

Логика ML и новые алгоритмы

Decision Tree и Random Forest

Политология, 3 курс
Татьяна Рогович


Решающие деревья и случайный лес



Какие бывают алгоритмы?

Алгоритмов ML достаточно много. Причем у многих есть вариации, которые позволяют решать и задачи регрессии, и классификации. Такие узконаправленные специалисты, как логистическая регрессия, скорее исключения.

Поговорим про два алгоритма, с помощью которых можно решать очень много задач (и решать достаточно неплохо!)



Решающее дерево

Решающее дерево (Decision tree) — решение задачи обучения с учителем, основанный на том, как решает задачи прогнозирования человек.

В общем случае — это k -ичное дерево с *решающими правилами* в нелистовых вершинах (узлах) и некотором заключении о целевой функции в листовых вершинах (*прогнозом*).

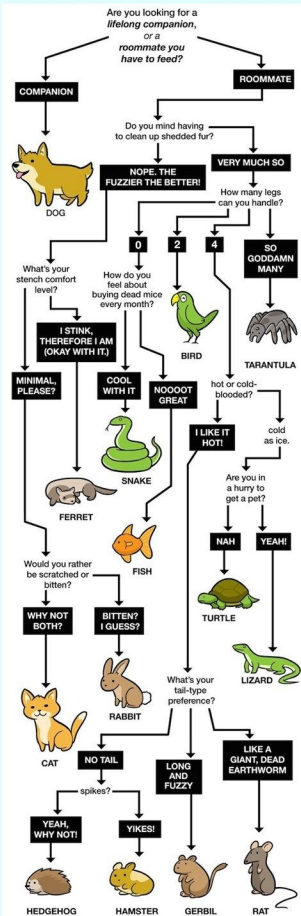
Решающее правило — некоторая функция от объекта, позволяющая определить, в какую из дочерних вершин нужно поместить рассматриваемый объект.

В листовых вершинах могут находиться разные объекты: класс, который нужно присвоить попавшему туда объекту (в задаче классификации), вероятности классов (в задаче классификации), непосредственно значение целевой функции (задача регрессии).

К какой форме отнести публикацию?



Which Pet Should You Get?

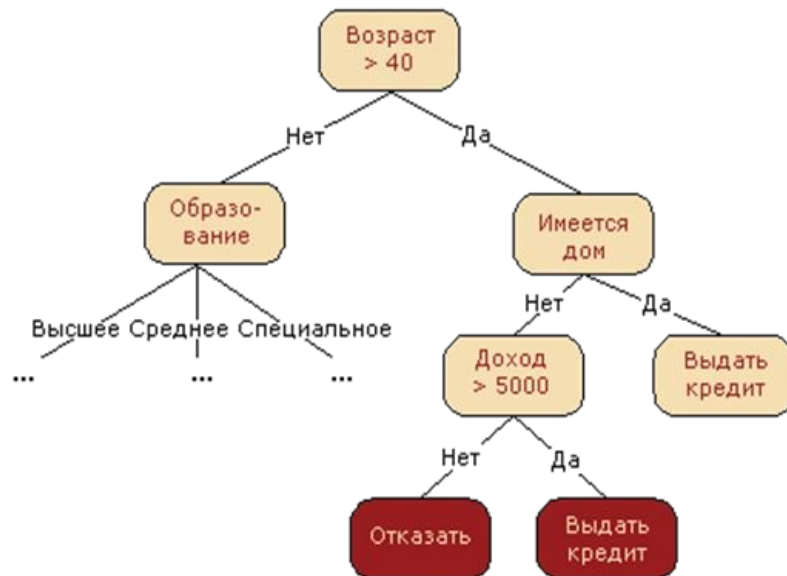


bellygifs.com

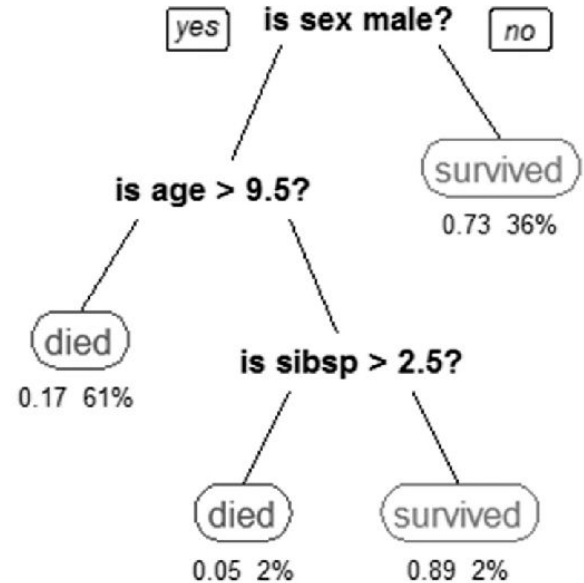
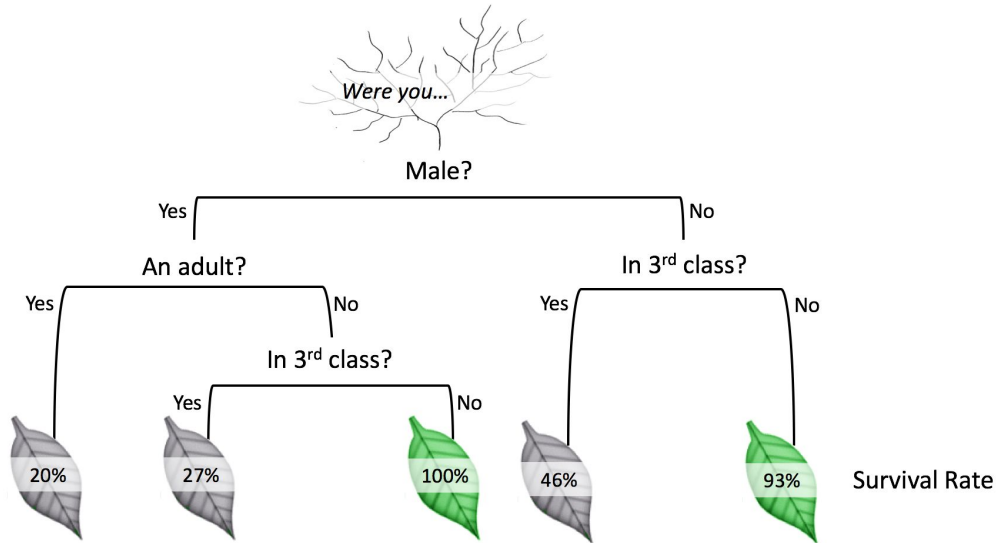
Персонаж подходит : Майкл Фарадей

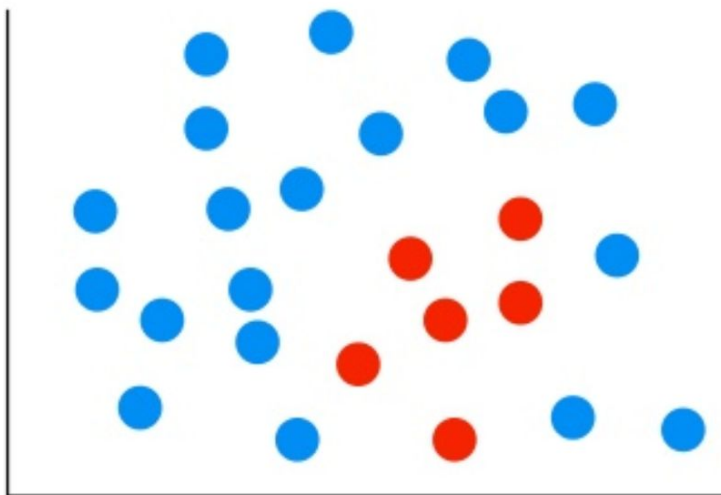
Вопрос	Ответ дан	Ответ ожидается
Ваш персонаж существовал в реальности?	Да	Да
Ваш персонаж мужчина?	Да	Да
Ваш персонаж живет в России?	Нет	Нет
Ваш персонаж известен в кино?	Нет	Нет
Ваш персонаж старше 40 лет?	Да	Да
Ваш персонаж политик?	Нет	Нет
Ваш персонаж умер?	Да	Да
Ваш персонаж связан с миром музыки?	Нет	Нет
Ваш персонаж ученый?	Да	Да
Ваш персонаж - профессор?	Я не знаю	Нет
У вашего персонажа немецкая фамилия?	Нет	Нет
Ваш персонаж говорит по-английски?	Да	Да
Ваш персонаж связан с компьютерами?	Нет	Нет
Ваш персонаж имеет отношение к физике?	Да	Да
У этого персонажа белые волосы?	Да	Я не знаю
У вашего персонажа есть усы?	Нет	Нет
Когда мы думаем о вашем персонаже, мы ассоциируем его с яблоком?	Нет	Нет
Ваш персонаж русский?	Нет	Нет
Ваш персонаж гражданин Соединенных Штатов?	Нет	Нет
Ваш персонаж жил в двадцатом веке?	Нет	Нет
Ваш персонаж имеет работу, связанную с искусством?	Нет	Нет
Фамилия вашего персонажа состоит из 2-х слогов?	Нет	Нет
Ваш персонаж связан с электричеством?	Да	Да
Ваш персонаж имеет форму шара?	Нет	NONE
Есть ли волосы у вашего персонажа?	Да	NONE

Пример решающего дерева



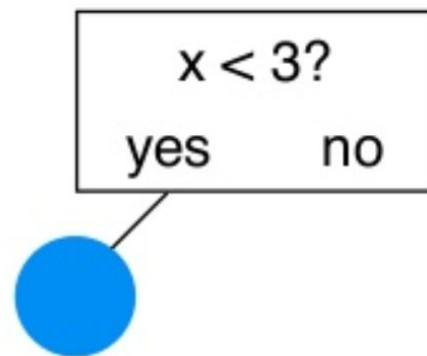
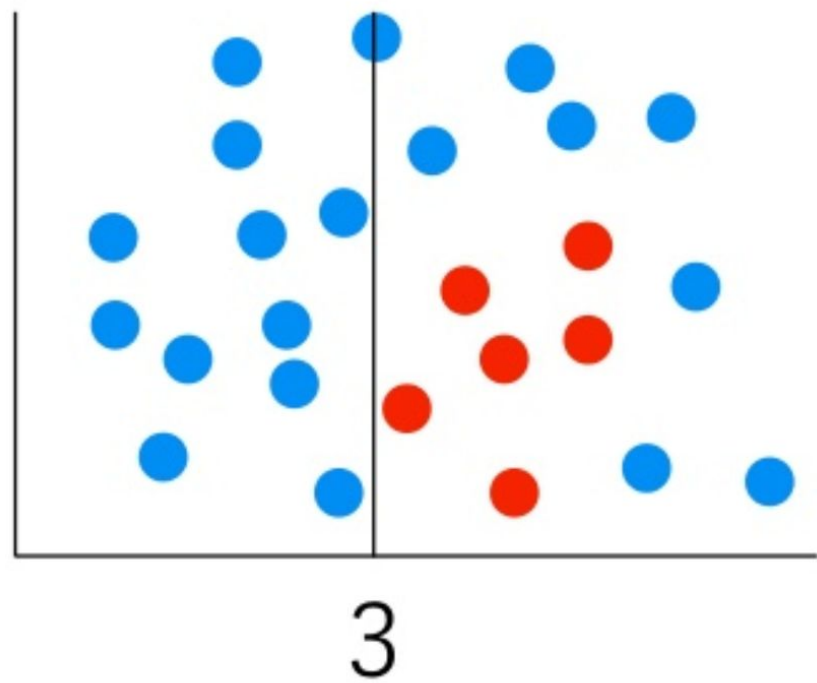
И для Титаника (куда же без него)

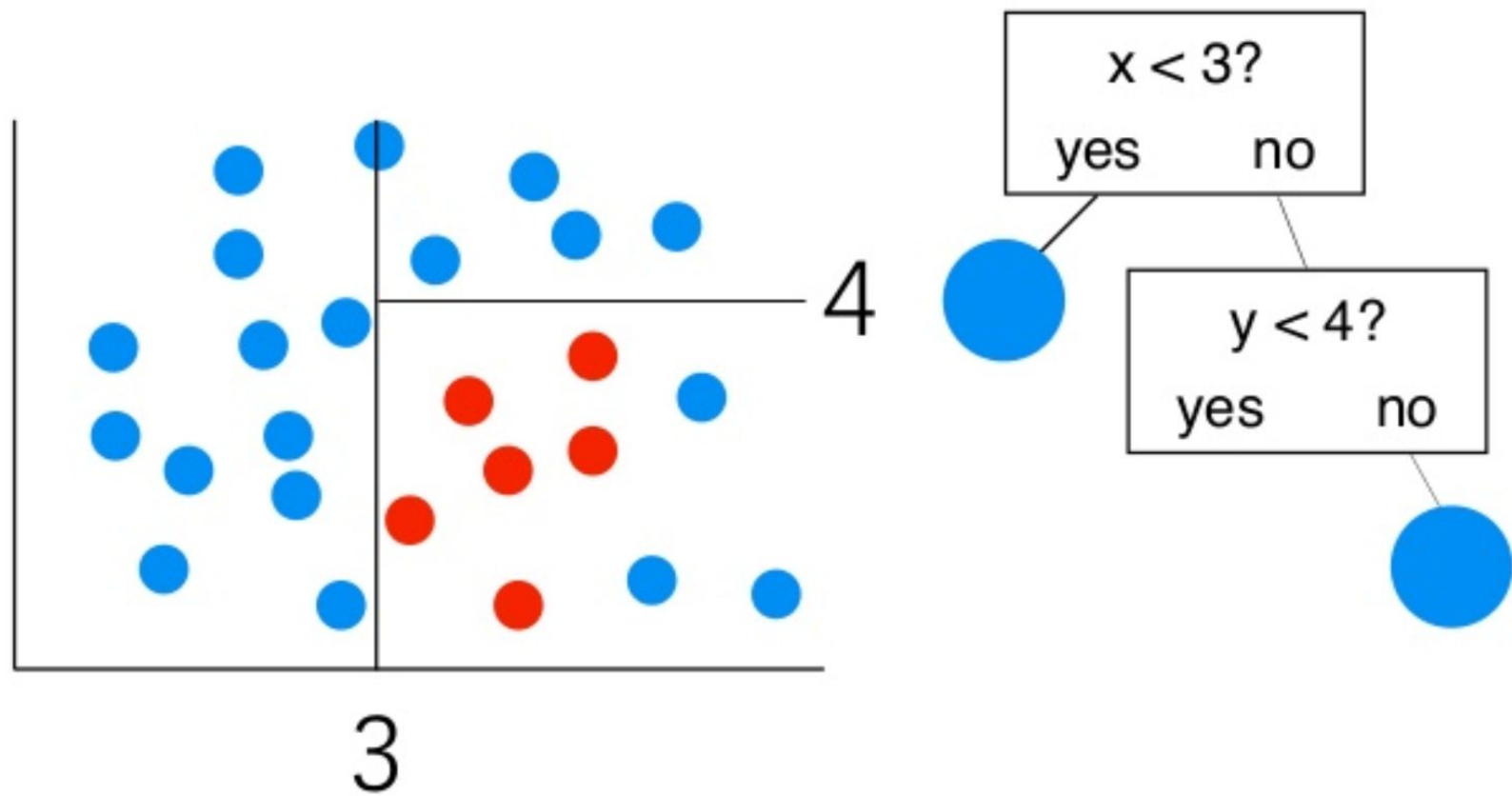


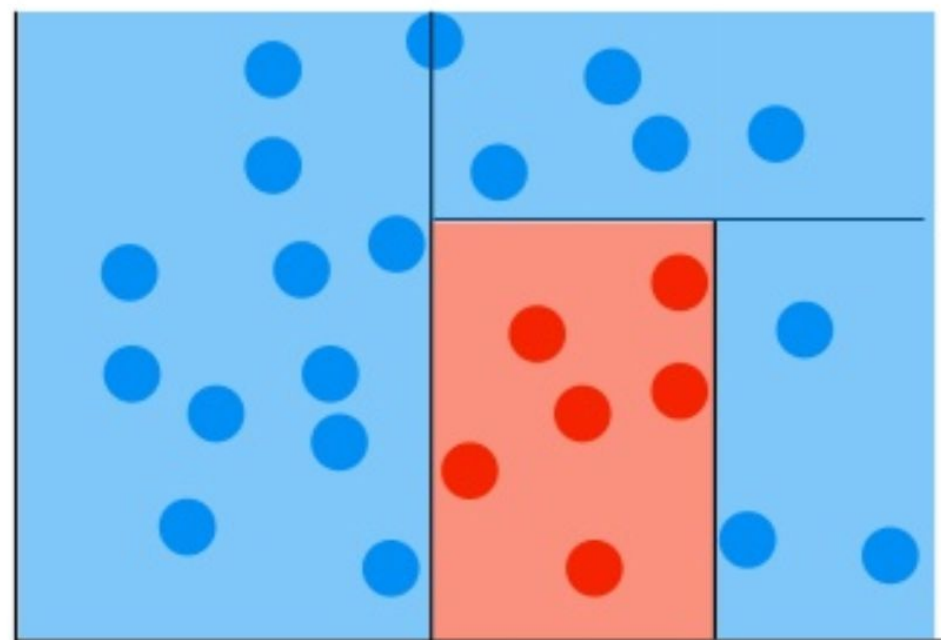


Нам нужно нарисовать такие линии, которые лучше всего изолируют красные точки от синих.

Давайте начнем со шкалы X. Где нужно нарисовать линию, чтобы слева и справа от нее состав точек был максимально однородным (насколько это возможно)?







3

5

4

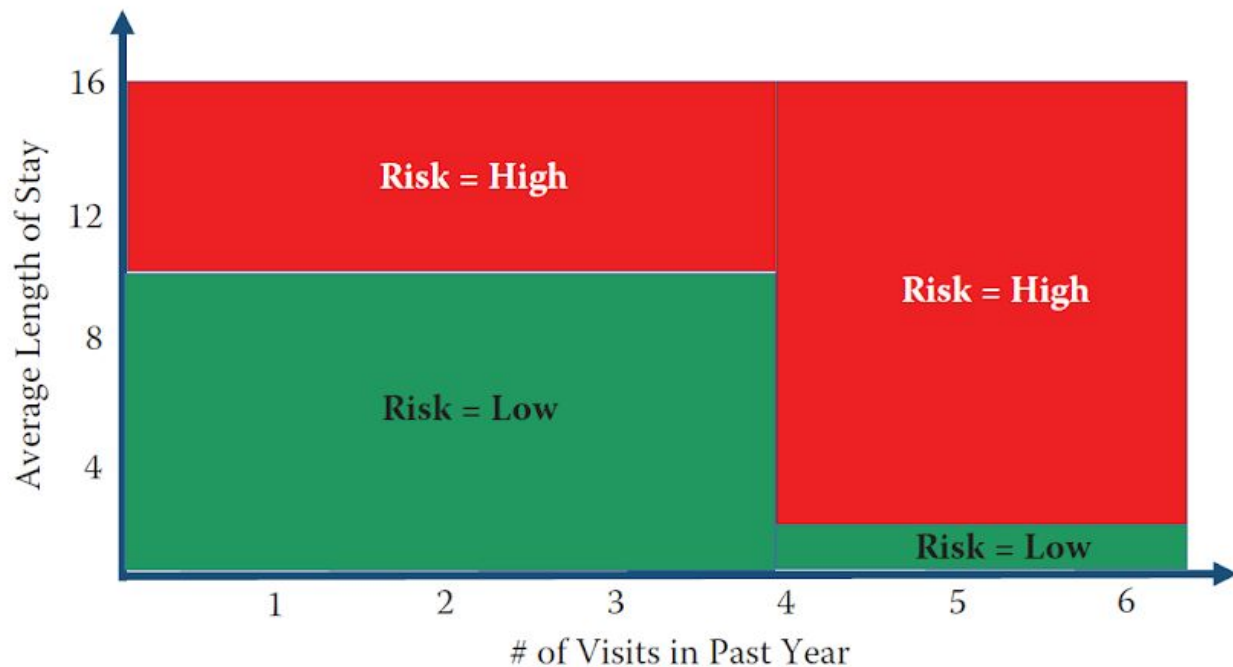
$x < 3?$
yes no

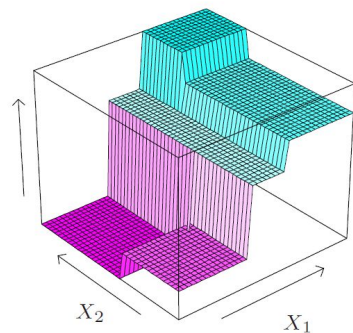
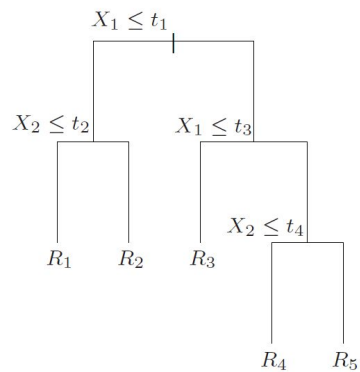
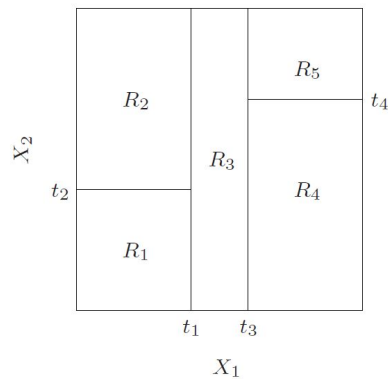
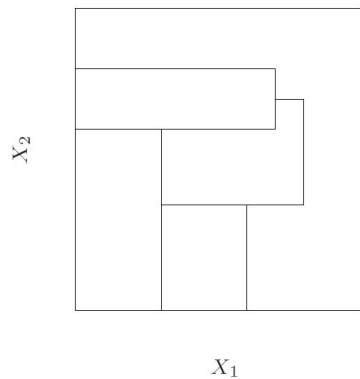
$y < 4?$
yes no

$x < 5?$
yes no



В реальной жизни



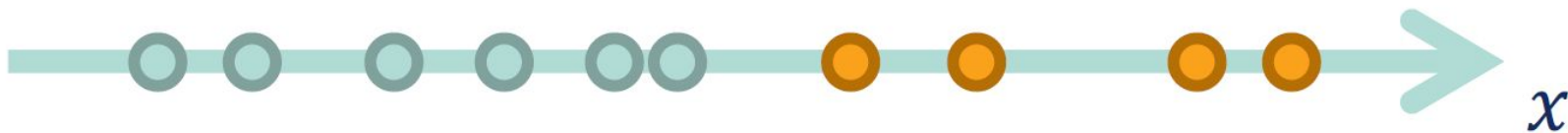


**Как дерево выглядит
в пространстве
признаков**

Давайте попробуем понять, как это работает

Рассмотрим выборку объектов с одним признаком x :

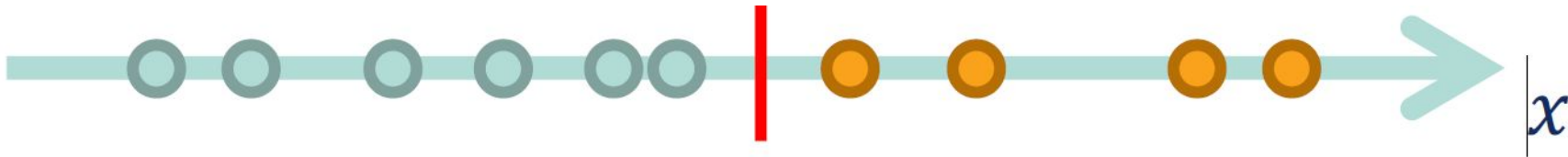
Как подобрать порог по признаку в задаче бинарной классификации?



Давайте попробуем понять, как это работает

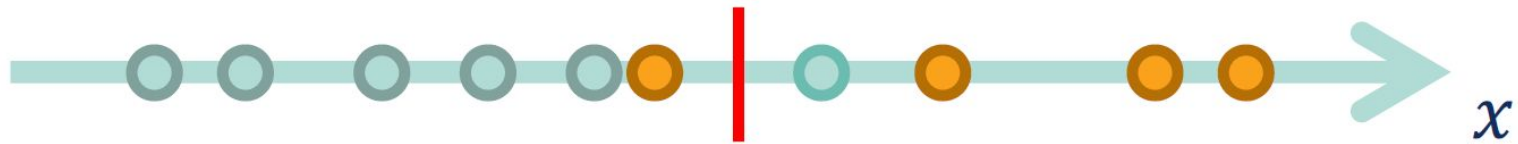
Рассмотрим выборку объектов с одним признаком x :

Как подобрать порог по признаку в задаче бинарной классификации?

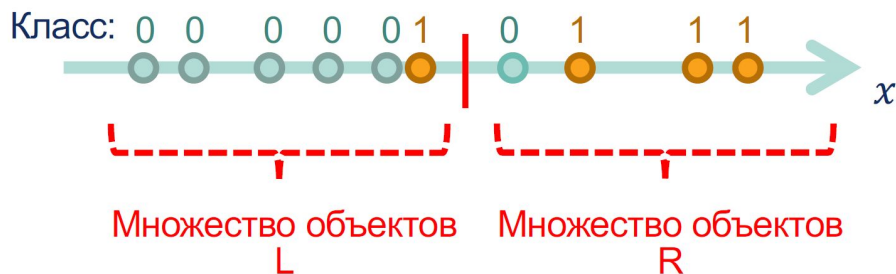




А что делать тут?



А что делать тут?



Чтобы разделить классы хорошо – нужно, чтобы и в L и в R преобладал только один класс.

$$S = - \sum_{i=1}^N p_i \log_2 p_i$$

Энтропия – метрика хаоса. => уменьшение энтропии – это прирост информации



Алгоритм построения дерева решений

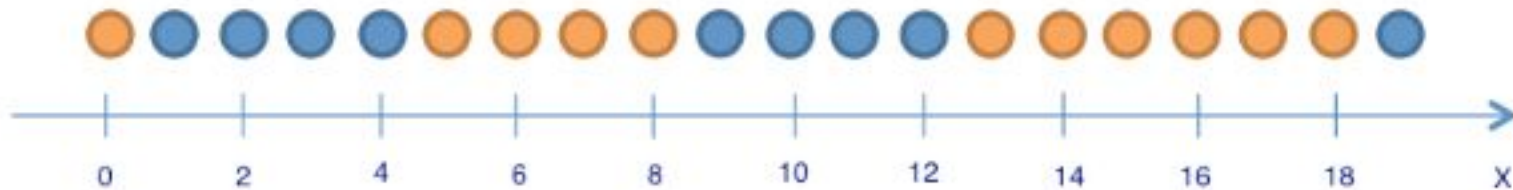
s_0 = вычисляем энтропию исходного множества

- Если $s_0 == 0$ значит:
все объекты исходного набора, принадлежат к одному классу
сохраняем этот класс в качестве листа дерева
- Если $s_0 \neq 0$ значит:
ищем предикат, который разбивает исходное множество таким образом
чтобы уменьшилось среднее значение энтропии
- найденный предикат является частью дерева принятия решений, сохраняем его

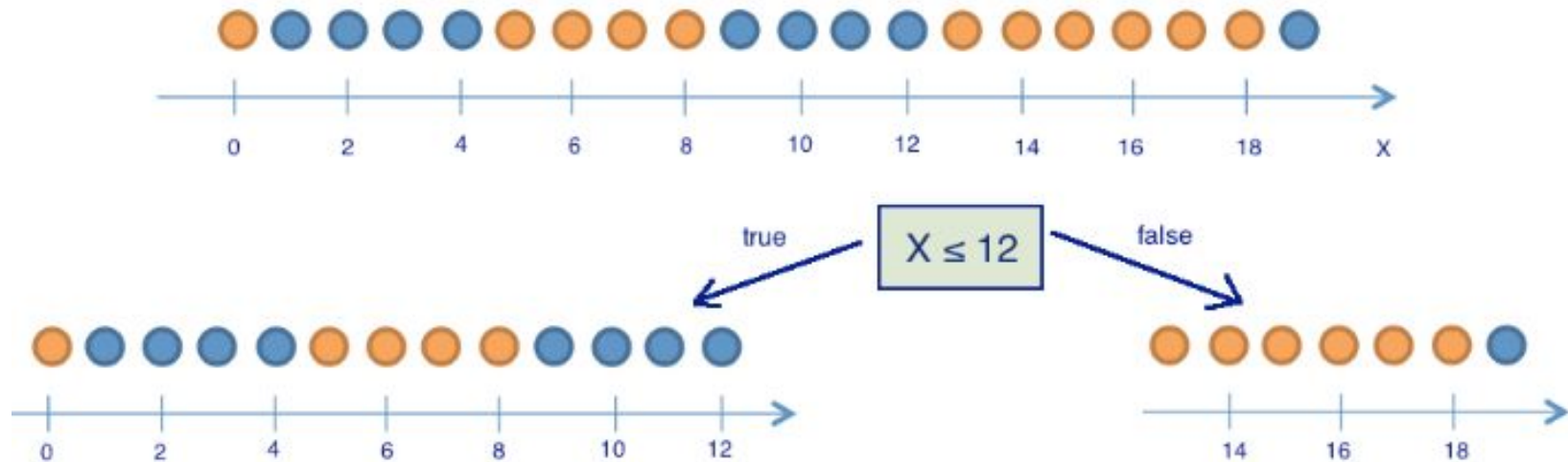
Разбиваем исходное множество на подмножества, согласно предикату
Повторяем данную процедуру рекурсивно для каждого подмножества



Потренируемся



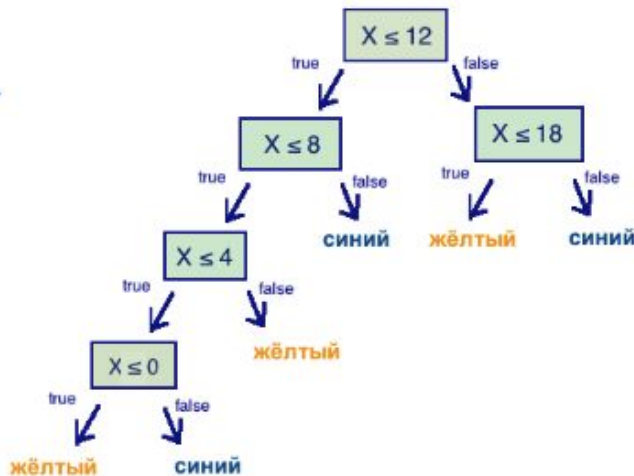
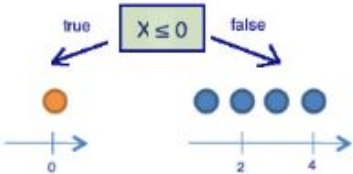
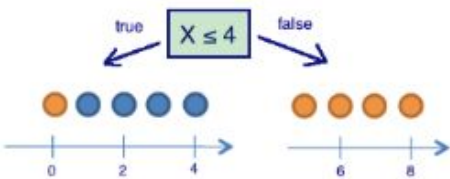
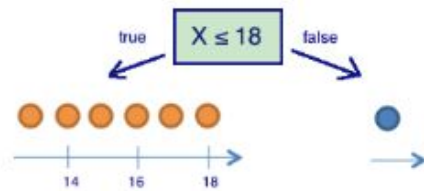
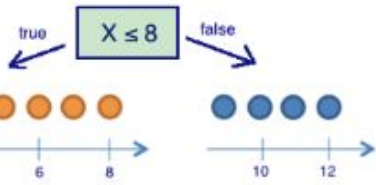
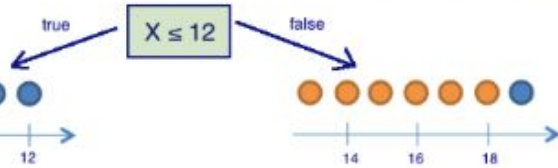
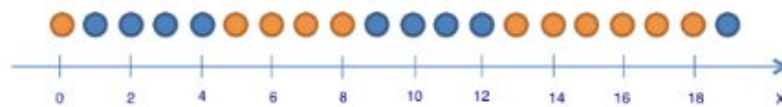
Потренируемся

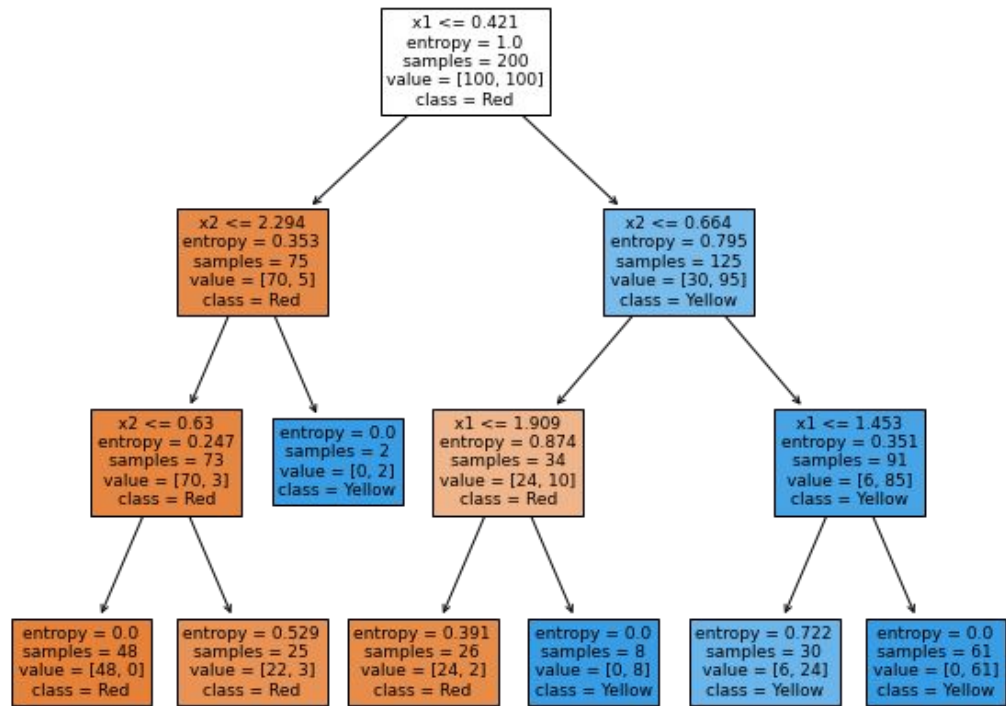
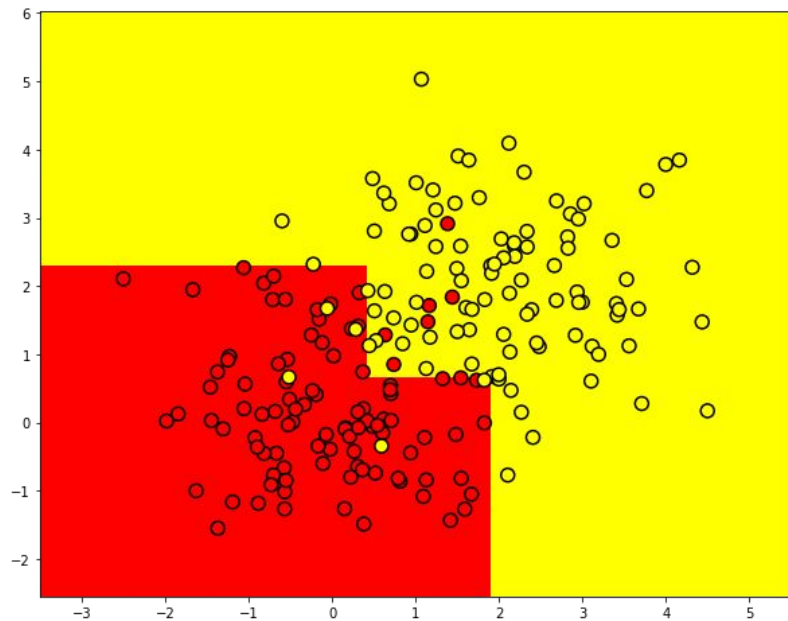





Прирост информации (information gain)

$$IG(Q) = S_O - \sum_{i=1}^q \frac{N_i}{N} S_i,$$








Решающее дерево и количественный признак

	Возраст	Невозврат кредита
0	17	1
1	64	0
2	18	1
3	20	0
4	38	1
5	49	0
6	55	0
7	25	1
8	29	1
9	31	0
10	33	1



Решающее дерево и количественный признак

	Возраст	Невозврат кредита
0	17	1
2	18	1
3	20	0
7	25	1
8	29	1
9	31	0
10	33	1
4	38	1
5	49	0
6	55	0
1	64	0



Настройка решающего дерева

Основные параметры класса `sklearn.tree`

- `max_depth` – максимальная глубина дерева
- `max_features` — максимальное число признаков, по которым ищется лучшее разбиение в дереве (это нужно потому, что при большом количестве признаков будет "дорого" искать лучшее (по критерию типа прироста информации) разбиение среди *всех* признаков)
- `min_samples_leaf` – минимальное число объектов в листе. У этого параметра есть понятная интерпретация: скажем, если он равен 5, то дерево будет порождать только те классифицирующие правила, которые верны как минимум для 5 объектов



Немного интерактива

<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>



Случайный лес

RF (random forest) — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству.

Один из немногих универсальных алгоритмов: задач, на который случайный лес плохо бы себя показал — меньшинство.

Что внутри?

- Выбирается подвыборка обучающей выборки размера `samplesize` (м.б. с возвращением) — по ней строится дерево (для каждого дерева — своя подвыборка).
- Для построения каждого расщепления в дереве просматриваем `max_features` случайных признаков (для каждого нового расщепления — свои случайные признаки).
- Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса).



Теорема о жюри присяжных

Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5 , то тогда вероятность правильного решения присяжных в целом возрастает с увеличением количества членов жюри и стремится к единице.



Мудрость топлы

800 человек пытались угадать вес быка, который стоял перед ними. Бык весил 1198 фунтов. Ни один крестьянин не угадал точный вес быка, но если посчитать среднее от их предсказаний, то получим 1197 фунтов.

Тут обойдемся одной иллюстрацией

