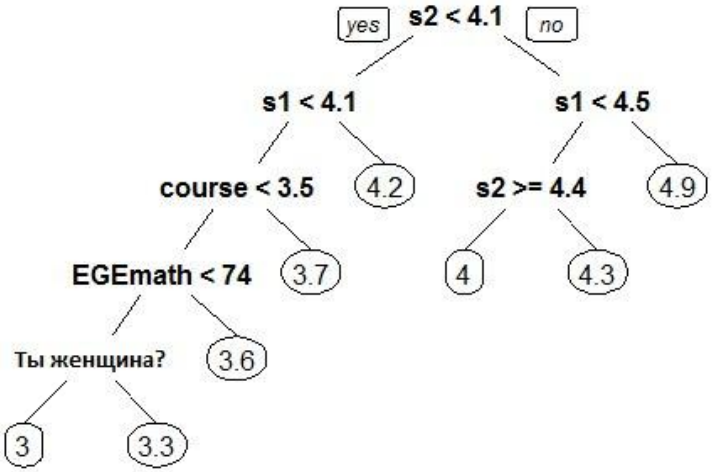
Деревья

Лекция 8

Пример

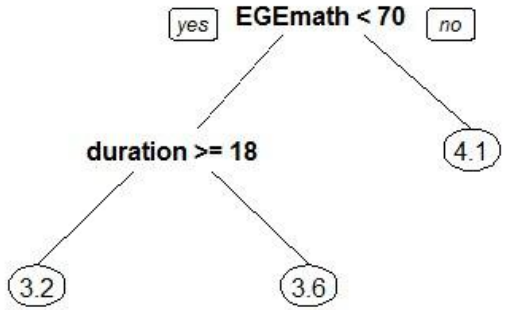
Пример дерева для классификации

Это дерево предсказывает средний балл студента на ближайшей сессии (задача регрессии). Здесь s1 (s2) – средний балл за предпоследнюю (последнюю) сессию. Еще использованы «номер курса», «баллы ЕГЭ по математике», «пол».



Предсказание для первокурсников

(у них нет данных о предыдущих сессиях)
duration – сколько времени (мин.)
студент тратит на дорогу «дом-универ»

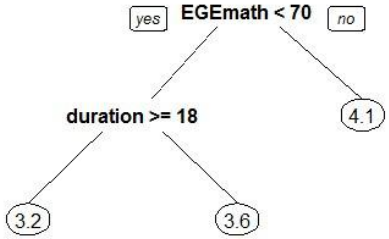


Терминология

корень

лист

В «не-листьях» стоят условия на ветвление.
В листах стоят предсказания модели.



Основной вопрос

Как найти оптимальное условие для ветвления???

Для этого нужно:

1. Найти оптимальный признак.
2. Найти значение этого признака для ветвления.

Если в п.1 был выбран бинарный признак, то п.2 можно не делать.

Для простоты далее будем предполагать, что все признаки бинарные (случай не бинарных признаков см. в [1]). Например, вот задача о предсказании результата матча:

V – соперник выше в турнирной таблице
D – играем дома
L – лидеры команды участвуют в матче
R –во время матча идет дождь

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	1	0
2	1	1	1	0	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	1	1
7	1	0	1	1	0

Как найти оптимальный признак для ветвления

Есть несколько критериев, каждый з которых вычисляет свою величину:

1. Энтропия [1]
2. Неопределенность Джини (изучим на этой лекции)
3. Статистическая информативность [2]
4. и т.д.

Неопределенность Джини (Gini impurity)

Для признака П она считается по формуле (Y – целевой признак):

$$Gini(П)=Pr(П=0)*Pr(Y=0|П=0)*Pr(Y=1|П=0)+$$
$$+Pr(П=1)*Pr(Y=0|П=1)*Pr(Y=1|П=1)$$

Факт: Gini определяет разброс значений Y при фиксированном значении признака П.

Правило: для ветвления нужно брать признак с минимальным Gini.

Вычисляем

Pr(V=0)=… Pr(V=1)=…

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	1	0
2	1	1	1	0	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	1	1
7	1	0	1	1	0

Вычисляем

$Pr(V=0)=3/7$ $Pr(V=1)=4/7$

$Pr(Y=0|V=0)=...$
$Pr(Y=1|V=0)=...$
$Pr(Y=0|V=1)=...$
$Pr(Y=1|V=1)=...$

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	1	0
2	1	1	1	0	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	1	1
7	1	0	1	1	0

Вычисляем

$Pr(V=0)=3/7$ $Pr(V=1)=4/7$

$Pr(Y=0|V=0)=1/3$
$Pr(Y=1|V=0)=2/3$
$Pr(Y=0|V=1)=2/4$
$Pr(Y=1|V=1)=2/4$

$Gini(V)=3/7*1/3*2/3+$
$4/7*2/4*2/4=0.22$

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	1	0
2	1	1	1	0	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	1	1
7	1	0	1	1	0

Вычисляем Джини для второго признака

Pr(D=0)=2/7 Pr(D=1)=5/7

Pr(Y=0|D=0)=2/2
Pr(Y=1|D=0)=0/2
Pr(Y=0|D=1)=2/5
Pr(Y=1|D=1)=3/5

Gini(D)=2/7*2/2*0/2+
5/7*2/4*3/4=0.17

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	1	0
2	1	1	1	0	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	1	1
7	1	0	1	1	0

Аналогично для других признаков

Gini(L)=0.24
Gini(R)=0.24

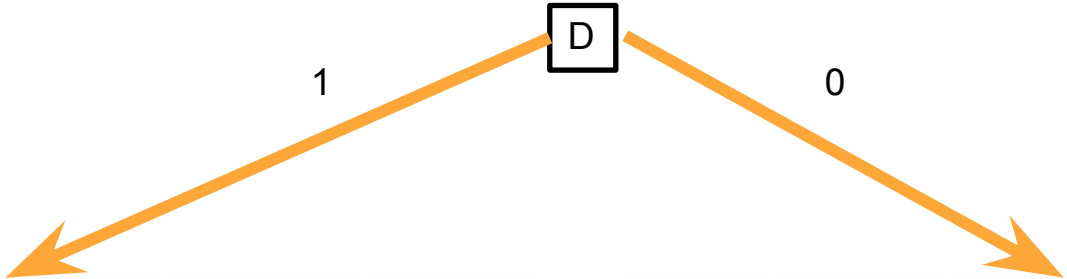
№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	1	0
2	1	1	1	0	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	0	0	0	0
6	0	1	0	1	1
7	1	0	1	1	0

Находим признак с минимальным Джини

Gini(V)=0.22
Gini(D)=0.17 (!)
Gini(L)=0.24
Gini(R)=0.24

Этот признак идет в вершину дерева. Тренировочная выборка разбивается на 2 части

Находим признак с минимальным Джини



№ матча	Выше? (V)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	0
2	1	1	0	1
3	1	0	0	0
4	0	0	0	1
6	0	0	1	1

№ матча	Выше? (V)	Лидеры? (L)	Дождь? (R)	Победа (Y)
5	0	0	0	0
7	1	1	1	0

В этом листе все объекты принадлежат одному классу (Y=0). Здесь ветвления заканчиваются.

Теперь работаем с объектами из левой вершины

$Pr(V=0)=2/5$ $Pr(V=1)=3/5$

$Pr(Y=0|V=0)=0/2$
$Pr(Y=1|V=0)=2/2$
$Pr(Y=0|V=1)=2/3$
$Pr(Y=1|V=1)=1/3$

$Gini(V)=2/5*0/2*2/2+ 3/5*2/3*1/3=0.13$

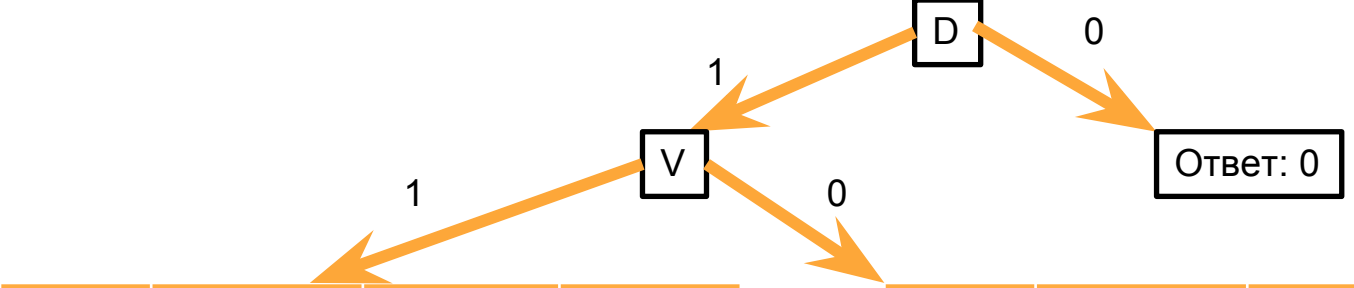
№ матча	Выше? (V)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	0
2	1	1	0	1
3	1	0	0	0
4	0	0	0	1
6	0	0	1	1

Теперь работаем с объектами из левой вершины

Получаем
Gini(V)=0.13
Gini(L)=0.23
Gini(R)=0.23

Выбираем для ветвления признак V

№ матча	Выше? (V)	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	1	0
2	1	1	0	1
3	1	0	0	0
4	0	0	0	1
6	0	0	1	1



№ матча	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	0
2	1	0	1
3	0	0	0

Будем строить ветвление здесь

№ матча	Лидеры? (L)	Дождь? (R)	Победа (Y)
4	0	0	1
6	0	1	1

В этом листе все объекты принадлежат одному классу (Y=1). Здесь ветвления заканчиваются.

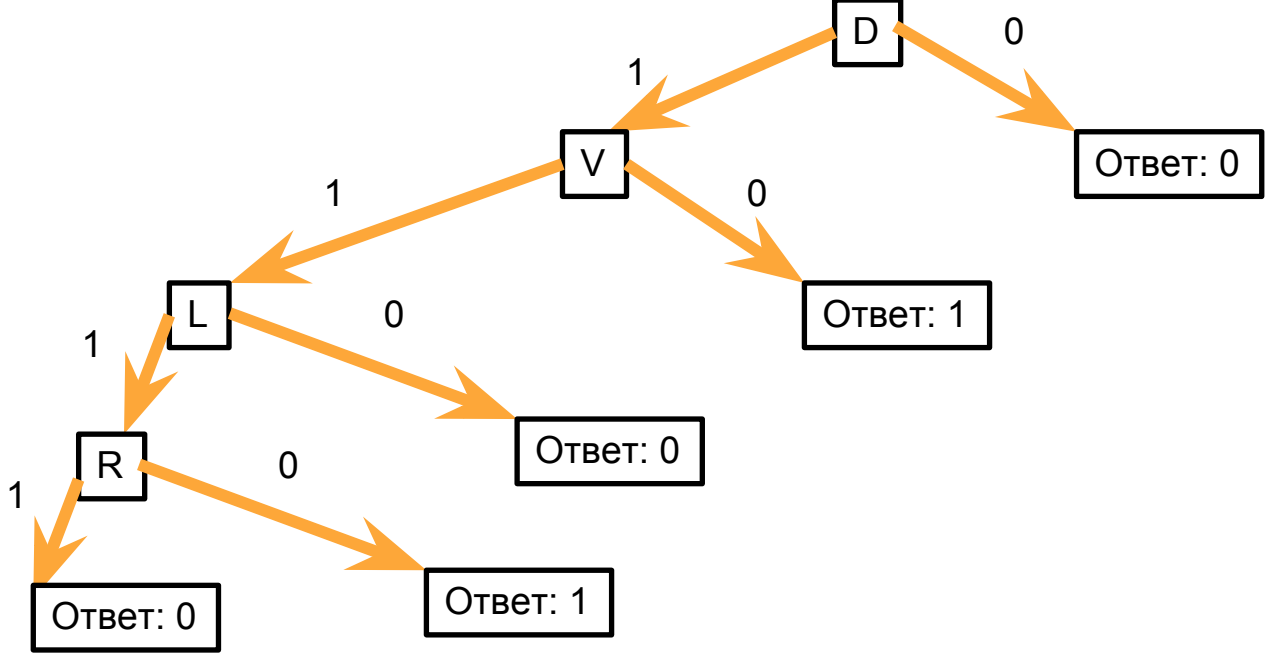
Теперь работаем с объектами из левой вершины

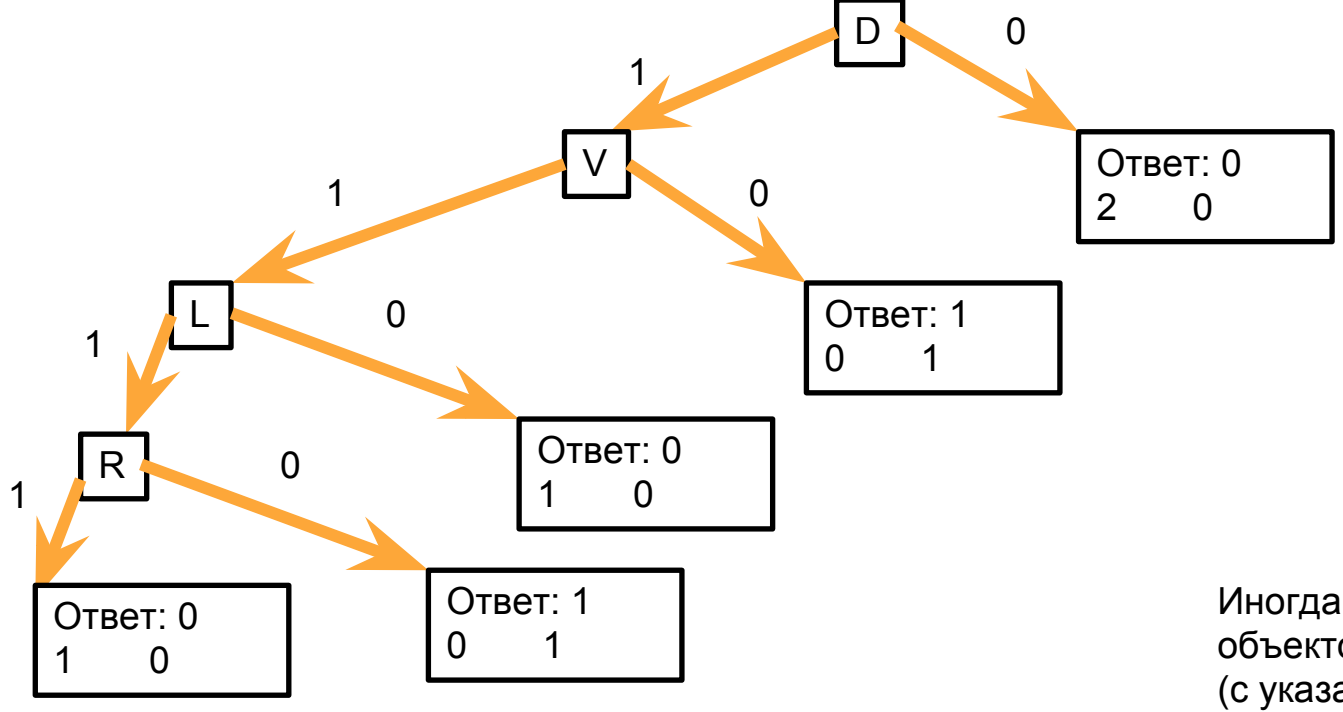
Здесь получаем

Gini(L)=0.16
Gini(R)=0.16

Можно выбрать любой из признаков.

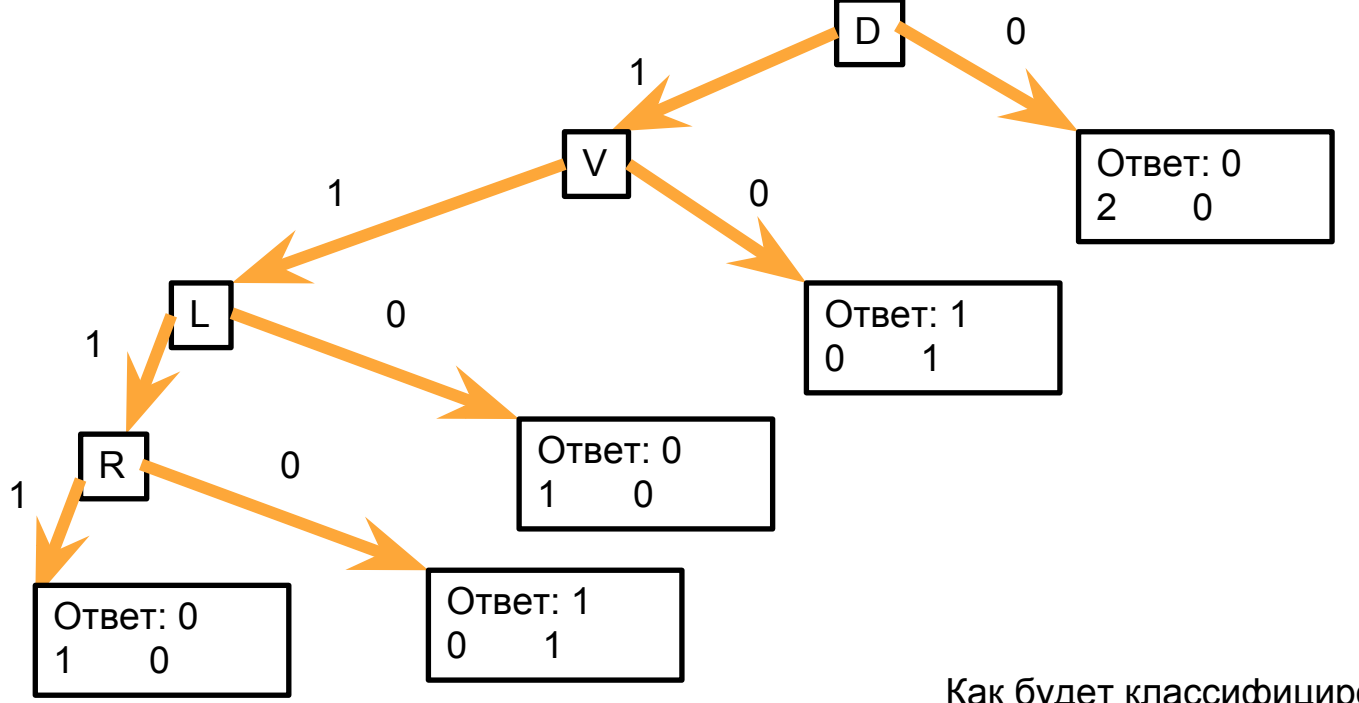
№ матча	Лидеры? (L)	Дождь? (R)	Победа (Y)
1	1	1	0
2	1	0	1
3	0	0	0


D
0
1
Ответ: 0
V
1
0
Ответ: 1
L
1
0
Ответ: 0
R
1
0
Ответ: 0
Ответ: 1



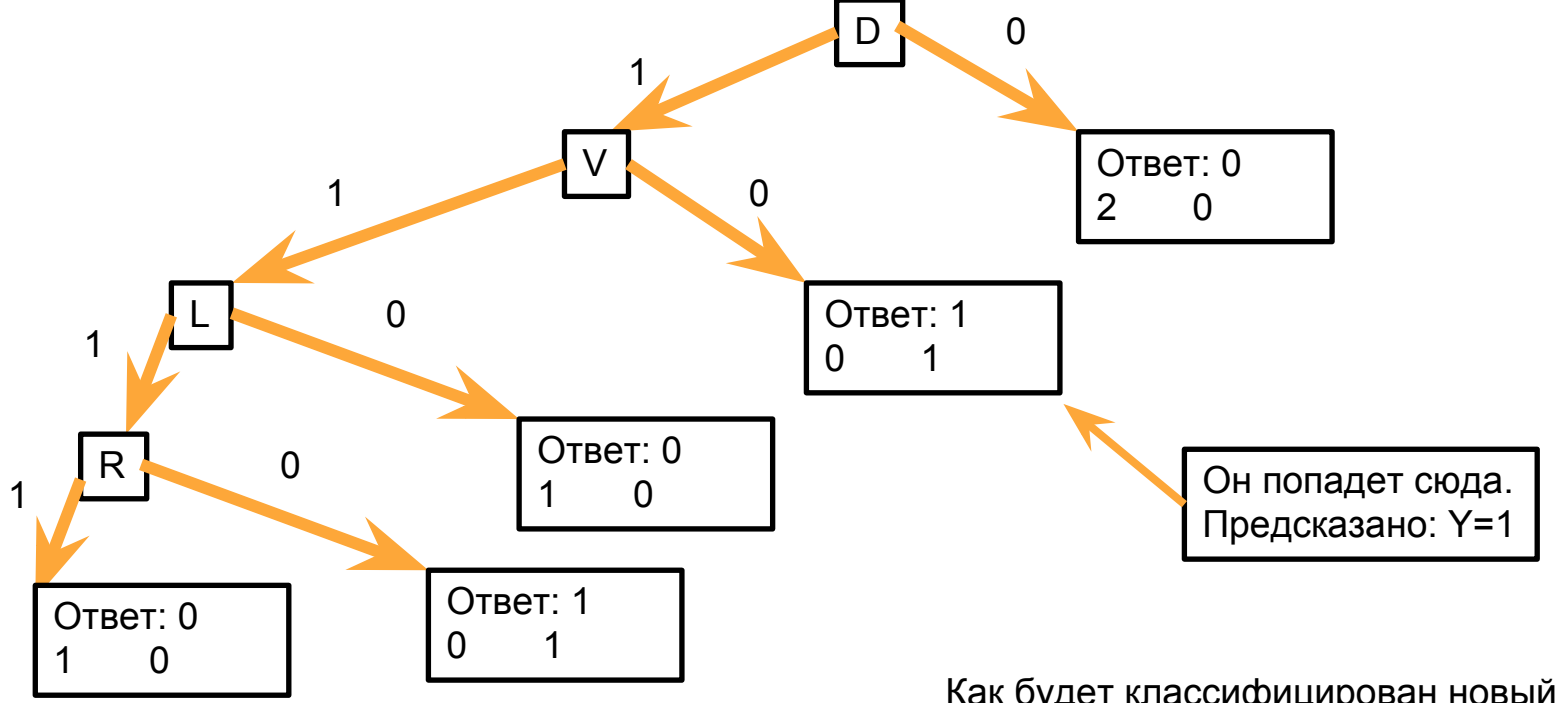
Иногда удобно писать количество объектов тренировочной выборки (с указанием их классов), попавших в каждый лист

Если в лист попали тренировочные объекты разных классов, то лист относит тестовые объекты к преобладающему классу.



Как будет классифицирован новый объект?

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
8	0	1	1	0	?



Как будет классифицирован новый объект?

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
8	0	1	1	0	?

Деревья для задачи регрессии

На прошлых слайдах было построено дерево для задачи классификации

А как строить дерево для решения задачи регрессии???

Нужно ответить на вопросы:

1. По какому правилу находить признак для ветвления?
2. Как выбрать предсказываемое значение для листа?

Пример задачи регрессии

Как выбирать признак?

Правило: выбирай признак, который минимизирует величину

$Pr(П=0)*s_{Y(П=0)}+Pr(П=1)*s_{Y(П=1)}$

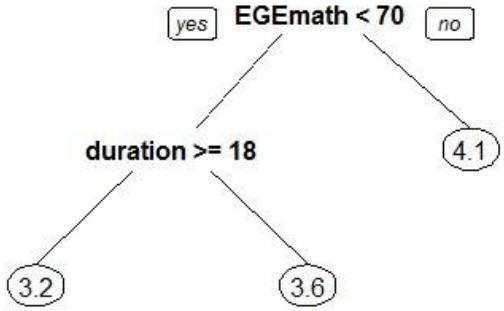
где

$s_{Y(П=i)}$ – отклонение значений Y, для которых П=i

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Число зрителей(Y)
1	1	1	1	1	23
2	1	1	1	0	10
3	1	1	0	0	5
4	0	1	0	0	54
5	0	0	0	0	14
6	0	1	0	1	22
7	1	0	1	1	20

В задаче регрессии значение, которое выдает лист – это

…среднее арифметическое значений Y тренировочных объектов, попавших в этот лист.



Деревья vs пропуски данных

Деревья могут предсказывать…

… значение целевого признака даже для объектов с пропусками.

Для этого нужно:

Объект обрабатывается деревом обычным способом, при попадании в вершину, в которой используется пропущенный признак…

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
8	1	?	1	0	?

Для этого нужно:

Объект обрабатывается деревом обычным способом, при попадании в вершину, в которой используется пропущенный признак происходит ветвление процесса: мы начинаем идти по обеим ветвям из этой вершины.
В итоге мы можем попасть более чем в один лист.

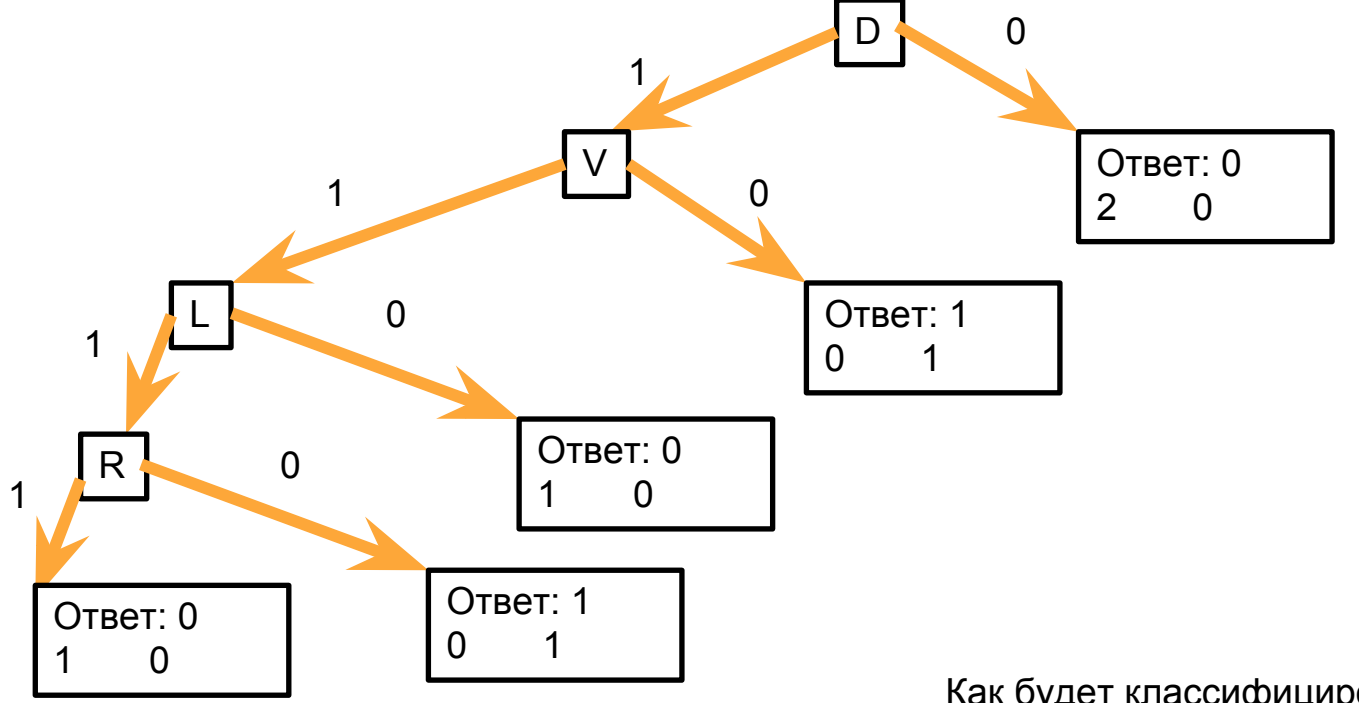
Какой выдать итоговый ответ? Для классификации? Для регрессии?

Итоговый ответ, если попали в несколько листьев

Пусть M - множество объектов тренировочной выборки, которые попадают в эти листья.

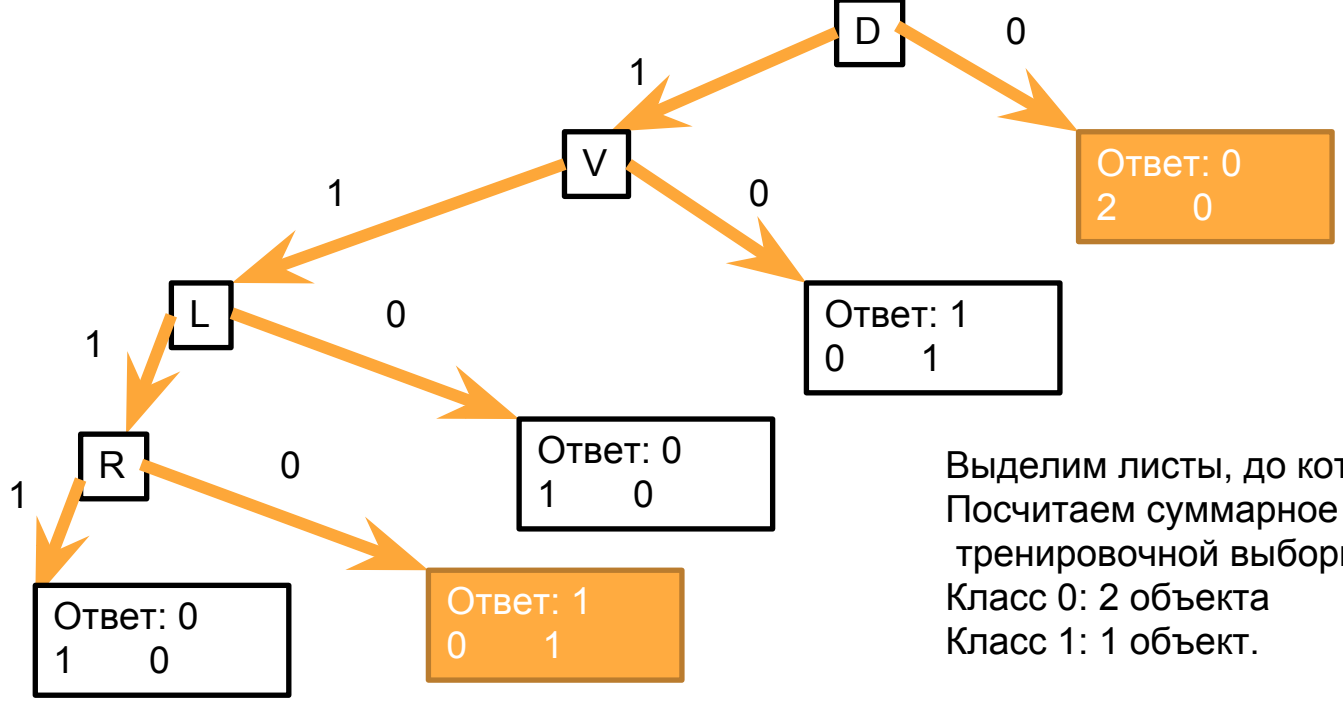
При предсказании числового признака (регрессия) в качестве ответа нужно взять среднее значение признака Y в M.

При предсказании метки класса в качестве ответа нужно взять метку преобладающего класса в M.



Как будет классифицирован объект с пропусками данных?

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
8	1	?	1	0	?



Выделим листы, до которых дошёл процесс.
Посчитаем суммарное число объектов тренировочной выборки из этих листьев.
Класс 0: 2 объекта
Класс 1: 1 объект.

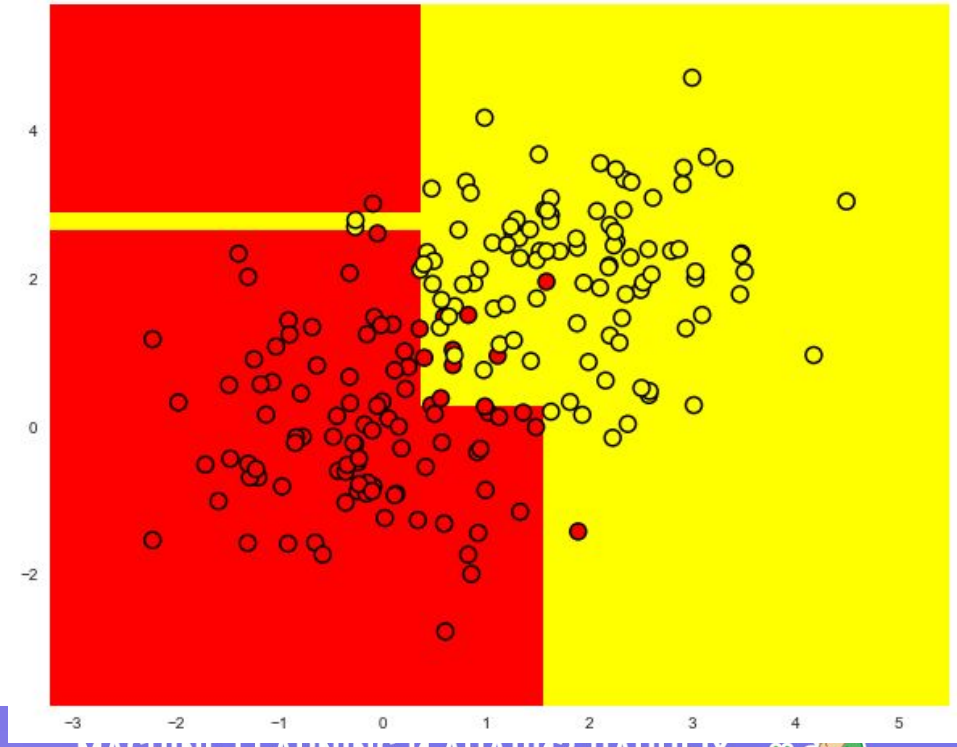
Решение принимает по доминирующему классу.
То для нового объекта будет предсказано Y=0

№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
8	1	?	1	0	?

Поиск выбросов с помощью деревьев (изолирующий лес)

Дерево разбивает пространство на прямоугольные секторы

Это можно использовать в задаче поиска выбросов.
Выброс должен лежать на периферии – следовательно, в секторе с выбросами будет мало элементов.



Изолирующий лес

1. Нужно построить N деревьев.
2. Каждое дерево строится до исчерпании выборки
3. Для построения ветвления в дереве: выбирается случайный признак и случайное значение для расщепления.

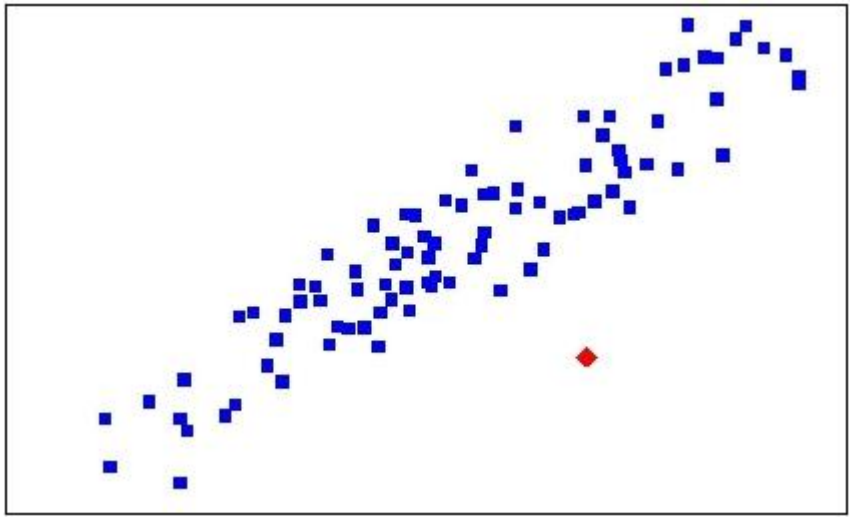
Для каждого объекта мера его нормальности – среднее арифметическое глубин листьев, в которые он попал (изолировался)

Изолирующий лес

Логика алгоритма простая: при «случайном» способе построения деревьев выбросы будут попадать в листья на ранних этапах (на небольшой глубине дерева), т.е. выбросы проще «изолировать». Для изоляции выбросов требуется меньшее число условий

Действительно,

чтобы вычленить (изолировать) красную точку, требуется 2 условия. А чтобы изолировать точку из центра выборки, нужно (в лучшем случае) четыре условия с очень жесткими ограничениями.



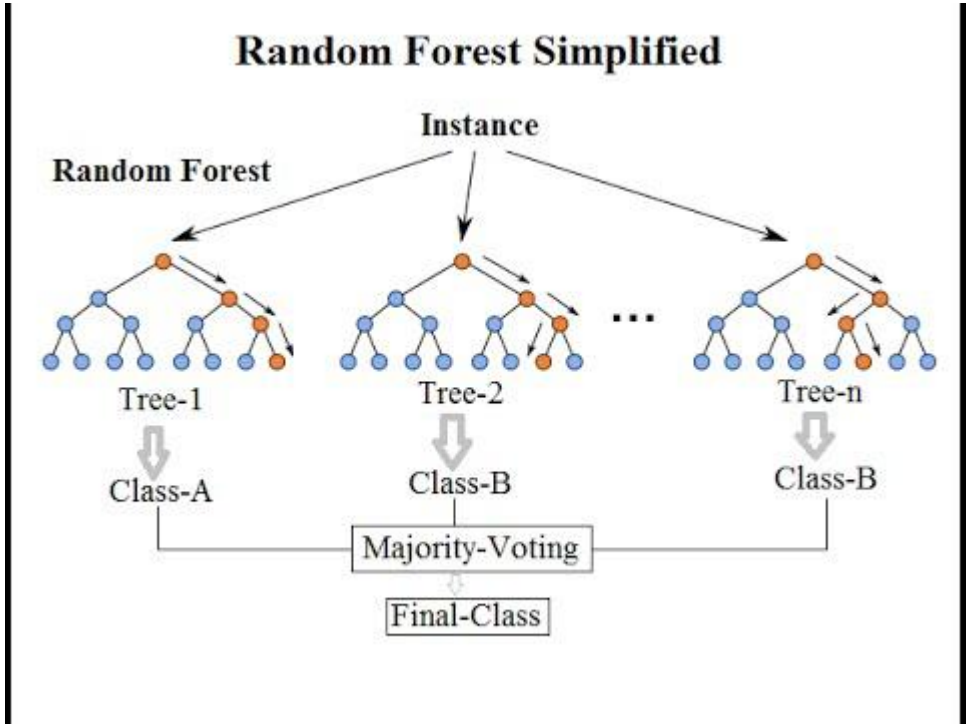
Случайный лес (Random forest)

Идея:

Построить несколько деревьев. Собрать их ответы для тестируемого объекта A. В качестве окончательного ответа выдать:

1. Среднее значение (если предсказывается числовой признак).
2. Метку преобладающего класса (если предсказывается категориальный признак).

А как построить несколько деревьев по одной выборке?


Random Forest Simplified
Instance
Random Forest
Tree-1
Tree-2
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class

Строительство случайных деревьев

1. Выбирается подвыборка обучающей выборки – по ней строится дерево (для каждого дерева — своя подвыборка).
2. Для построения каждого расщепления в дереве просматриваем p случайных признаков (для каждого нового расщепления — свои случайные признаки).
3. Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию) и т.д.

Использованная литература

1. https://habrahabr.ru/company/ods/blog/322534/ (здесь же про энтропию и не бинарные признаки)
2. Лекции Воронцова по деревьям http://www.machinelearning.ru/wiki/images/3/3e/Voron-ML-Logic.pdf
3. https://logic.pdmi.ras.ru/~sergey/teaching/ml/notes-01-dectrees.pdf (здесь пример про предсказание результата матча, но дерево строится с помощью энтропии (а не Джини))