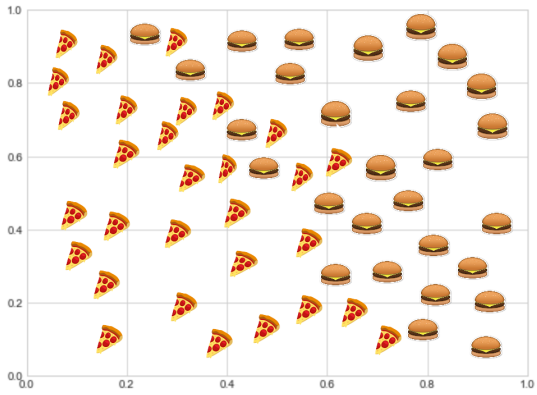
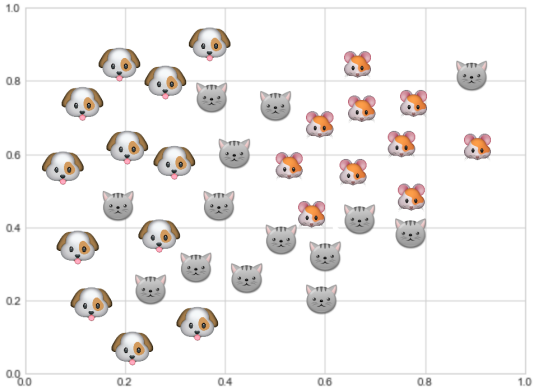
Семинар 7-8 (часть 2): Соседи, деревья, кросс-валидация

Задача 1 (классификация в картинках)

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 компуктер.





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 (KNN, кросс-валидация)

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 несовпадающих прогнозов.

Задача 3 (древо для классификации)

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 ошибок[1]. 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 (древо для регрессии)

Миша работает в маленькой кофейне. Харио Малабар Монсун является фирменным напитком этой кофейни. Мише интересно узнать как именно ведёт себя спрос на напиток y_i в зависимости от температуры за окном t_i. Четыре дня Миша записывал свои наблюдения:

t_i	y_i
21	1
19	2
12	8
8	8

Сегодня он решил обучить регрессионное дерево. В качестве функции потерь он использует

$$\sum (y_i - \hat{y}_i)^2.$$

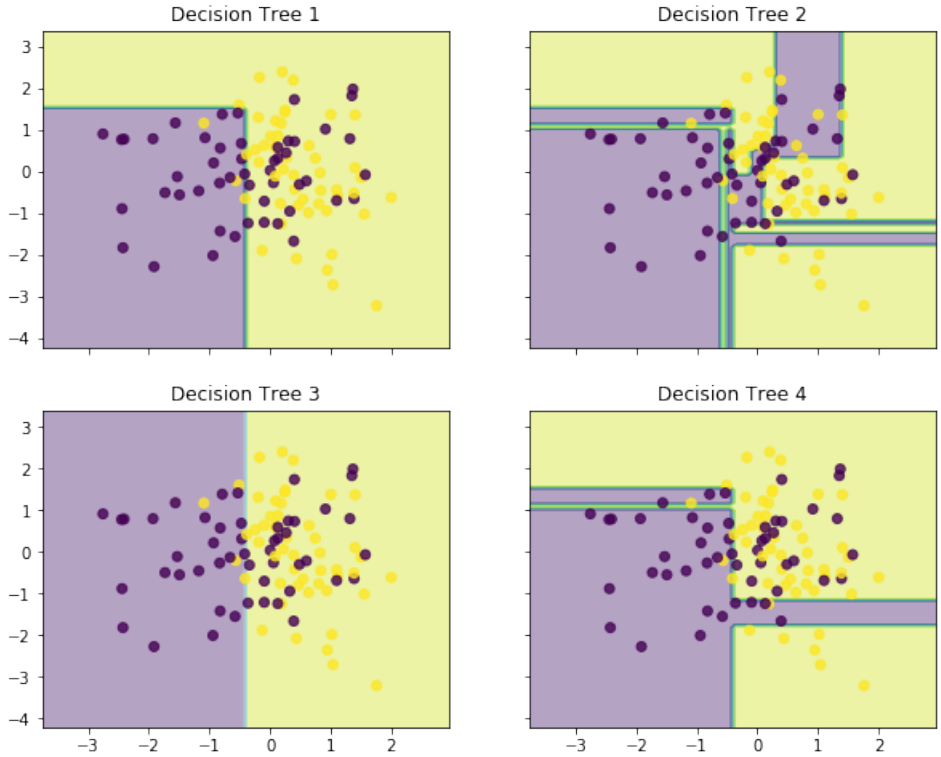
а) Обучите регрессионное дерево.

б) Какой прогноз на сегодня сделает дерево Миши, если за окном 13 градусов?

Задача 5

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

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



Ещё задачи

Тут лежит ещё несколько задач для самостоятельного решения. Возможно, похожие будут в самостоятельной работе...

Задача 6

Выращиваем регрессионное дерево в домашних условиях! Вот вам выборка для этого:

x_i	y_i
0	5
1	6
2	4
3	100

Критерий деления вершины — минимизация квадратичной функции потерь. Критерий остановки — три листа. Зачем нужен критерий остановки? Как дерево ведёт себя с выбросами?

Задача 7

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

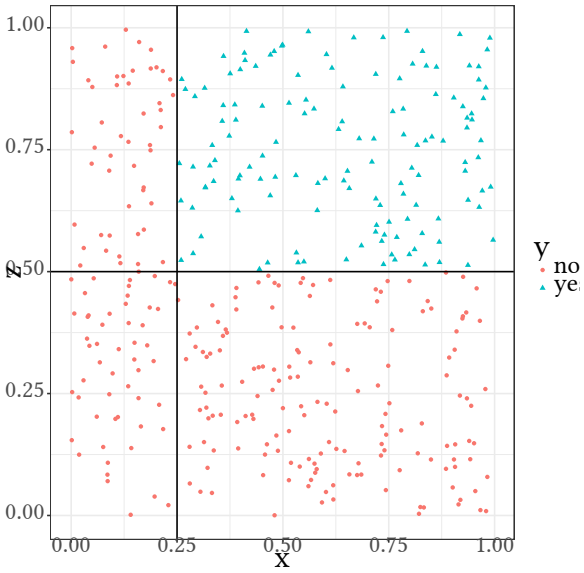
а) Пятачок собирается оценить дерево по всей выборке. Помогите очень маленькому существу сделать это.

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 пяточку построить лес и оценить качество его работы на тестовой выборке.

Задача 8

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



Задача 9

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?