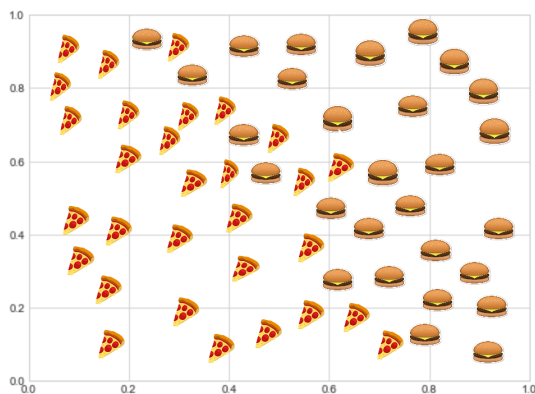
Задача 0

Ликбез! Со всей глубиной и знанием дела дайте ответы на следующие вопросы:

- Почему нумерация в этом семинаре начинается с нуля?
- Что такое машинное обучение и что оно позволяет делать?
- Что такое обучение без учителя и чем оно отличается от обучения с учителем?
- Чем задача классификации отличается от задачи кластеризации?
- Чем задача классификации отличается от задачи регрессии?
- Как оценить модель? Что для этого нужно?
- Что такое метрика качества модели? Какие метрики вы знаете? Как правильно измерить качество модели?
- Что такое кросс-валидация? Как объяснить это бабушке? В смысле не знаете? Мы целый модуль этим занимались!

Задача 1

Нам нужно научиться отделять пиццу от бургеров, а также котиков от пёсиков и от мышек. Проведите на картинках линии, которые отделят одни классы от других. Да, это и есть машинное обучение. Но обычно кривые рисуем не мы, а компютер.



Почему нельзя провести между пиццей и бургерами слишком подробную и извилистую границу? В чём проблема самого правого верхнего котика? Что такое переобучение? Как понять переобучились ли мы?

Задача 2

На плоскости расположены колонии рыжих и чёрных муравьёв. Рыжих колоний три и они имеют координаты $(-1, -1)$, $(1, 1)$ и $(3, 3)$. Чёрных колоний тоже три и они имеют координаты $(2, 2)$, $(4, 4)$ и $(6, 6)$.

- а) Чем KNN отличается от K-means?
- б) Поделите плоскость на «зоны влияния» рыжих и чёрных используя метод одного ближайшего соседа.
- в) Поделите плоскость на «зоны влияния» рыжих и чёрных используя метод трёх ближайших соседей.
- г) С помощью кросс-валидации с выкидыванием отдельных наблюдений выберите оптимальное число соседей k перебрав $k \in \{1, 3, 5\}$. Целевой функцией является количество несоответствующих прогнозов.

Задача 3

Машка пять дней подряд гадала на ромашке, а затем выкладывала очередную фотку «Машка с ромашкой» в инстаграмчик. Результат гадания — переменная y_i , количество лайков у фотки — переменная x_i . Постройте классификационное дерево для прогнозирования y_i с помощью x_i на обучающей выборке:

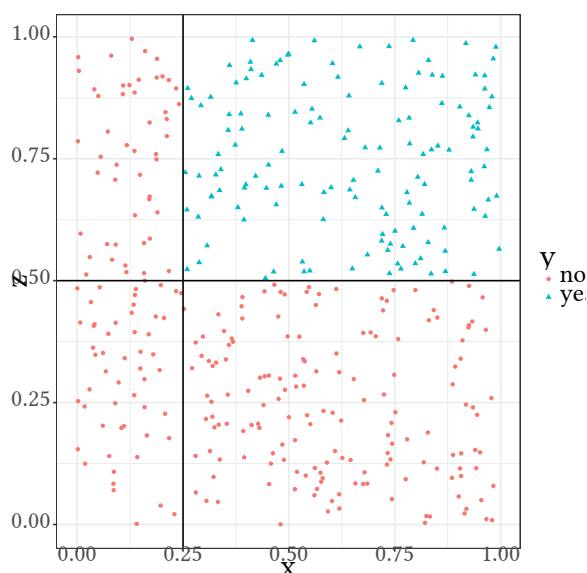
y_i	x_i
плюнет	10
поцелует	11
поцелует	12
к сердцу прижмёт	13
к сердцу прижмёт	14

Дерево строится до идеальной классификации. Критерий деления узла на два — минимизация числа допущенных ошибок¹. Правило прогнозирования в каждой вершине: в качестве прогноза выдаем тот класс, представителей которого в вершине больше. Предположим, что под фоткой стоит 15 лайков, каков будет результат гадания?

Задача 4

По данной диаграмме рассеяния постройте классификационное дерево для зависимой переменной y :

¹На самом деле на практике так не делают. Обычно для разбиения узла при строительстве классификационных деревьев используют энтропию. О том, что это такое, можно погуглить.



Задача 5

Предположим, что мы собираемся выпустить на рынок новое приложение. Нам хотелось бы, чтобы оно привлекло как можно больше пользователей и постоянно оставалось популярным. Предположим, что у нас в руках есть выборка о том, какие приложения есть в данный момент в App Store, какими они обладают параметрами и какой у них рейтинг.

1. Как можно было бы использовать эту выборку?
2. Какие подводные камни нас, скорее всего, ожидают?

Ещё задачи!

Задача 6

Пятачок собрал данные о визитах Винни-Пуха в гости к Кролику. Здесь x_i - количество съеденного мёда в горшках, а y_i - бинарная переменная, отражающая застревание Винни-Пуха при входе

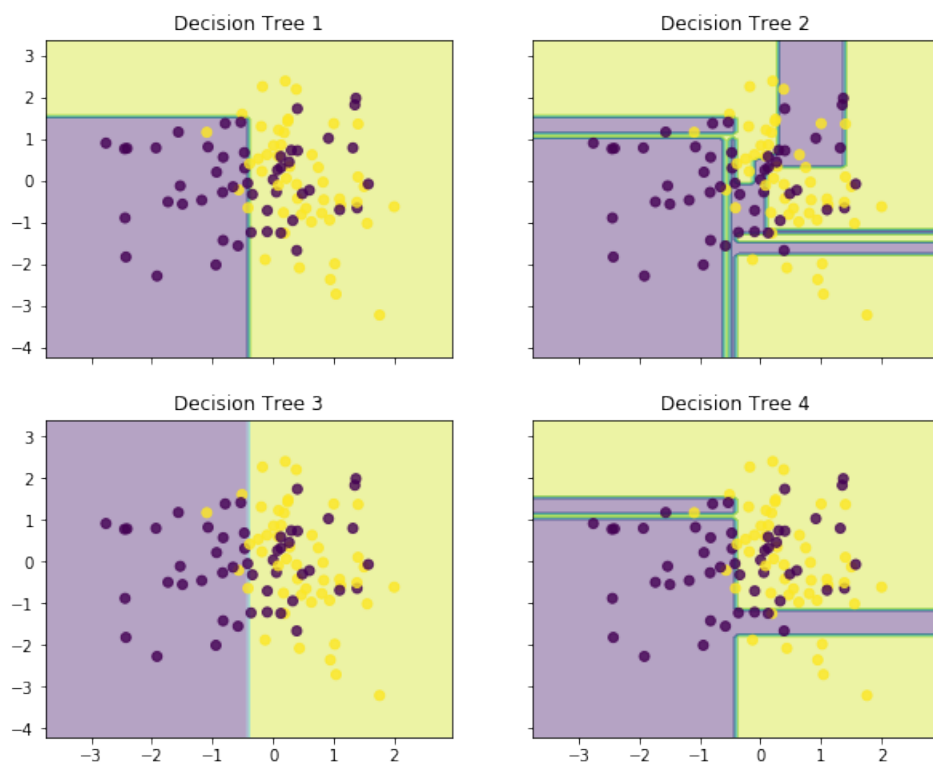
y_i	x_i
0	1
1	4
1	2
0	3
1	3
0	1

- а) Пятачок собирается оценить дерево по всей выборке. Помогите очень маленькому существу сделать это.

- б) Пятачок узнал у Иа-Иа, что оказывается выборку надо делить на тренировочную и тестовую. Поэтому он отложил последние два наблюдения для теста. Оцените дерево по первым четырём наблюдениям и проверьте его работоспособность по последним двум.
- в) Пятачок поговорил с Совой и узнал, что деревья часто переобучаются. Она рассказала ему, что над деревьями надо строить ансамбли. Например, случайный лес. Пятачок решил построить лес из двух деревьев. Первое дерево он строит на наблюдениях с первого по третье, второе на наблюдениях со второго по четвёртое. Третье дерево на наблюдениях 1, 2, 4. Помогите пятачку построить лес и оценить качество его работы на тестовой выборке.

Задача 7

Ниже изображены разделяющие поверхности для задачи бинарной классификации, соответствующие решающим деревьям разной глубины. Какое из изображений соответствует наиболее глубокому дереву? Какой примерной глубине дерева соответствует каждая из картинок?



Задача 8

Рассмотрим обучающую выборку для прогнозирования y с помощью x и z :

y_i	x_i	z_i
y_1	1	2
y_2	1	2
y_3	2	2
y_4	2	1
y_5	2	1
y_6	2	1
y_7	2	1

Будем называть деревья разными, если они выдают разные прогнозы на обучающей выборке. Сколько существует разных классификационных деревьев для данного набора данных?