МАШИННОЕ ОБУЧЕНИЕ

ОБУЧЕНИЕ С УЧИТЕЛЕМ

МИХАИЛ ЛИПКОВИЧ

ПЛАН

- Логистическая регрессия
- Деревья принятия решений
- Ансамбли моделей

МОТИВАЦИЯ

- У нас есть линейная регрессия, которой мы можем предсказать число (задача регрессии)
- Теперь хотим решить задачу бинарной классификации
- А почему бы просто линейной регрессией не предсказать вероятность принадлежности к классу?
- Не получим вероятность [0, 1] - значения линейной регрессии лежат в (-inf, +inf)
- А давайте тогда просто применим к выходу линейной регрессии функцию, которая переведет все в [0, 1]

СИГМОИД

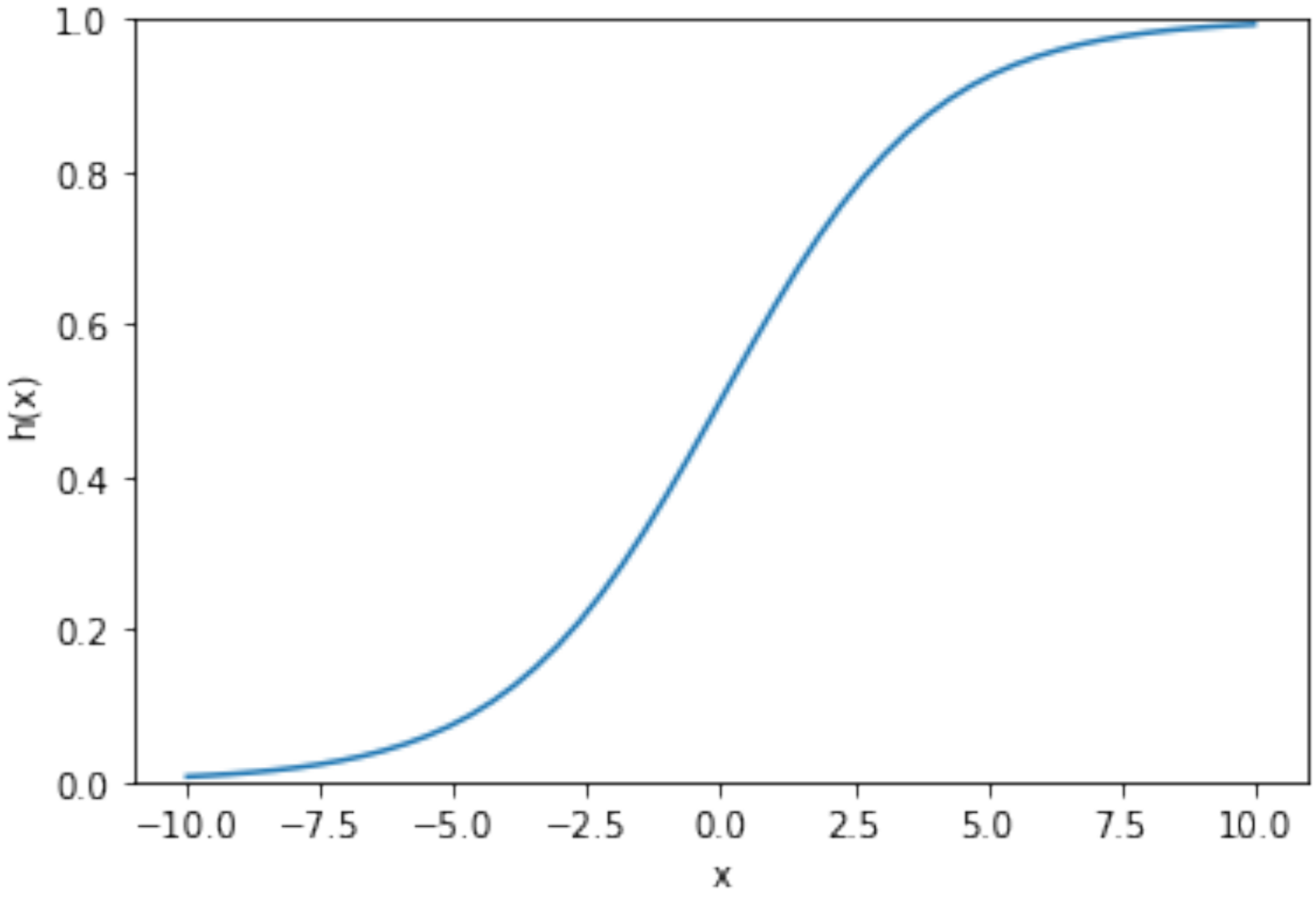
В качестве такой функции можно взять сигмоид:

$$h(x) = \frac{1}{1 + e^{-x}}$$

САМ НАРИСОВАЛ!

В качестве такой функции можно взять сигмоид:

$$h(x) = \frac{1}{1 + e^{-x}}$$



КАК БУДЕМ ОБУЧАТЬ?

КАК БУДЕМ ОБУЧАТЬ?

Введем функцию потерь, а далее градиентный спуск

$$LogLoss = \sum_{x,y \in D} [-y \log(y') - (1-y) \log(1-y')] =$$

$$= \sum_{x,y \in D} \log(1 + e^{-yw^T x})$$

DECISION BOUNDARY

Логистическая регрессия дает на выходе вероятность.

Пока что будем считать что

$h(x) < 0.5 \Rightarrow$ класс 0

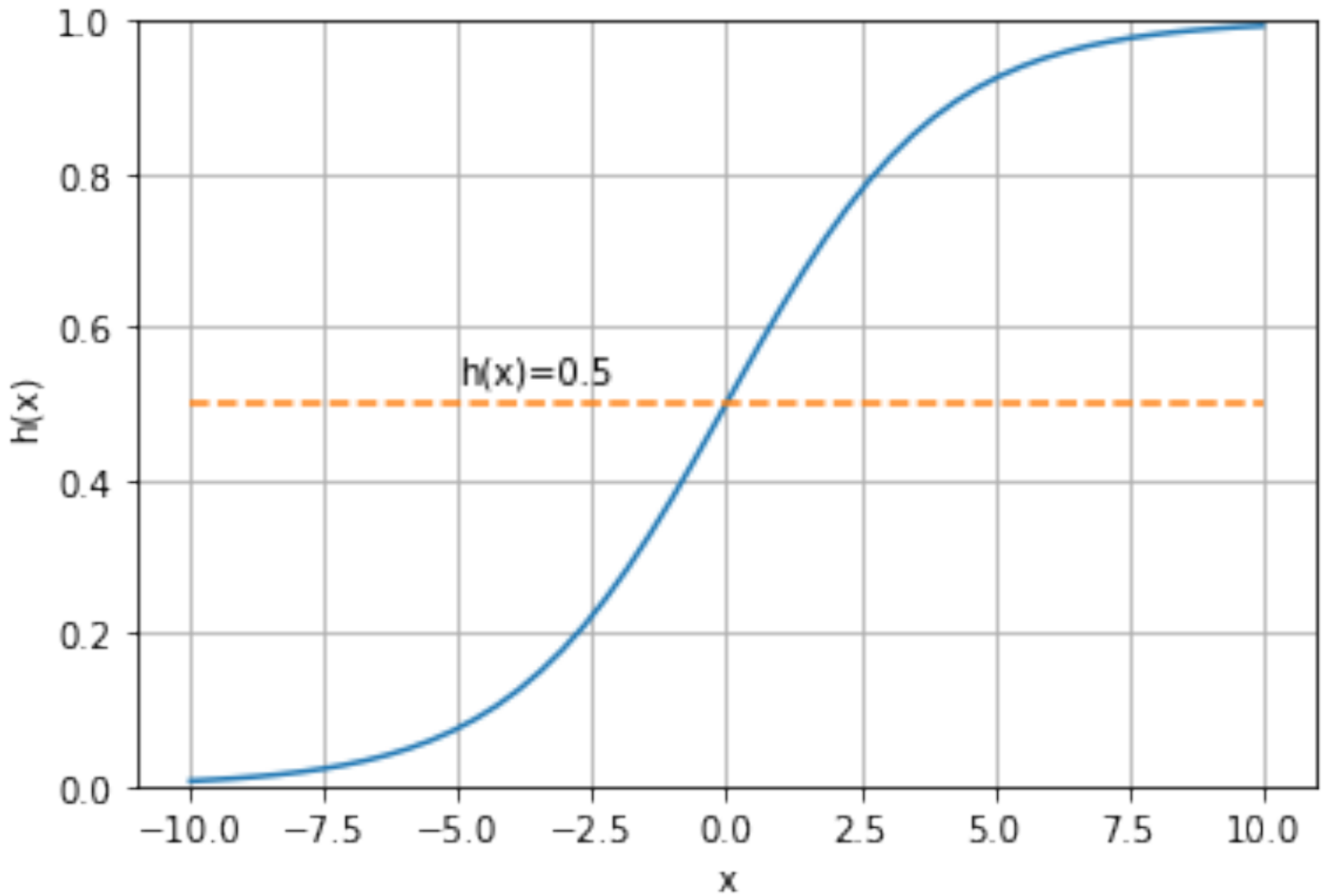
$h(x) \geq 0.5 \Rightarrow$ класс 1

DECISION BOUNDARY

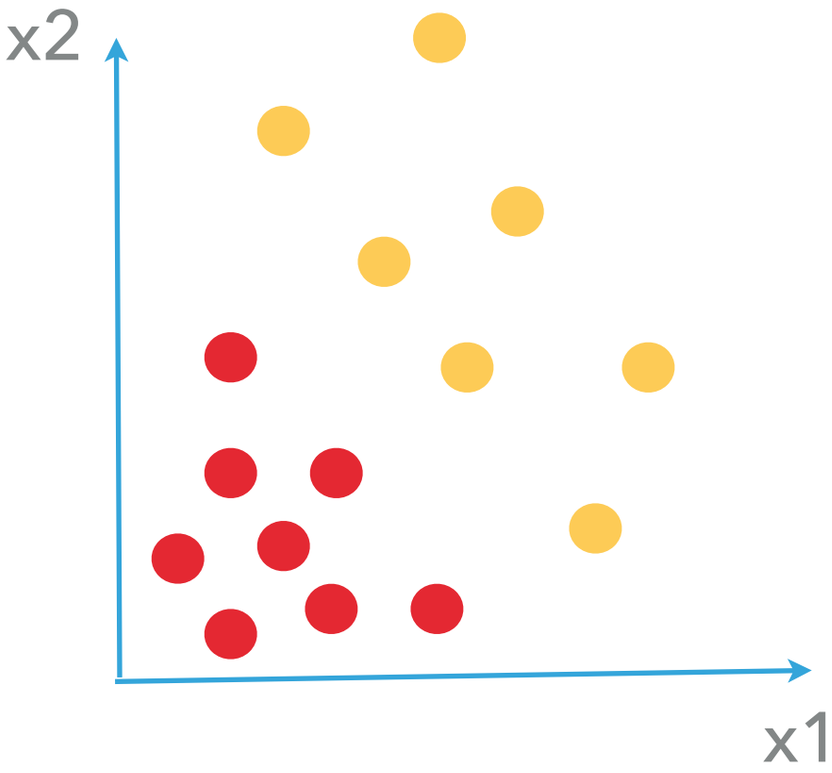
Иными словами:

$x < 0 \Rightarrow$ класс 0

$x >= 0 \Rightarrow$ класс 1



DECISION BOUNDARY



мы подбираем такие коэффициенты w

чтобы прямая $w^T x = 0$

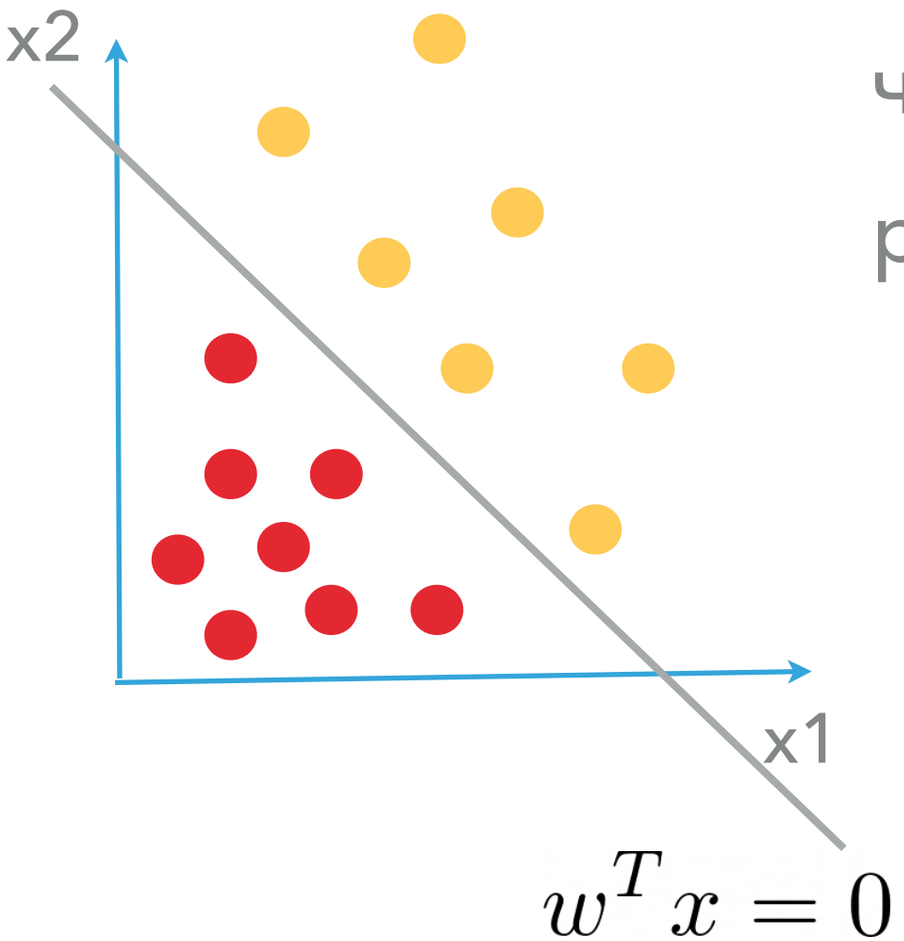
разделила классы

DECISION BOUNDARY

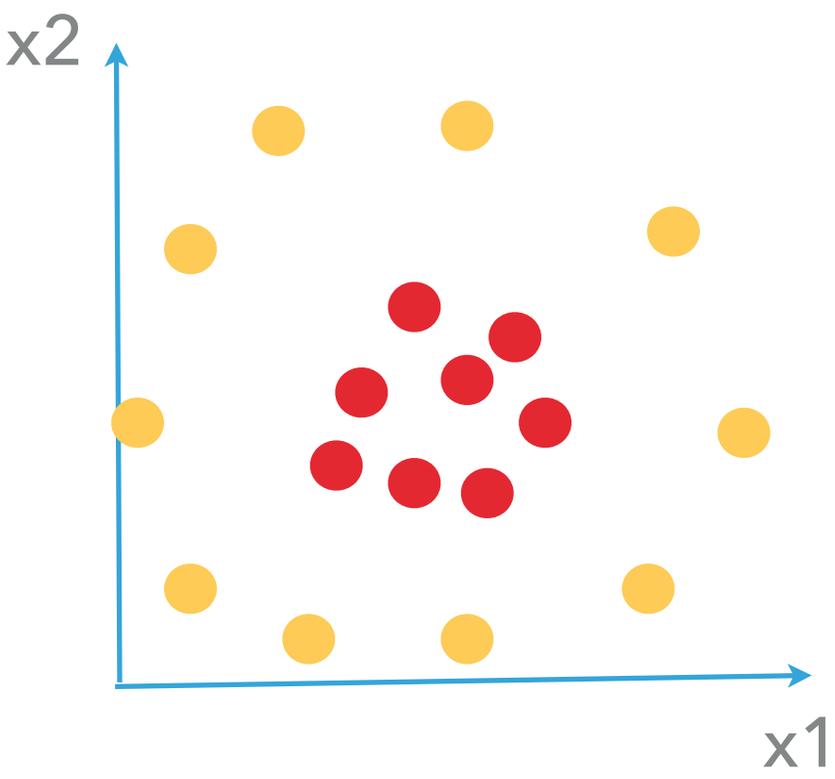
мы подбираем такие коэффициенты w

чтобы прямая $w^T x = 0$

разделила классы

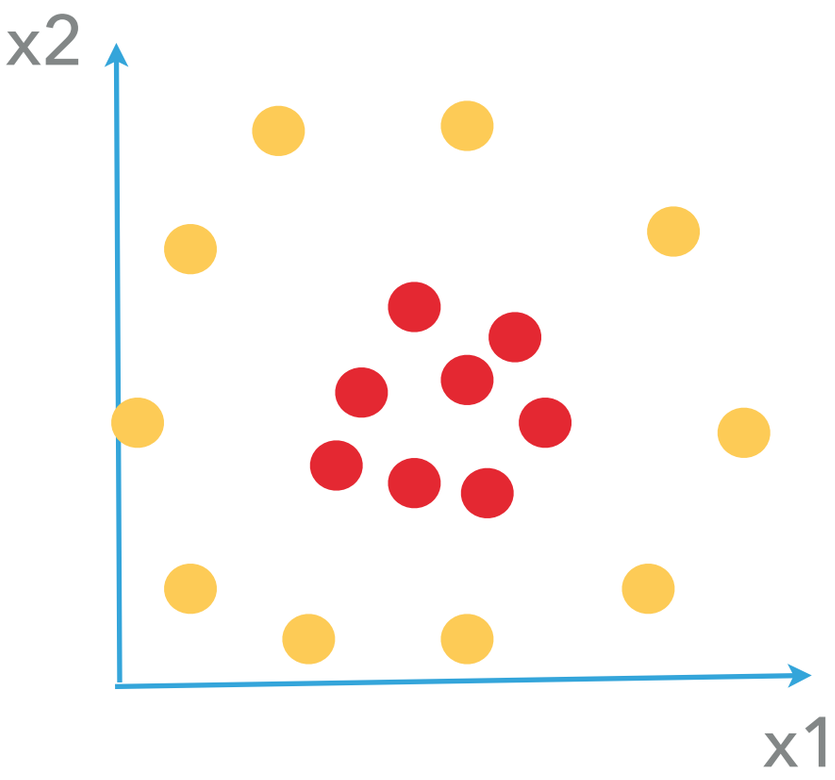


DECISION BOUNDARY



а такое сможем разделить
логистической регрессией?

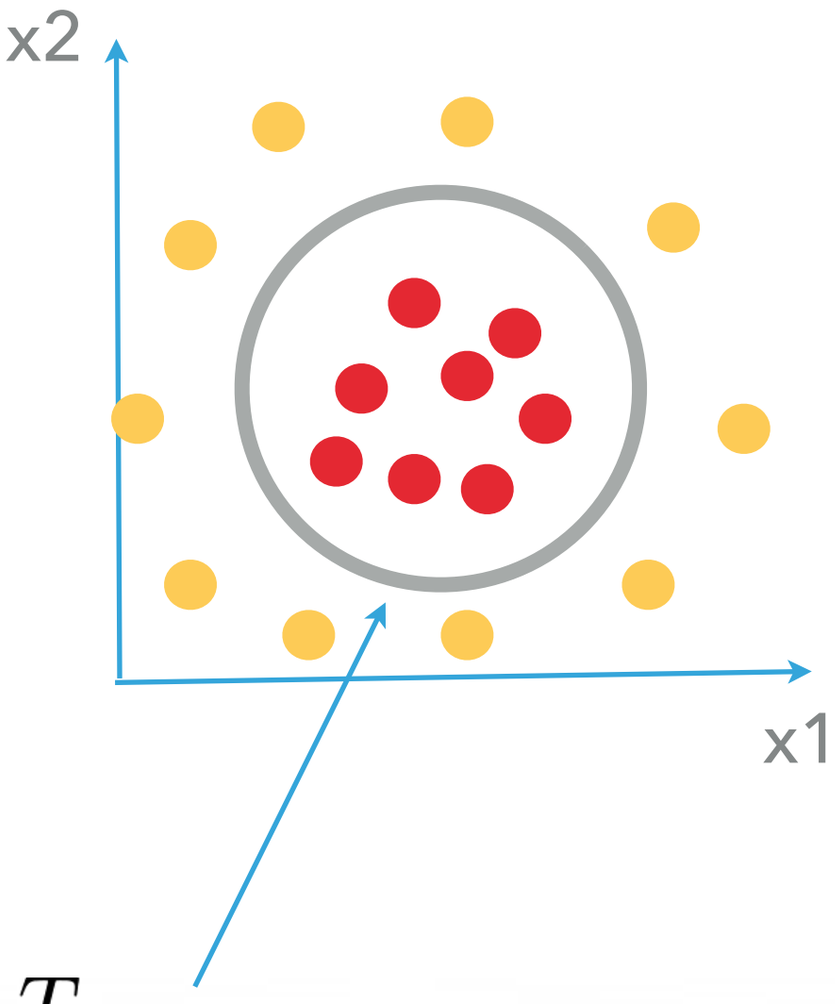
DECISION BOUNDARY



а такое сможем разделить
логистической регрессией?

сможем, если сами добавим
нелинейности от X

DECISION BOUNDARY



а такое сможем разделить
логистической регрессией?

сможем, если сами добавим
нелинейности от X

$$w^T x = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 = 0$$

ВЫБОР ПОРОГОВОГО ЗНАЧЕНИЯ (THRESHOLD)

- Для простоты рассматривали порог, равный 0.5
- Но такой порог не всегда имеет смысл:
  - Могут быть разные размеры классов
  - Разная цена ошибки между классами
- Поэтому это дополнительный параметр, который выбираем исходя из метрик на валидационном множестве (см. также ROC-кривые)

СВОЙСТВА

- Новый класс моделей, работающих сразу и для регрессии, и для классификации
- Хорошая интерпретируемость результатов
- Не будет градиентного спуска
- С категориями и ненормированными признаками работают "из коробки"
- С многоклассовой классификацией работают "из коробки"
- Служат основной для более сложных моделей с нелинейными decision boundary

ДЕРЕВЬЯ ПРИНЯТИЯ РЕШЕНИЙ

- Рассмотрим такой датасет:

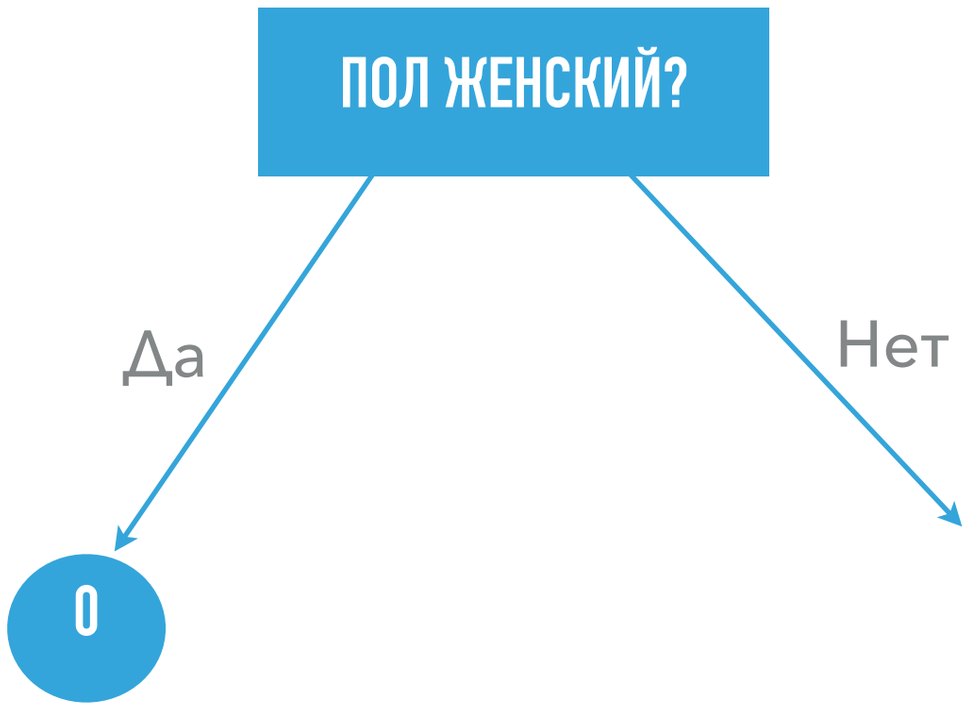
Пол	Средний балл	Пойдет в армию?
ж	4.3	0
ж	3.0	0
м	4.7	0
м	2.3	1
м	2.4	1
м	3.5	0

ДЕРЕВЬЯ ПРИНЯТИЯ РЕШЕНИЙ

- Будем строить модель, последовательно задавая вопросы:

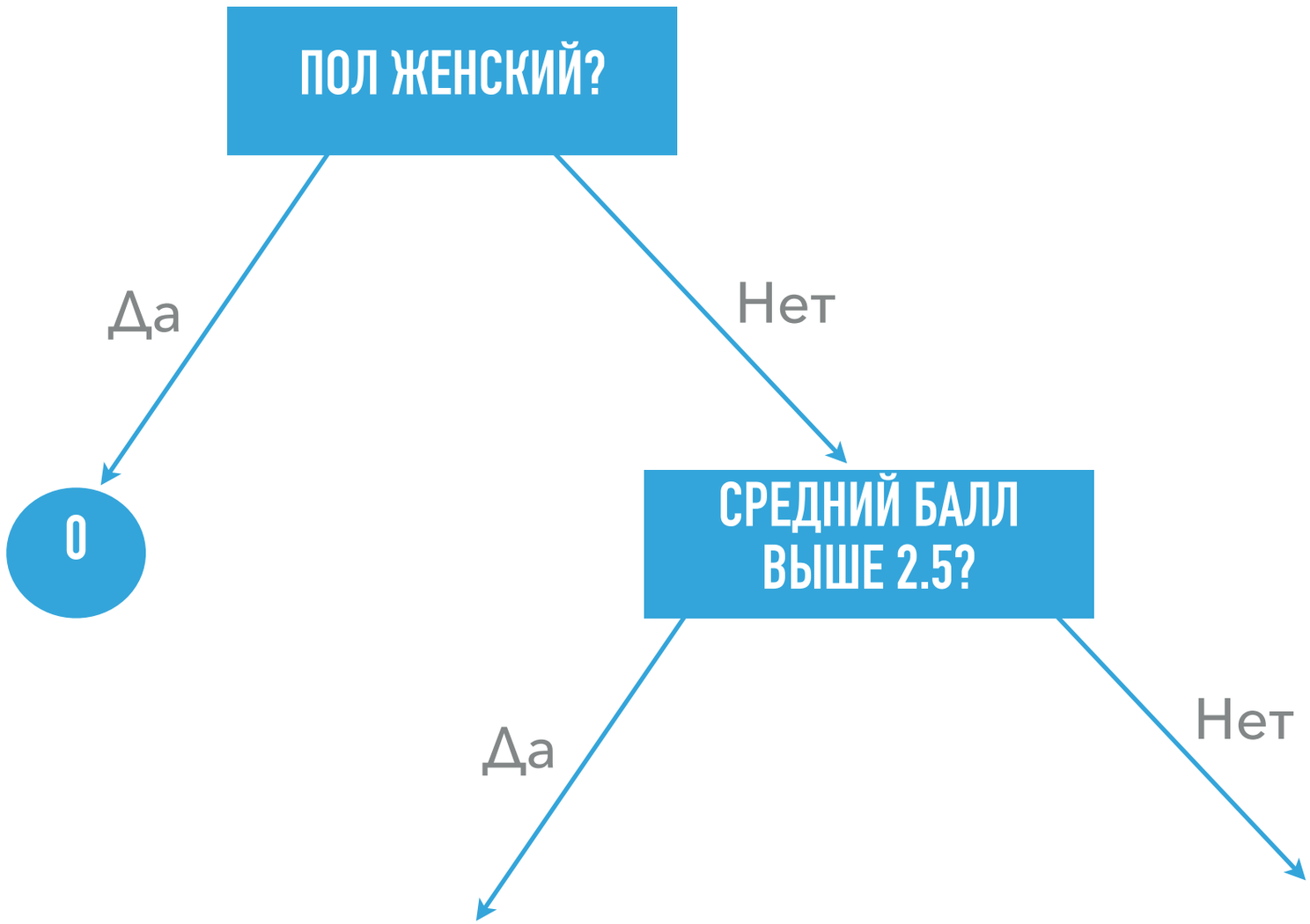
ДЕРЕВЬЯ ПРИНЯТИЯ РЕШЕНИЙ

- Будем строить модель, последовательно задавая вопросы:



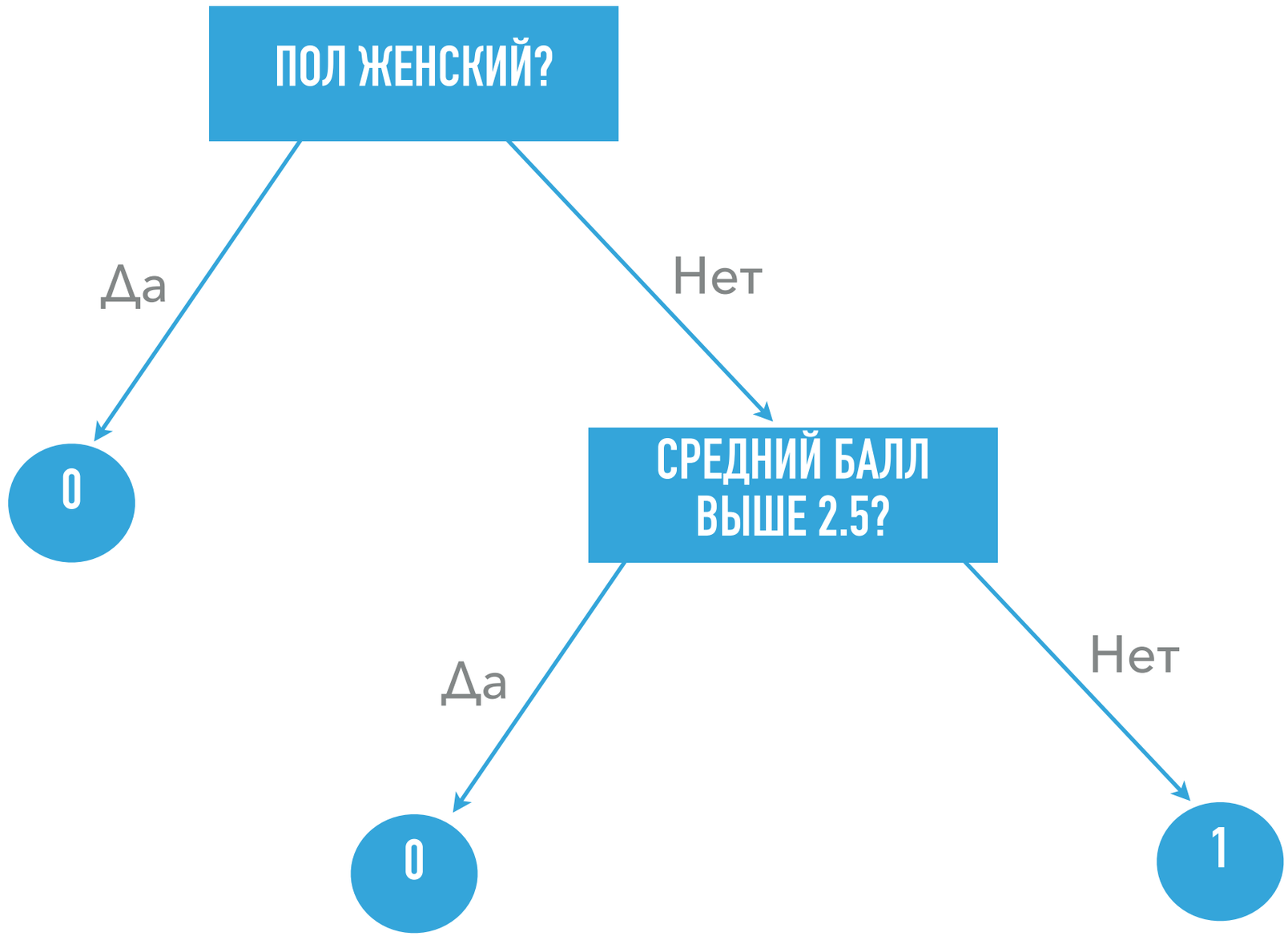
ДЕРЕВЬЯ ПРИНЯТИЯ РЕШЕНИЙ

- Будем строить модель, последовательно задавая вопросы:



ДЕРЕВЬЯ ПРИНЯТИЯ РЕШЕНИЙ

- Будем строить модель, последовательно задавая вопросы:



ДЕРЕВЬЯ ПРИНЯТИЯ РЕШЕНИЙ

- То есть строим бинарное дерево
- В листьях дерева у нас итоговые классы для задачи классификации. Для задачи регрессии берем в листья среднее значение целевого признака семлов что в этот лист попали
- Если признак категориальный, то вопрос ставится в том какое значение он принимает
- Если признак численный, то вопрос ставится в том больше (или меньше) ли он чем какое-то значение

ОТКРЫТЫЕ ВОПРОСЫ

- В каком порядке выбирать признаки?
- Выбрав признак, по какому значению его проверять на больше/меньше (в случае чисел) или по какому значению категорий
- Что с переобучением?

В КАКОМ ПОРЯДКЕ ВЫБИРАТЬ ПРИЗНАКИ

- Допустим, что в предыдущем датасете у нас 1000 девочек (никто не будет служить) и 200 мальчиков (из них будет служить половина)

В КАКОМ ПОРЯДКЕ ВЫБИРАТЬ ПРИЗНАКИ

- Допустим, что в предыдущем датасете у нас 1000 девочек (никто не будет служить) и 200 мальчиков (из них будет служить половина)

Какой признак возьмем сначала?

В КАКОМ ПОРЯДКЕ ВЫБИРАТЬ ПРИЗНАКИ

- Допустим, что в предыдущем датасете у нас 1000 девочек (никто не будет служить) и 200 мальчиков (из них будет служить половина)

Какой признак возьмем сначала?

Возьмем сначала "пол", т.к. мы в этом случае сделаем данные по разным сторонам узла более однородными

В КАКОМ ПОРЯДКЕ ВЫБИРАТЬ ПРИЗНАКИ

- Допустим, что в предыдущем датасете у нас 1000 девочек (никто не будет служить) и 200 мальчиков (из них будет служить половина)

  Какой признак возьмем сначала?

  Возьмем сначала "пол", т.к. мы в этом случае сделаем данные по разным сторонам узла более однородными

- Формальные критерии однородности: индекс Джини, энтропия/количество информации/deviance

ИНДЕКС ДЖИНИ (GINI INDEX)

- Пусть у нас есть k классов
- Мы находимся в некотором узле, в котором частоты семплов по разным классам такие:

$$(p_1, \ldots, p_k)$$

ИНДЕКС ДЖИНИ (GINI INDEX)

- Пусть у нас есть k классов
- Мы находимся в некотором узле, в котором частоты семплов по разным классам такие:

$$(p_1, \ldots, p_k)$$

- Критерий такой:

$$GINI = \sum_{j,j' \in \{1,\ldots,k\}: j \neq j'} p_j p_{j'} = 1 - \sum_{1}^{k} p_j^2$$

ИНДЕКС ДЖИНИ (GINI INDEX)

- Мы находимся в некотором узле, в котором частоты семплов по разным классам такие:

$$(p_1, \ldots, p_k)$$

- Критерий такой:

$$GINI = \sum_{j,j' \in \{1,\ldots,k\}: j \neq j'} p_j p_{j'} = 1 - \sum_{1}^{k} p_j^2$$

- Максимален когда классы равномерны
- Минимален когда все семплы лежат в одном классе

ИНДЕКС ДЖИНИ (GINI INDEX)

- Мы можем посчитать текущий индекс Джини
- Можем перебрать все признаки и все потенциальные разбиения и посчитать каким будет индекс Джини в нодах ниже
- Выбираем то разбиение, которое в наибольшей степени минимизирует индекс (можно выписать не очень сложную формулу, но не будем)

ИНДЕКС ДЖИНИ (GINI INDEX)

- Мы можем посчитать текущий индекс Джини
- Можем перебрать все признаки и все потенциальные разбиения и посчитать каким будет индекс Джини в нодах ниже
- Выбираем то разбиение, которое в наибольшей степени минимизирует индекс (можно выписать не очень сложную формулу, но не будем)
- Другие критерии работают аналогично

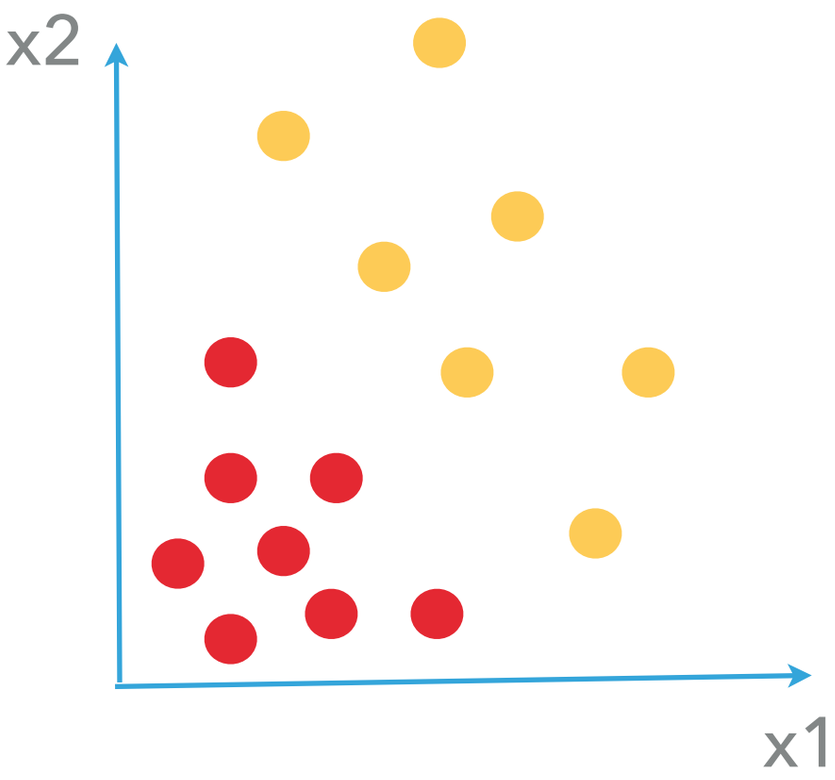
ИНДЕКС ДЖИНИ (GINI INDEX)

- Деревья могут идеально описать любой training set => очень склонны к переобучению :(

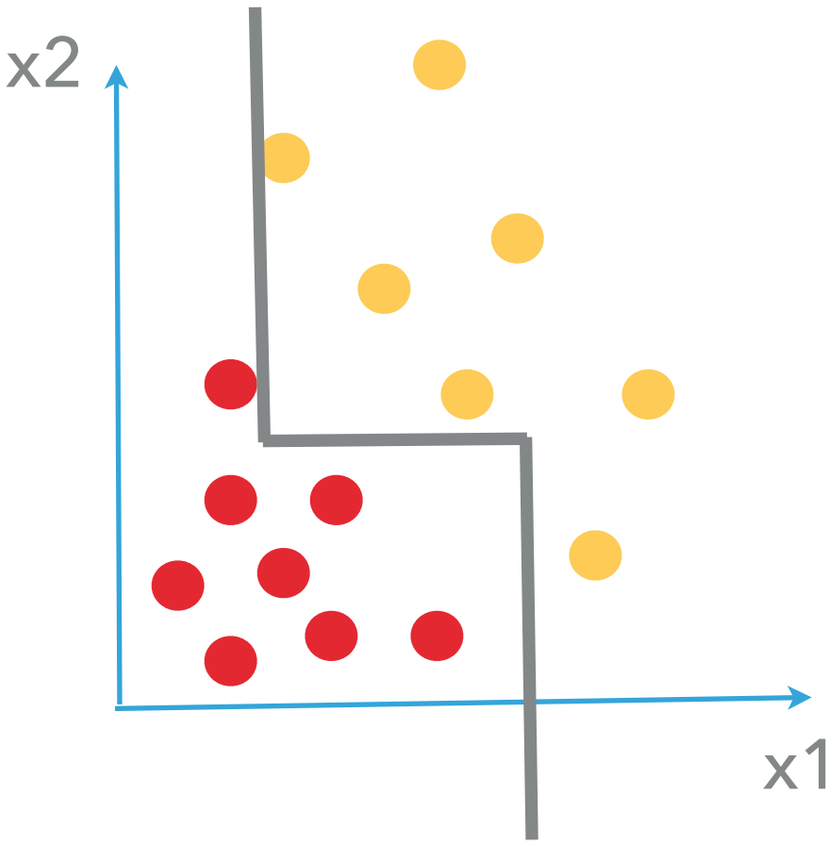
ИНДЕКС ДЖИНИ (GINI INDEX)

- Деревья могут идеально описать любой training set => очень склонны к переобучению :(
- Но как узнаем далее, на самом деле нам все равно на переобучение деревьев

DECISION BOUNDARY



DECISION BOUNDARY



В отличие от регрессии, где проводили прямую линию, здесь строим блоки

КАК ОПРЕДЕЛИТЬ ВЕС БЫКА?

- Есть бык
- Есть толпа обывателей. Оценка толпы - среднее их независимых прогнозов
- Есть несколько экспертов по быкам

КАК ОПРЕДЕЛИТЬ ВЕС БЫКА?

- Есть бык
- Есть толпа обывателей. Оценка толпы - среднее их независимых прогнозов
- Есть несколько экспертов по быкам

Чья оценка веса будет точнее?

КАК ОПРЕДЕЛИТЬ ВЕС БЫКА?

- Есть бык
- Есть толпа обывателей. Оценка толпы - среднее их независимых прогнозов
- Есть несколько экспертов по быкам

Чья оценка веса будет точнее?
По слухам толпа будет лучше

ПРИЧЕМ ЗДЕСЬ MACHINE LEARNING?

Оказывается, идея взять кучу слабых классификаторов и объединить их оценки работает очень хорошо

Называется подход "ансамблирование"

ОСНОВНЫЕ ПОДХОДЫ К АНСАМБЛИРОВАНИЮ

- Bagging - берем одинаковые классификаторы, обучаем независимо на разных подвыборках, результаты усредняем

- Boosting - обучаем одинаковые классификаторы последовательно, где каждый последующий акцентирует внимание на ошибках предыдущего

- Stacking - обучаем любые модели независимо, а потом обучаем еще одну мета-модель, которая берет результаты моделей на вход и предсказывает итоговый результат

ДЕРЕВЬЯ

- Деревья - очень хороший кандидат на "слабую модель":
  - Быстро обучаются
  - Легко переобучаются
  - Легко получать отличающиеся модели с помощью них

ДЕРЕВЬЯ

- Деревья - очень хороший кандидат на "слабую модель":
  - Быстро обучаются
  - Легко переобучаются
  - Легко получать отличающиеся модели с помощью них
- Одна из самых популярных моделей ансамбля - случайный лес

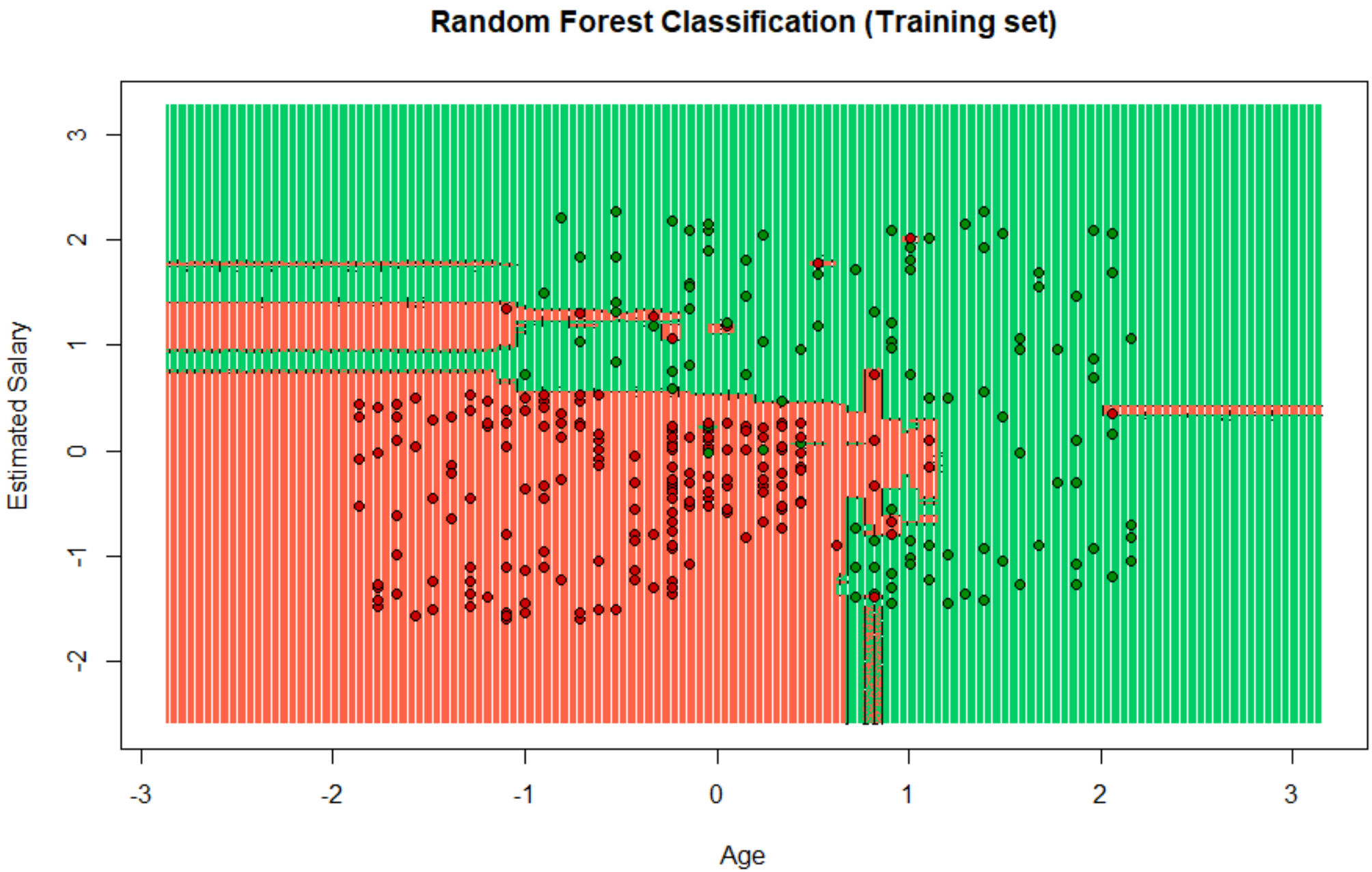
СЛУЧАЙНЫЙ ЛЕС

- Случайный лес он скорее относится к Bagging, но имеет нюансы за счет специфики деревьев
- Обучение случайного леса:
  - Выбираем число K - количество деревьев
  - Для каждого дерева проводим процедуру:
    - Делаем семплирование датасета с повторами (bootstrapping)
    - Фиксируем случайным образом часть фичей
    - Обучаем на семпле данных и этом наборе фичей дерево

СЛУЧАЙНЫЙ ЛЕС

- Случайный лес он скорее относится к Bagging, но имеет нюансы за счет специфики деревьев
- Обучение случайного леса:
  - ....
- Благодаря такому подходу получаем действительно разные модели
- Предсказание случайным лесом:
  - Опросили каждое дерево, взяли среднее в задаче регрессии или самый частый класс в задаче классификации

DECISION BOUNDARY



https://medium.com/analytics-vidhya/random-forest-a-model-designed-to-provide-structure-in-chaos-e267d559ca04

БУСТИНГИ

- Самый модный (или был до недавнего времени) подход - Gradient Boosting
- Основная реализация - XGBoost
- Сложные и тяжелые модели, являющиеся победителями большинства соревнований kaggle
- В реальной жизни раньше использовались редко, т.к. сложно настраивать, сложно интерпретировать, да и вообще, но сейчас уже начинают требовать везде