

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! Слабо дать ответ на каждый вопрос с помощью одного ёмкого слова?

Задача 1

Есть четыре человека и Вася. У всех профили с кинопоиска.

Фильмы какого жанра скорее всего понравятся Васе, если доверить выбор методу одного ближайшего соседа? А если выбирать по двум ближайшим соседям? (АХТУНГ!) А если выбирать по трём соседям? Почему нельзя брать очень много соседей?

Дайте ответы на вопросы выше, используя евклидово расстояние.

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 $(6, 6)$.

1. Чем KNN отличается от K-means?

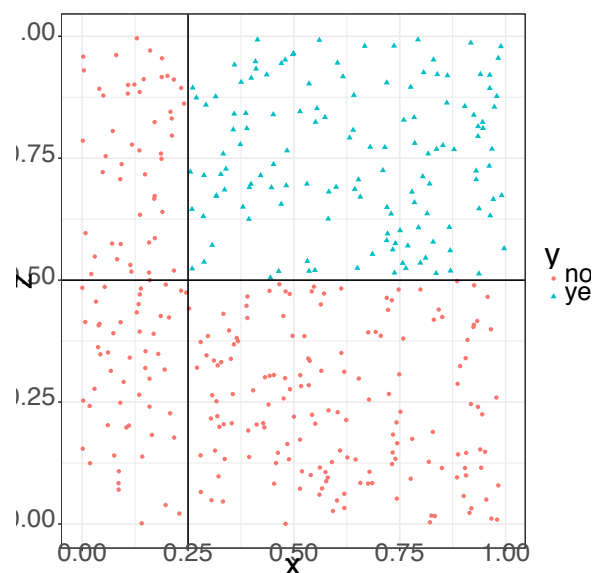
2. Поделите плоскость на «зоны влияния» рыжих и чёрных используя метод одного ближайшего соседа.
3. Поделите плоскость на «зоны влияния» рыжих и чёрных используя метод трёх ближайших соседей.
4. С помощью кросс-валидации с выкидыванием отдельных наблюдений выберите оптимальное число соседей k перебрав $k \in \{1, 3, 5\}$. Целевой функцией является количество несовпадающих прогнозов.

Задача 3

Обучаем дерево своими руками. Правило разбиения вершин: минимизация числа допущенных ошибок. Правило прогнозирования в каждой вершине: в качестве прогноза выдаем тот класс, представителей которого в вершине больше.

Задача 4

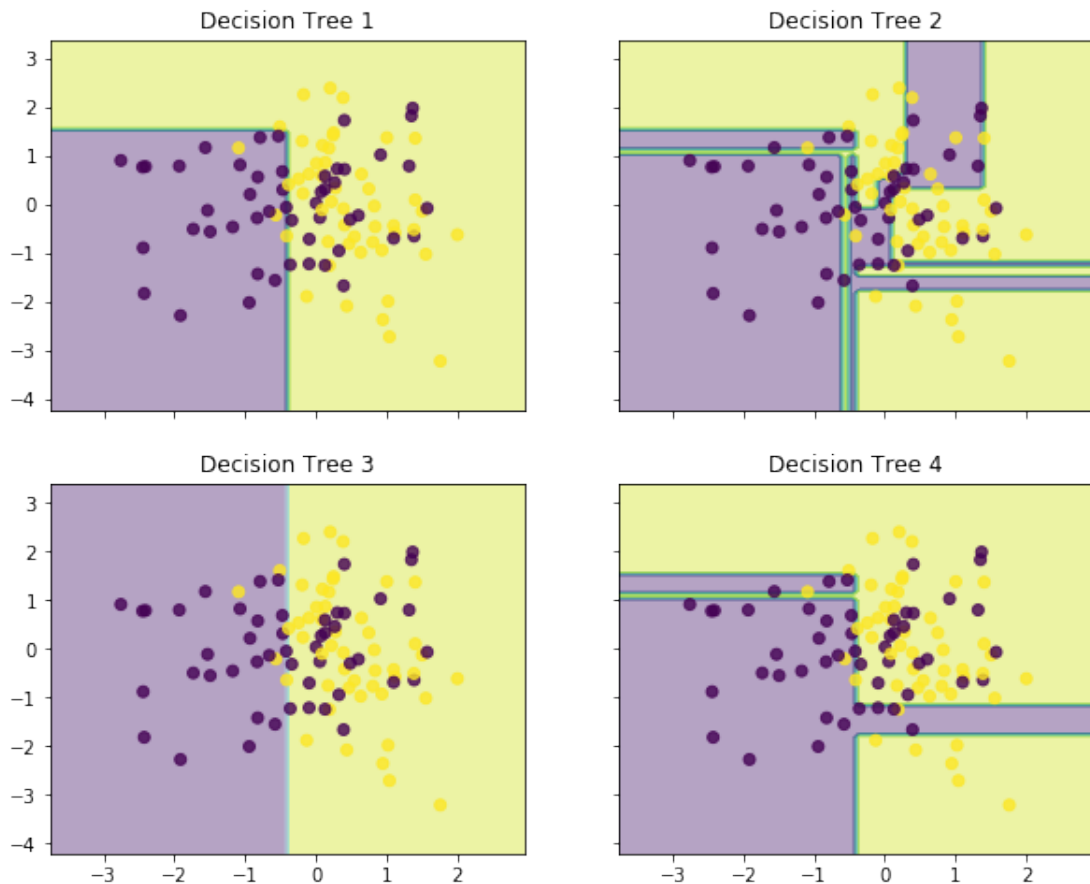
По данной диаграмме рассеяния постройте классификационное дерево для зависимой переменной y :



1. Ещё задачи

Задача 5

Ниже изображены разделяющие поверхности для задачи бинарной классификации, соответствующие решающим деревьям разной глубины. Какое из изображений соответствует наиболее глубокому дереву?



Задача 6

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?

Задача 7

Решите первую и вторую задачи, используя манхеттенское расстояние вместо евклидова. Можно ли подбирать метрику для подсчёта расстояния с помощью кросс-валидации, как мы делали это с параметром k ?