

# Деревья в машинном обучении

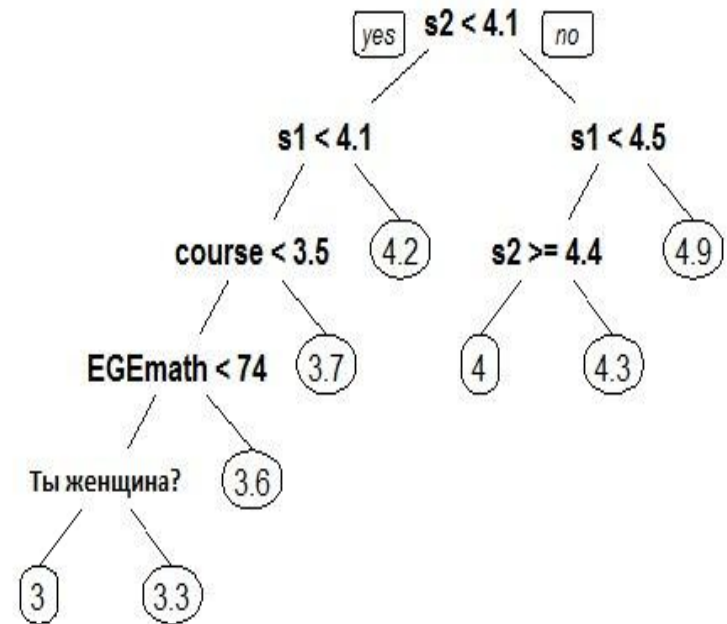
Лекция №7

Лектор: Артём Шевляков

**Здесь лучше начать с примера**

## 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».

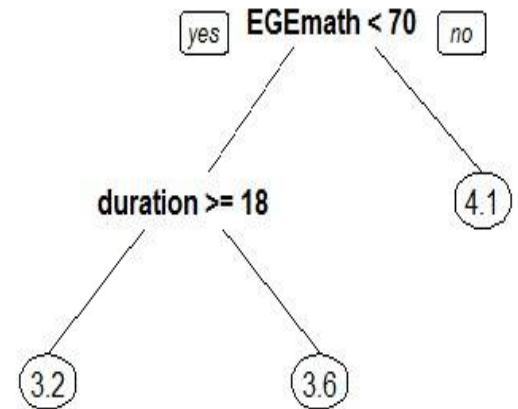


## Предсказание для первокурсников

(у них нет данных о предыдущих сессиях)

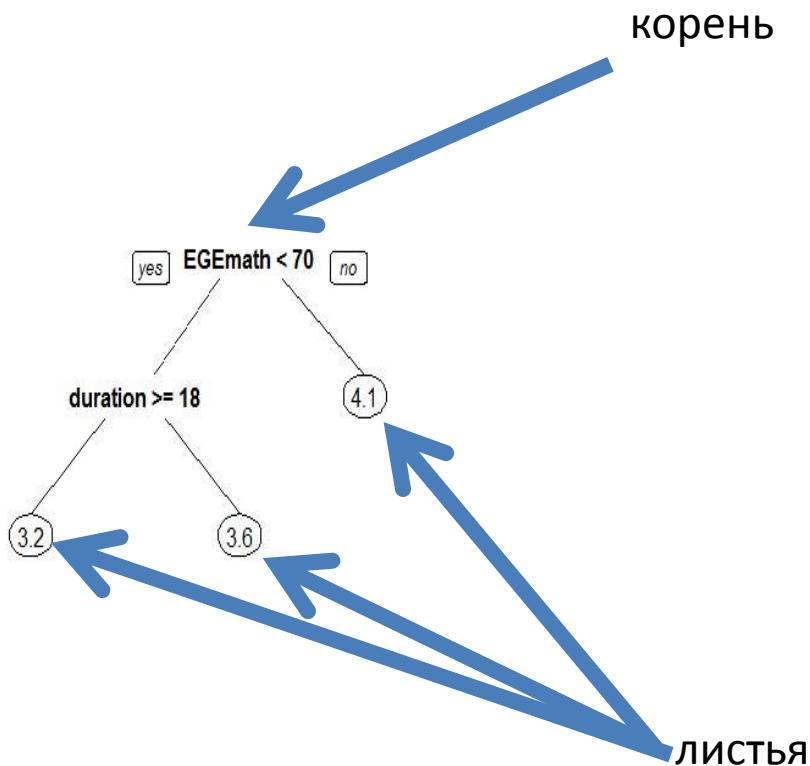
duration – сколько времени (мин.)

студент тратит на дорогу «дом-универ»



# Терминология

В «не-листьях» стоят условия на ветвление.  
В листьях 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 из которых вычисляет свою величину:

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



# Но смысл всех этих умных критериев состоит в следующем:

как определить, что один из признаков определяет  $Y$   
лучше, чем другой?

Какой тут из признаков лучше всего определяет  
признак  $Y$ ?

P1	P2	P3	P4	Y
0	1	0	0	0
0	1	1	1	0
1	0	0	1	1
1	0	1	1	1

А как это формализовать?

## Неопределенность Джини (Gini impurity)

Для признака  $\Pi$  она считается по формуле ( $Y$  – целевой признак):

$$\text{Gini}(\Pi) = \Pr(\Pi=0) * \Pr(Y=0 | \Pi=0) * \Pr(Y=1 | \Pi=0) + \\ + \Pr(\Pi=1) * \Pr(Y=0 | \Pi=1) * \Pr(Y=1 | \Pi=1)$$

## Неопределенность Джини (Gini impurity)

$$\text{Gini}(\Pi) = \Pr(\Pi=0) * \Pr(Y=0 | \Pi=0) * \Pr(Y=1 | \Pi=0) + \\ + \Pr(\Pi=1) * \Pr(Y=0 | \Pi=1) * \Pr(Y=1 | \Pi=1)$$

Для таблицы

P1	P2	P3	P4	Y
0	1	0	0	0
0	1	1	1	0
1	0	0	1	1
1	0	1	1	1

имеем

$$\text{Gini}(P1) = 0.5 * 1 * 0 + 0.5 * 0 * 1 = 0$$

$$\text{Gini}(P2) = 0.5 * 0 * 1 + 0.5 * 1 * 0 = 0$$

$$\text{Gini}(P3) = 0.5 * 0.5 * 0.5 + 0.5 * 0.5 * 0.5 = 0.25$$

$$\text{Gini}(P4) = 0.25 * 1 * 0 + 0.75 * 1/3 * 2/3 = 1/6$$

## Неопределенность Джини (Gini impurity)

Последний пример показывает, что чем меньше Gini, тем лучше признак.

**Правило:** при построении дерева для ветвления нужно брать признак с минимальным Gini.

## Вычисляем

$\Pr(V=0)=\dots$   $\Pr(V=1)=\dots$

№ 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

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

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

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

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

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

№ 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

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

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

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

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

$$\Pr(Y=1|V=1)=1/4$$

$$\text{Gini}(V)=3/7*1/3*2/3+4/7*3/4*1/4=0.21$$

№ 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

$$\Pr(D=0)=2/7 \quad \Pr(D=1)=5/7$$

$$\Pr(Y=0 | D=0)=2/2$$

$$\Pr(Y=1 | D=0)=0/2$$

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

$$\Pr(Y=1 | D=1)=3/5$$

$$\text{Gini}(D)=2/7*2/2*0/2+5/7*2/4*3/4=0.17$$

№ 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

$\text{Gini}(L)=0.24$

$\text{Gini}(R)=0.24$

№ 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

$$\text{Gini}(V)=0.22$$

$$\text{Gini}(D)=0.17 \text{ (!)}$$

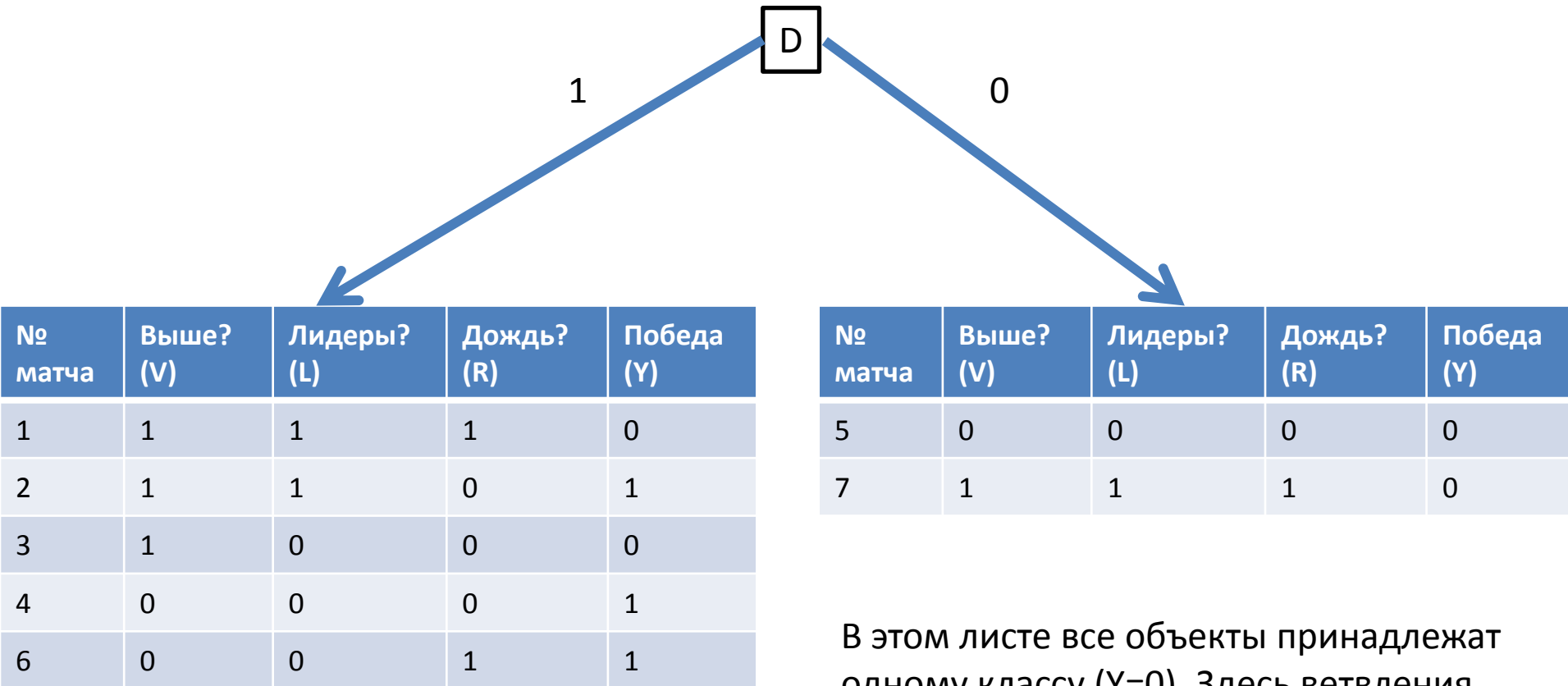
$$\text{Gini}(L)=0.24$$

$$\text{Gini}(R)=0.24$$

Этот признак идет в  
вершину дерева.

Тренировочная  
выборка разбивается  
на 2 части

## Находим признак с минимальным Джини



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

Теперь работаем с объектами из левой вершины

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

$$\Pr(Y=0|V=0)=0/2$$

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

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

$$\Pr(Y=1|V=1)=1/3$$

$$\text{Gini}(V)=2/5*0/2*2/2+3/5*2/3*1/3=0.13$$

№ 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

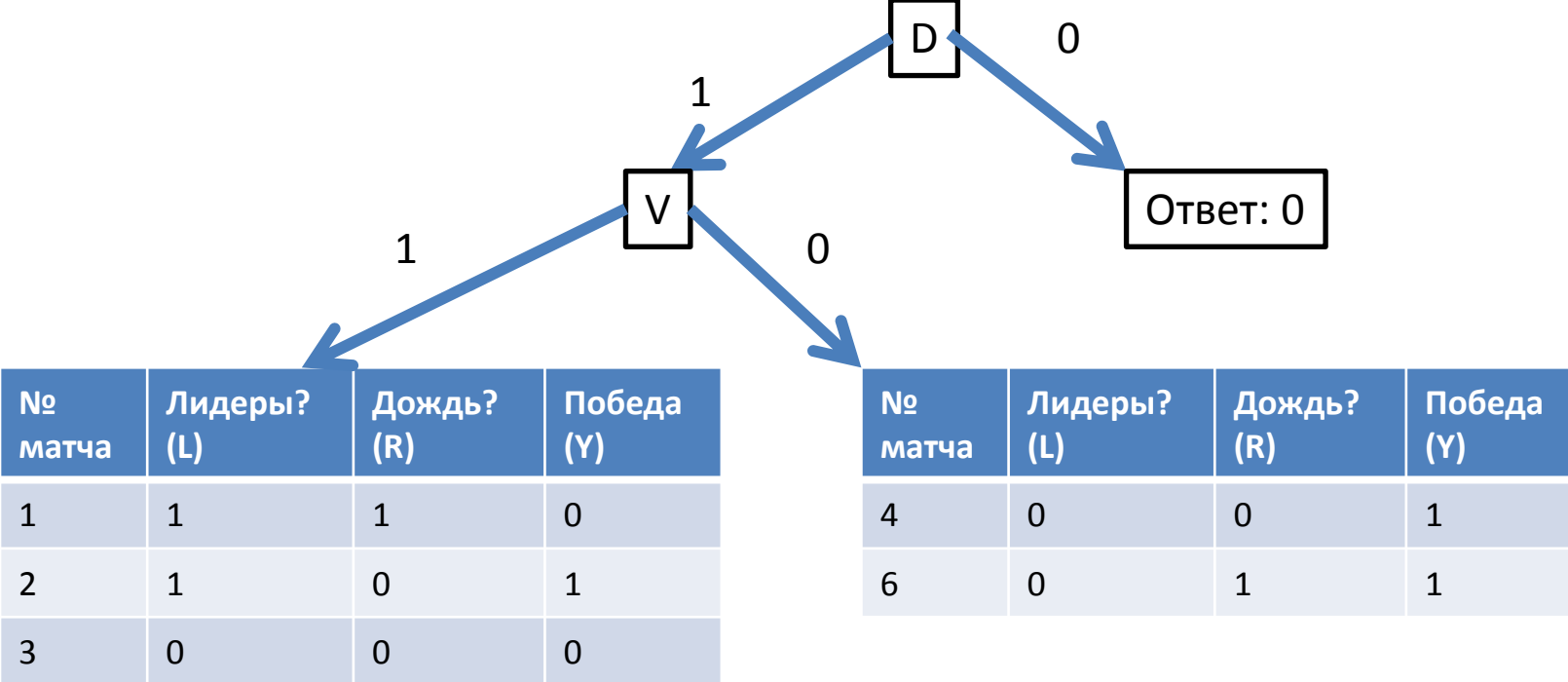
$$\text{Gini}(V)=0.13$$

$$\text{Gini}(L)=0.23$$

$$\text{Gini}(R)=0.23$$

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



Будем строить ветвление здесь

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

Теперь работаем с объектами из левой вершины

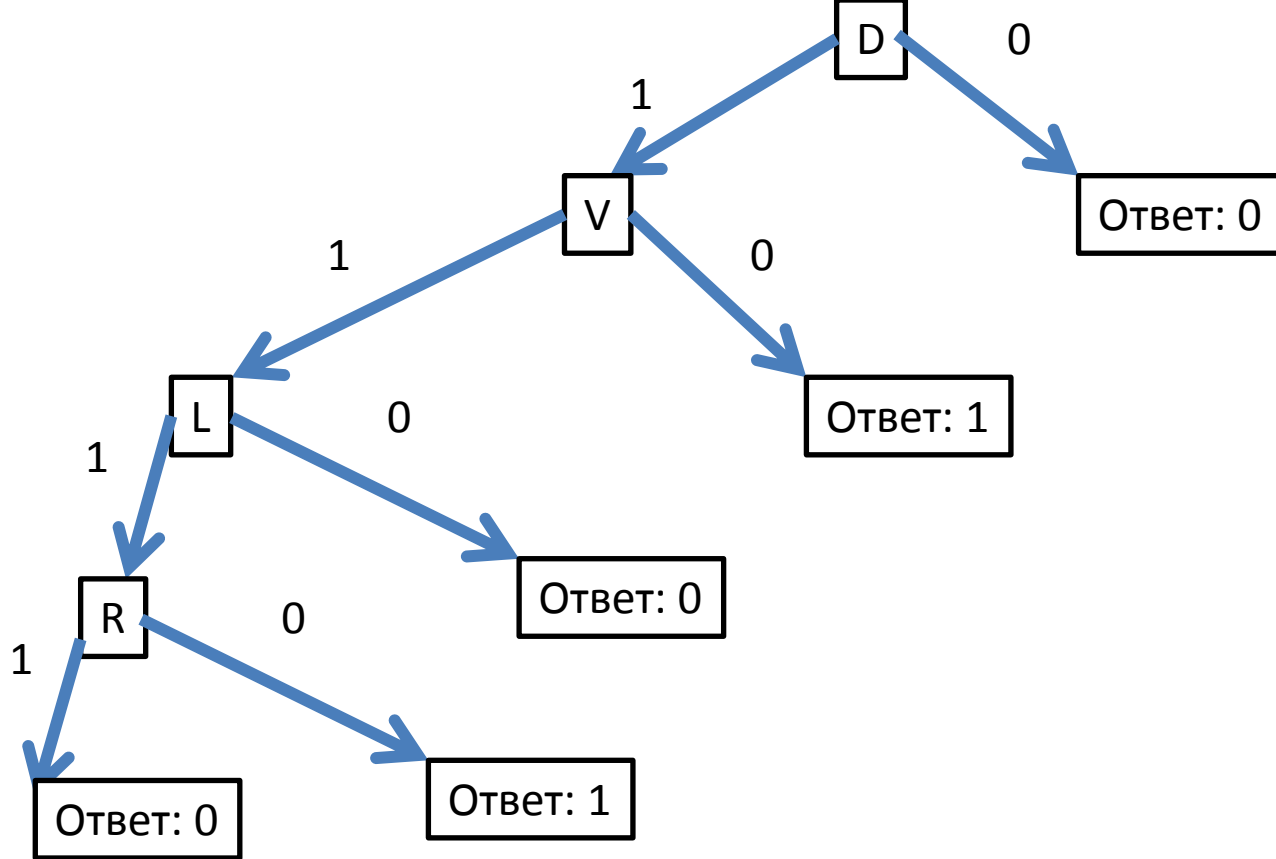
Здесь получаем

$$\text{Gini}(L)=0.16$$

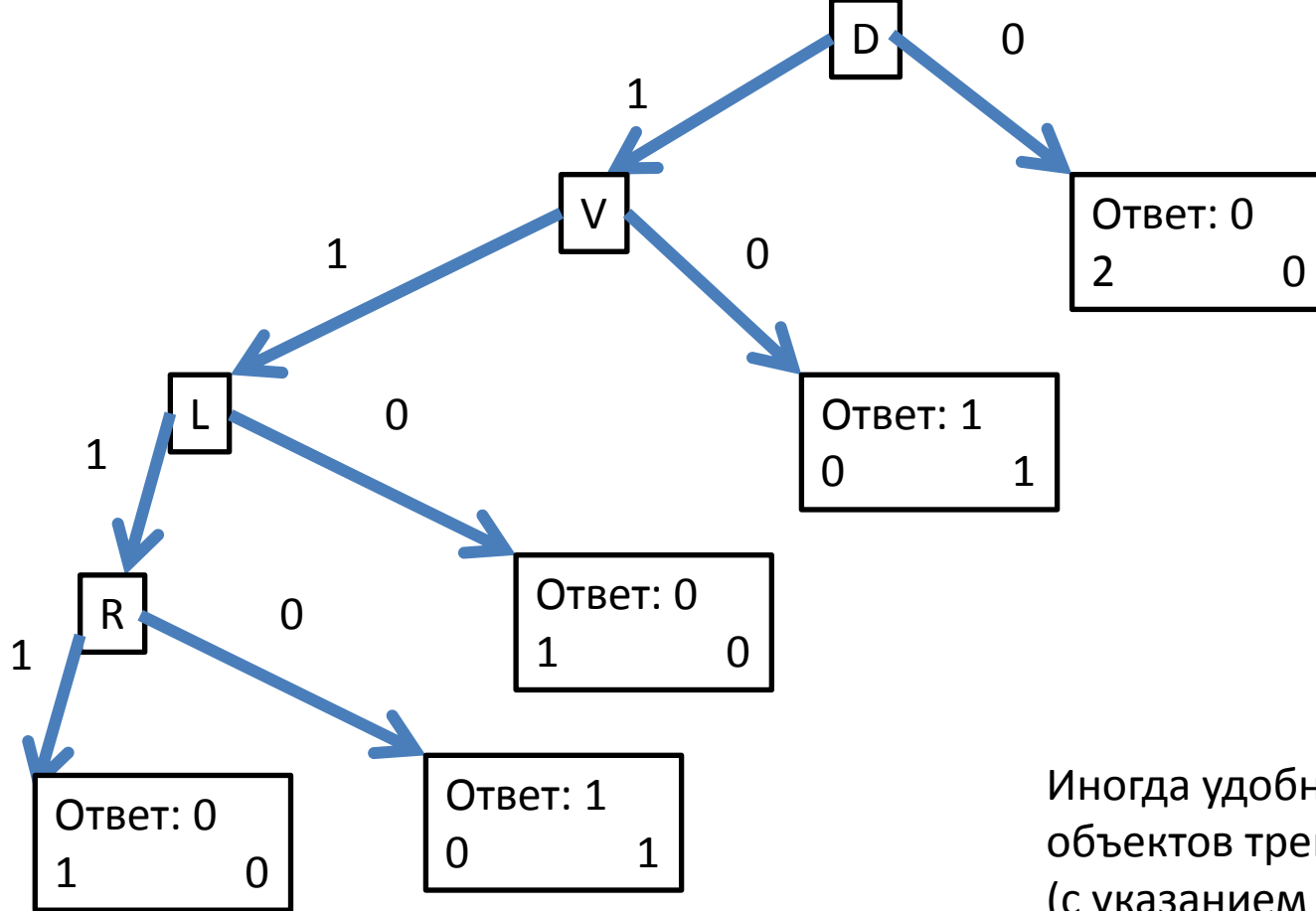
$$\text{Gini}(R)=0.16$$

Можно выбрать любой из признаков.

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

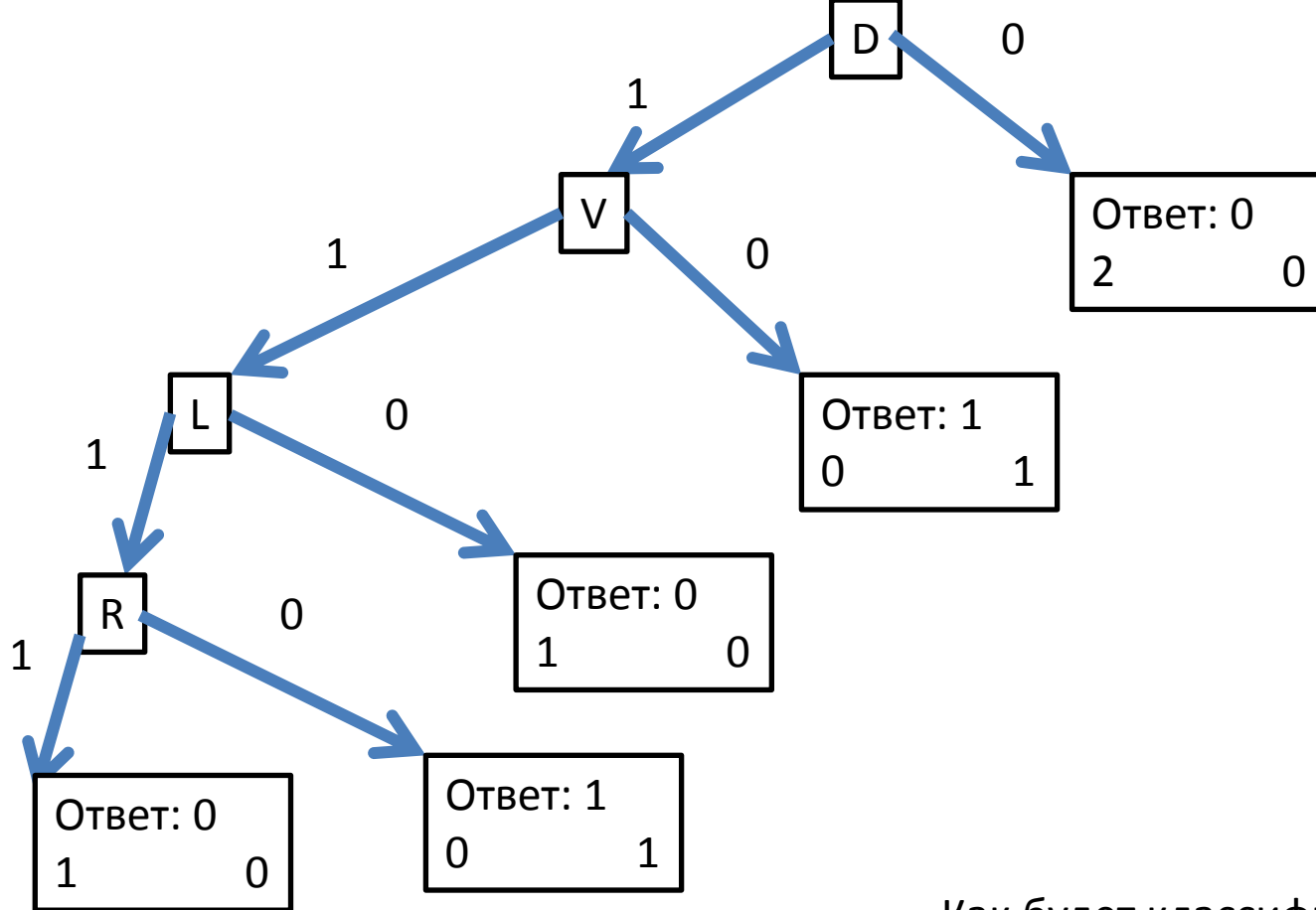






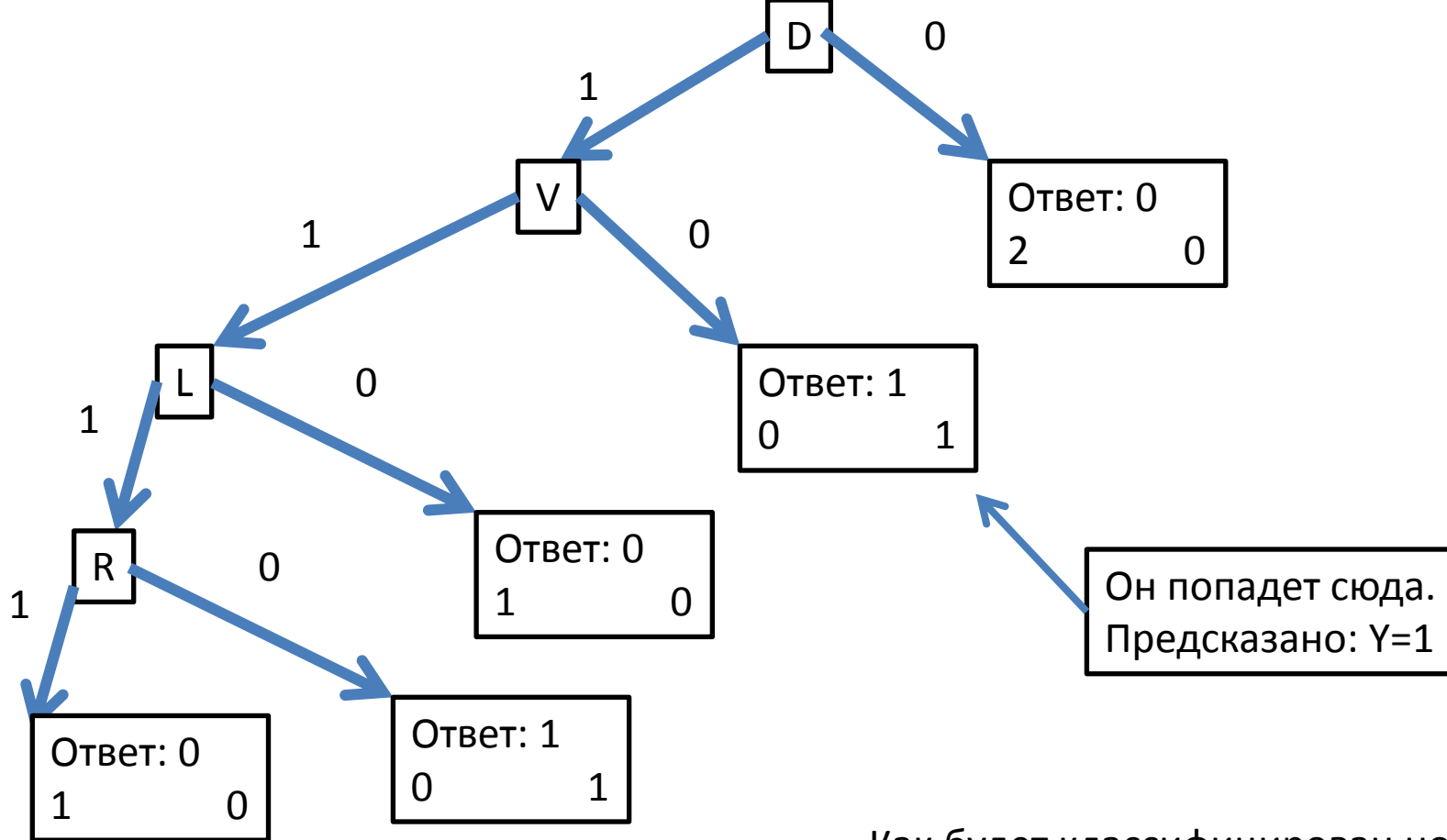
Иногда удобно писать количество объектов тренировочной выборки (с указанием их классов), попавших в каждый лист

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 (самое простое):** выбирай признак, который минимизирует величину

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

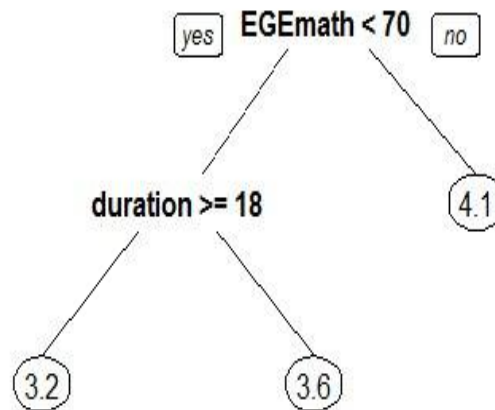
где

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

№ 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?

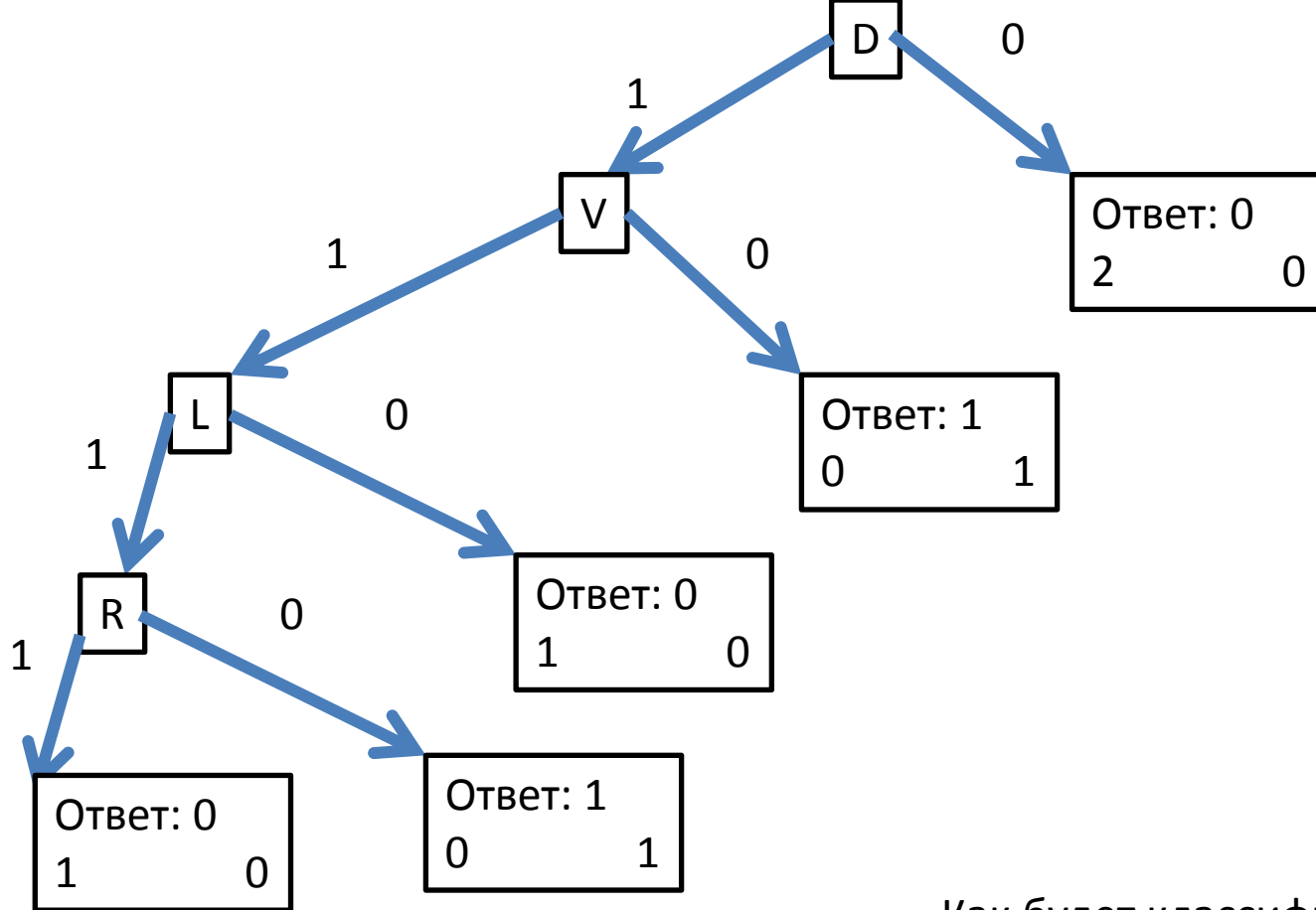
№ матча	Выше? (V)	Дома? (D)	Лидеры? (L)	Дождь? (R)	Победа (Y)
8	1	?	1	0	?

Итоговый ответ, если попали в несколько листьев

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

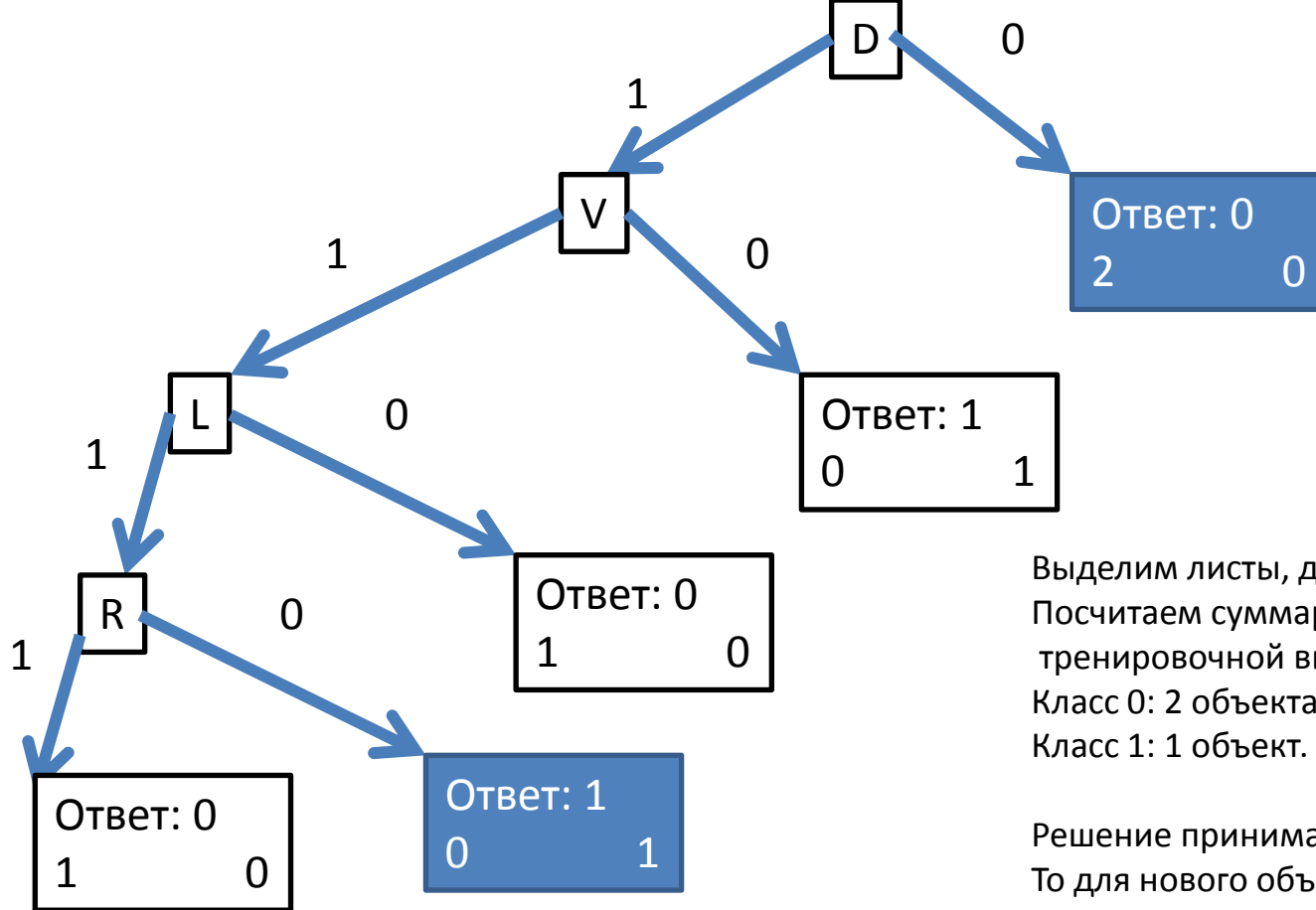
При **предсказании числового признака** (регрессия) в качестве ответа нужно взять среднее значение признака  $Y$  в  $M$ .

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

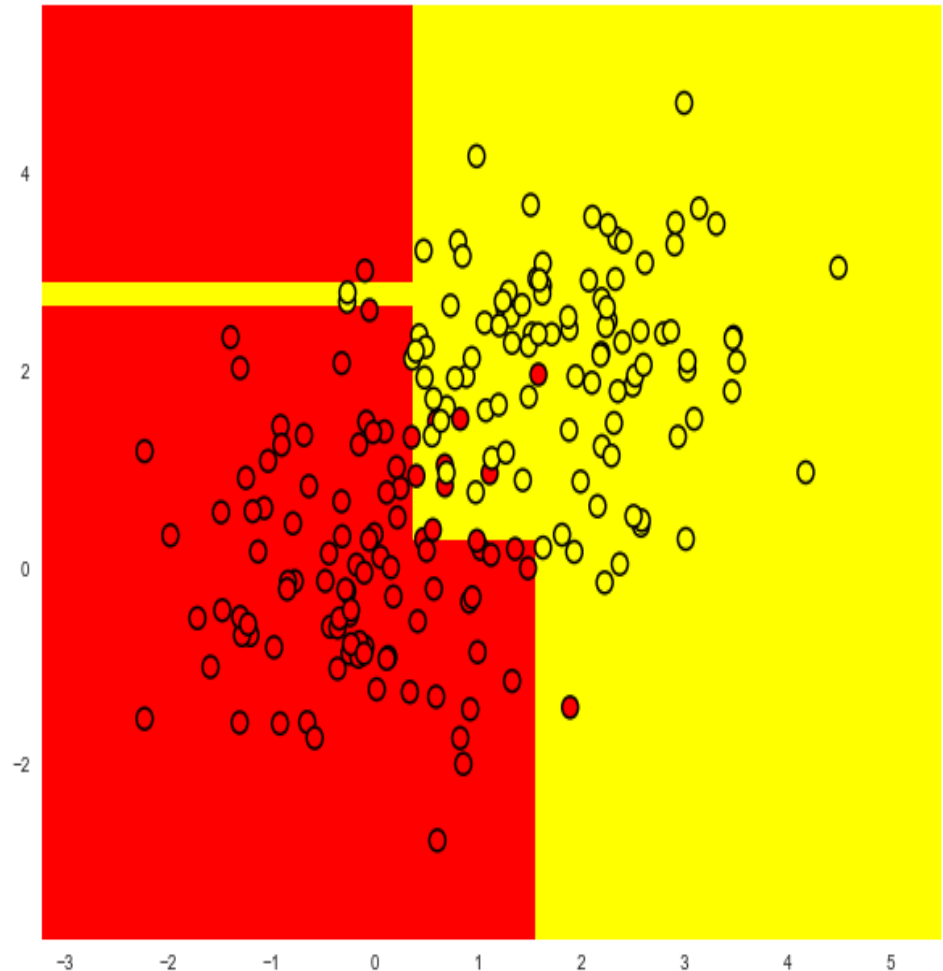
Решение принимает по доминирующему классу.  
То для нового объекта будет предсказано  $Y=0$

№ 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 исчерпаниия выборки
3. Для построения ветвления в дереве: выбирается случайный признак и случайное значение для расщепления.

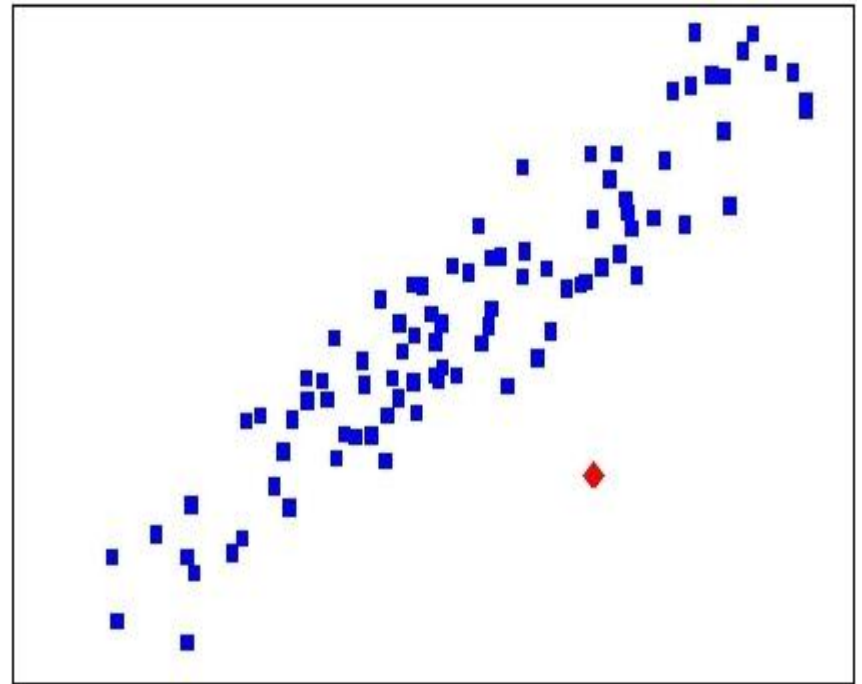
Для каждого объекта **мера его нормальности** – среднее арифметическое глубин листьев, в которые он попал (изолировался)

## 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.



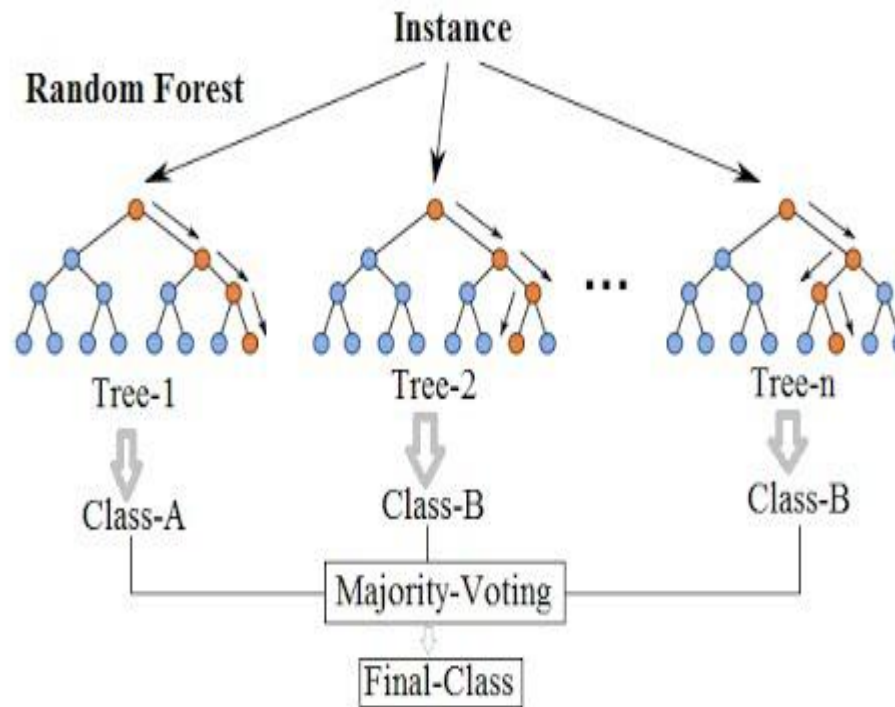
# **Случайный лес (Random forest)**

## Идея:

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

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

## Random Forest Simplified



А как построить несколько деревьев по одной выборке?

## 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 наилучший 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))