# Семинар 9: алгоритмы классификации

Есть два типа людей.

Макс Корж

На прошлом семинаре мы говорили про то, какими бывают метрики классификации. Обычно их используют, чтобы сравнивать между собой различные алгоритмы для классификации. Ни одного такого алгоритма мы пока ещё не знаем. Пришло время исправить это досадное упущение.

## Упражнение 1 (KNN и 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) Поделите плоскость на «зоны влияния» рыжих и чёрных муравьёв, используя метод одного ближайшего соседа.

б) Поделите плоскость на «зоны влияния» рыжих и чёрных муравьёв, используя метод трёх ближайших соседей.

в) С помощью кросс-валидации с выкидыванием отдельных наблюдений выберите оптимальное число соседей $k$ перебрав $k \in \{1,3,5\}$. Целевой функцией является количество верных предсказаний (accuracy).

## Упражнение 2 (древо для классификации)

У Маши есть инстаграмчик. На неё подписана целая куча парней. Недавно Дима поставил ей 10 лайков. Тогда Маша схватила ромашку и начала гадать. Она нагадала, что Дима плюнет в неё. То же самое она сделала с другими парнями.

Результат гадания — переменная $y_i$, количество лайков у фотки — переменная $x_i$.

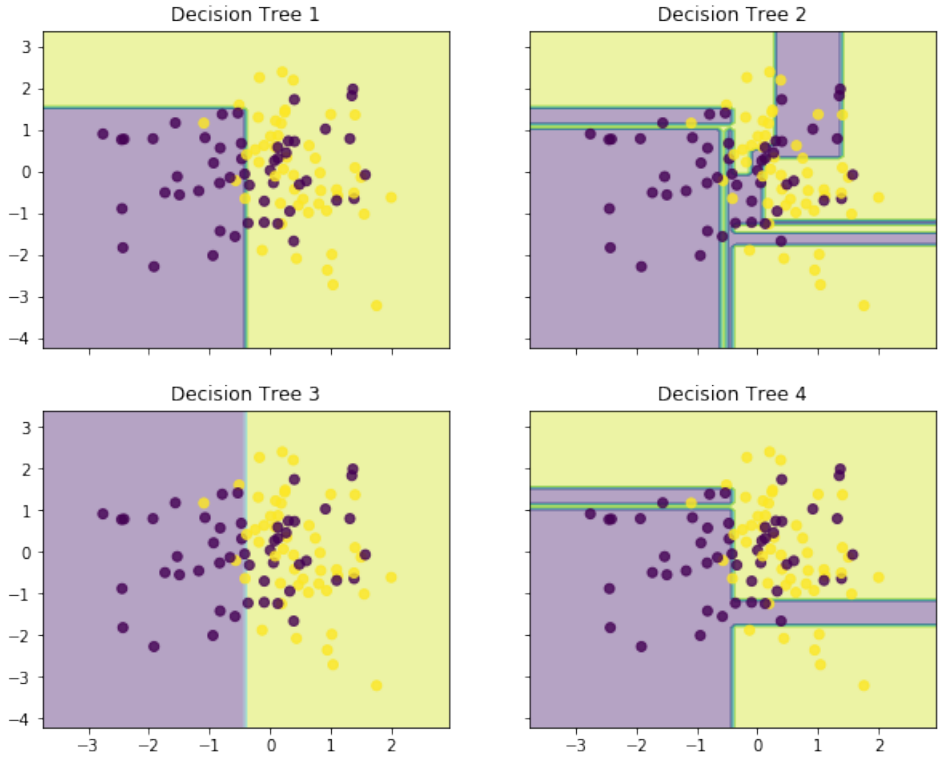
| $y_i$ | $x_i$ |
|---|---|
| плюнет | 10 |
| поцелует | 11 |
| поцелует | 12 |
| к сердцу прижмёт | 13 |
| к сердцу прижмёт | 14 |

Сегодня в Машином инстаграмме Джонни Деп поставил 15 лайков. Маше очень хочется понадать что же с ней сделает Джонни Деп, но у неё кончились ромашки. Поэтому она решила обучить классификационное дерево, которое поможет ей спрогнозировать $y_i$ по $x_i$.

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.

### Упражнение 3 (ещё деревья)

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?




## Ещё задачи

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

### Упражнение 4

На плоскости расположены колонии рыжих и чёрных муравьёв. Рыжих колоний три и они имеют координаты $(-1,1)$, $(1,-1)$ и $(1,1)$. Чёрных колоний одна и она имеет координаты $(0,0)$.

а) Поделите плоскость на «зоны влияния» рыжих и чёрных используя метод одного и трёх ближайших соседей.

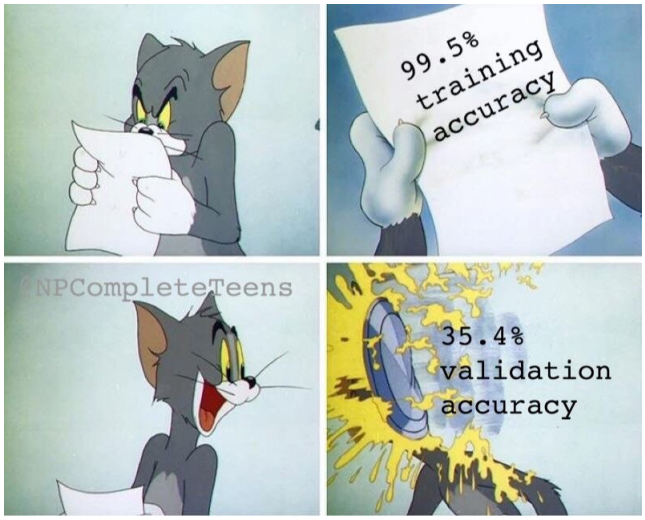
б) С помощью кросс-валидации с выкидыванием отдельных наблюдений выберите оптимальное число соседей $k$ перебрав $k \in \{1,3\}$. Целевой функцией является количество

[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.

несовпадающих прогнозов.

## Упражнение 5

Объясните мемас:




## Упражнение 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 |

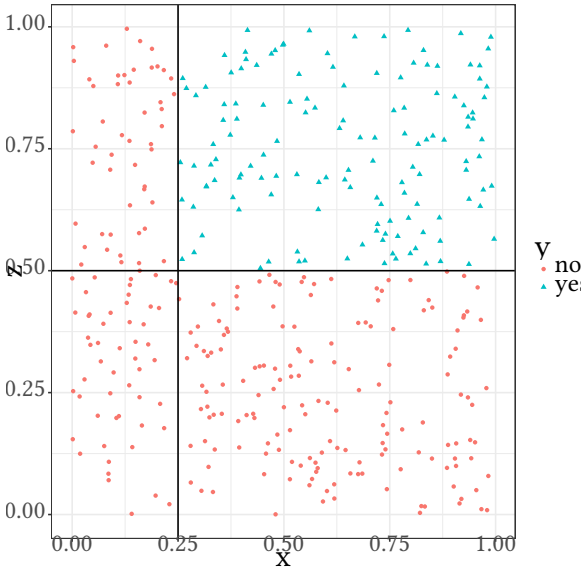
а) Пятачок собирается оценить дерево по всей выборке. Помогите очень маленькому существу сделать это.

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 трёх 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.

## Упражнение 7

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




## Упражнение 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?