



# Энтропия Деревья принятия решений

Высшая Школа Цифровой Культуры  
Университет ИТМО

[dc@itmo.ru](mailto:dc@itmo.ru)

# **Содержание**

<b>1</b>	<b>Введение</b>	<b>2</b>
<b>2</b>	<b>Энтропия</b>	<b>4</b>
2.1	Немного о степени неопределенности . . . . .	4
2.2	Энтропия Шеннона и проверка ожидаемых свойств . . . . .	6
<b>3</b>	<b>Условная энтропия</b>	<b>11</b>
3.1	Наводящие соображения . . . . .	11
3.2	Определение условной энтропии . . . . .	12
3.3	Пример на вычисление условной энтропии . . . . .	17
3.4	Энтропия и прирост информации . . . . .	19
3.5	Вычисление прироста информации на конкретном примере . . .	21
<b>4</b>	<b>Деревья принятия решений</b>	<b>23</b>
4.1	Немного о самих деревьях . . . . .	23
4.2	Алгоритм построения дерева и пример с кошками . . . . .	24
4.3	Бинарное дерево решений . . . . .	28
4.4	Типы признаков и их группировка . . . . .	29
4.4.1	Алгоритм построения бинарного дерева решений и пример с кошками . . . . .	31
4.4.2	Синтетический пример . . . . .	32
<b>5</b>	<b>Неопределенность Джини</b>	<b>36</b>
5.1	Определение и свойства . . . . .	36
5.2	Небольшое сравнение прироста Джини и энтропии . . . . .	42
5.3	Прирост Джини на данных . . . . .	42
<b>6</b>	<b>Деревья принятия решений на реальном примере</b>	<b>44</b>
<b>7</b>	<b>Заключение</b>	<b>47</b>

## 1 Введение

Здравствуйте, уважаемые слушатели! В данной лекции мы поговорим про такой популярный и эффективный подход к решению задачи классификации, как деревья принятия решений. На совсем бытовом уровне дерево принятия решений можно отождествить с подробной инструкцией, четко говорящей что, на каком этапе делать, и как в той или иной ситуации поступать. Посмотрите, для примера, на рисунок 1. На нем приведен чрезвычайно упрощенный алгоритм, опираясь на который, некоторый человек принимает решение: соглашаться на предложение о трудоустройстве, или нет. Этот



Рис. 1: Пример дерева решений

пример сразу демонстрирует и огромное преимущество деревьев принятия решений (из-за которого, в частности, они и завоевали столь большую популярность, причем не только в машинном обучении) – они интуитивно понятны без каких-либо дополнительных пояснений. Скорее всего вы догадались, что, согласно описанному алгоритму, первым делом человек смотрит на размер предлагаемой заработной платы, причем, в зависимости от последнего, возникает три варианта дальнейших действий. Так, если предлагаемая зарплата меньше, чем 50000, то предложение безоговорочно отклоняется, а если больше или равна ста тысячам, то сразу же принимается. Ситуация же, когда зарплата находится в диапазоне от 50000 до 100000, не является столь очевидной, как предыдущие: для принятия решения оказывается важным выяснить время транспортной доступности. И снова, если на дорогу придется затратить как минимум час, то предложение никуда не годится и отклоняется. Иначе же мы спускаемся ниже и видим, что окончательный ответ зависит от того: будут ли выплачиваться квартальные бонусы. Если да – то предложение принимается, а если нет – отклоняется.

Конечно, не смотря на всю свою простоту, написанный алгоритм вызывает вопросы. Например, что делать в случае, если информация о квартальных

бонусах по какой-либо причине не предоставляется? В описанном алгоритме нет варианта «не знаю»: есть либо «да», либо «нет». А что, если работа предполагает выезд на разные объекты? Тогда какие-то объекты могут быть расположены близко к дому, и дорога до них займет меньше, чем 60 минут, а какие-то, наоборот, далеко от дома, а значит на дорогу придется тратить больше часа. А ведь еще работа может подразумевать обязательные командировки! На все эти вопросы у предложенного алгоритма нет ответа, а значит с каждой такой ситуацией нужно разбираться отдельно. В этой лекции мы расскажем, в том числе, про так называемые бинарные (или двоичные) деревья принятия решений, которые помогают решить описанные проблемы.

Кроме того (что, наверное, вы поняли и из примера), при построении дерева важным является очередность признаков. В нашем модельном примере самый важный признак – это размер заработной платы, затем время транспортной доступности, а на последнем месте – наличие квартальных бонусов. В то же время, признаки и их очередь сильно зависят от задачи. Скажем, при решении вопроса о выдаче кредита конкретному человеку, банк, наверное, первым делом захочет узнать возрастную категорию заемщика, затем размер его дохода, образование, семейное положение и так далее. Вроде все логично и понятно, но чем больше признаков, тем сложнее построить их последовательность, и, в то же время, тем важнее построить эту последовательность правильно: мы же хотим оптимальный, быстро работающий алгоритм, а не цепочку вопросов «длинною в жизнь».

Последняя фраза хорошо поясняется на примере простой игры «данетки» или «20 вопросов». Смысл в играх одинаковый: игроки могут задавать вопрос ведущему, а тот, в свою очередь, отвечать «да» или «нет». Угадать необходимо фильм, сериал, да все что угодно, при этом тема игры известна. Очевидно, что если речь идет об актере, то, задав вопрос о его гендерной принадлежности, мы отсекаем примерно половину возможных вариантов, а если, допустим, сразу зададим вопрос: «это Николас Кейдж?», мы отсекаем всего один вариант. Это интуитивно соответствует понятию прироста информации, основанному на понятии энтропии (мере неопределенности события или эксперимента). Задав первый вопрос, мы получили куда больше информации, так как убрали много неверных вариантов, и неопределенность эксперимента по угадыванию актера сильно уменьшилась. Во втором же случае, мы отсекли лишь один вариант, тем самым неопределенность практически не изменилась.

Итак, первым делом при построении дерева принятия решений нужно понять, как выстроить последовательность признаков. Для этого признаки имеет смысл разделить на более и менее информативные. Но что же такое информативность и как она измеряется? Давайте с обсуждения этих вопросов и начнем.

## 2 Энтропия

### 2.1 Немного о степени неопределенности

В повседневной жизни мы постоянно задумываемся о различных событиях. При этом у какой-то части из них мы предугадываем исход с большей уверенностью, а у какой-то – с меньшей. Так, например, мы ожидаем толкотню в вагоне метро в час пик, и, скорее всего, будем уверены, что не будем стоять в пробках по пути домой после изнурительного рабочего дня в 10 часов вечера. Ну или на вопрос ребенка: а что за черная птица пытается отобрать у голубей кусок хлеба, мы с большой долей вероятности обвиним ворону, не так ли? К чему эти рассуждения? А к тому, что все описанные ситуации для нас оказываются «не очень сомнительными» и несут мало неопределенности.

В то же время бывают и другие, прямо-таки противоположные примеры. Например, как вы думаете, выйдя завтра из дома, первый попавшийся вам навстречу человек будет мужчиной или женщиной? Или, скажем, оглядывая попутчиков в вагоне метро, можете ли вы с уверенностью сказать, кто из них едет до конечной? Скорее всего нет. Все потому, что во всех предложенных ситуациях достаточно большая неопределенность.

Для дальнейшего нам оказывается важным построить какую-то математическую модель этой неопределенности. Даже больше, нам будет важно уметь получать ее, этой неопределенности, численное значение. Давайте немного приблизимся к математической постановке задачи (не в самом общем виде, а лишь в том, что потребуется нам) и «пощупаем» еще раз понятие неопределенности, но уже на числах.

Итак, предположим, что эксперимент имеет лишь  $n$  возможных мельчайших несовместных (то есть таких, которые не могут произойти одновременно) исходов  $\omega_1, \omega_2, \dots, \omega_n$ . Кроме того, для каждого исхода  $\omega_i$  определена его вероятность  $P_i \geq 0$ ,  $i \in \{1, \dots, n\}$ , так, что сумма этих вероятностей равна единице, ведь кроме исходов  $\omega_1, \omega_2, \dots, \omega_n$  произойти больше ничего не может:

$$\sum_{i=1}^n P_i = 1.$$

Обычно, для удобства и краткости исходы и их вероятности записывают следующей таблицей:

$\omega_1$	$\omega_2$	$\dots$	$\omega_n$
$P_1$	$P_2$	$\dots$	$P_n$

Рассмотрим простейший пример. Пусть эксперимент заключается в подбрасывании правильной монетки. Что значит правильной? Это значит, что

как у орла, так и у решки шансы выпасть одинаковы. Значит, в наших обозначениях, эксперимент имеет всего два исхода:  $\omega_1 = \text{Орел}$  (или выпал орел),  $\omega_2 = \text{Решка}$  (выпала решка), причем вероятности этих исходов одинаковы и, соответственно, равны  $\frac{1}{2}$ . Запишем данные в таблицу:

Орел	Решка
$\frac{1}{2}$	$\frac{1}{2}$

Что можно сказать про результат рассматриваемого эксперимента? На самом деле, ничего определенного. Какому исходу отдать предпочтение? Это совершенно непонятно, ведь у них, у этих исходов, одинаковые вероятности. Перед нами очень плохая, с точки зрения информативности, ситуация.

Куда лучше ситуация с фальшивой монеткой, описываемая следующей таблицей

Орел	Решка
$\frac{99}{100}$	$\frac{1}{100}$

Наверное, практически все с уверенностью скажут, что монета выпадет орлом, не так ли? Неопределенность эксперимента уменьшилась разительно. Хотя не стоит себя обнадеживать зря: шансы у решки тоже есть, хоть и маленькие, а значит неопределенность все равно присутствует.

А что если у монетки с двух сторон нарисован орел? Тогда эксперимент описывается такой вот простой таблицей:

Орел
1

и неопределенности нет вообще – мы точно знаем, что выпадет орел. Итак, худшей для нас ситуацией, с точки зрения информативности, является ситуация, когда исходы равновероятны. Лучшей же – когда у одного из исходов вероятность равна единице. Запомним это :)

Давайте обратим внимание еще вот на какой момент. Пусть подбрасывается, скажем, правильный игральный кубик. Тогда исходов 6 (исход  $\omega_i$  означает, что выпало число  $i$ ), их вероятности одинаковы и равны  $\frac{1}{6}$ , и весь эксперимент описывается таблицей

1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Сравните этот эксперимент с тем, когда подбрасывалась правильная монетка

Орел	Решка
$\frac{1}{2}$	$\frac{1}{2}$

Какой из них информативнее? Похоже, что с кубиком дела обстоят куда хуже, ведь теперь кроме того, что все исходы равновозможны, исходов стало больше, а значит неопределенность только увеличилась.

Можно приводить и множество других примеров, но кажется, что идея ясна, не так ли? Давайте теперь разберемся, как же все эти интуитивные представления свести к конкретному числу.

## 2.2 Энтропия Шеннона и проверка ожидаемых свойств

Давайте повторим некоторые рассуждения, проделанные выше, чтобы ввести унифицированные обозначения и определения для дальнейших выкладок. Итак, пусть в результате эксперимента  $\Omega$  могут произойти лишь  $n$  исходов  $\omega_1, \omega_2, \dots, \omega_n$ . Как уже было отмечено, исходы – это мельчайшие, неделимые и несовместные (то есть те, которые не могут произойти одновременно) результаты рассматриваемого эксперимента. Сопоставим каждому исходу  $\omega_i$  его вероятность  $P_i \geq 0$  так, что сумма всех вероятностей равна 1:

$$\sum_{i=1}^n P_i = P_1 + P_2 + \dots + P_n = 1.$$

Тогда приходим к следующему определению.

**Определение 2.2.1** Экспериментом  $\Omega$  назовем произвольное множество исходов  $\omega_1, \omega_2, \dots, \omega_n$ , каждому из которых сопоставлено число  $P_i \geq 0$ ,  $i \in \{1, 2, \dots, n\}$ , называемое вероятностью исхода  $\omega_i$ , причем

$$\sum_{i=1}^n P_i = P_1 + P_2 + \dots + P_n = 1.$$

Ясно, что эксперимент может быть описан (и даже отождествлен с) таблицей следующего вида

$\Omega$	$\omega_1$	$\omega_2$	$\dots$	$\omega_n$
$P$	$P_1$	$P_2$	$\dots$	$P_n$

В первой строке таблицы стоят возможные исходы эксперимента  $\Omega$ , а во второй – вероятности этих исходов. Оказывается, весьма удобной мерой неопределенности эксперимента  $\Omega$ , описываемого таблицей

$\Omega$	$\omega_1$	$\omega_2$	$\dots$	$\omega_n$
$P$	$P_1$	$P_2$	$\dots$	$P_n$

является величина

$$H(\Omega) = - \sum_{i=1}^n P_i \log P_i,$$

где в случае, если  $P_i = 0$ , значение выражения  $P_i \log P_i$  считается равным нулю, что мотивируется тем, что

$$\lim_{x \rightarrow 0^+} x \log x = 0.$$

**Замечание 2.2.1** Конечно, написанная мотивировка – чисто математическая. На интуитивном же уровне понятно, что добавление исхода  $\omega_i$ , которому сопоставляется число  $P_i$  (вероятность этого исхода), равное нулю, не может менять неопределенность эксперимента (а значит и не должно добавлять никаких слагаемых в сумму). Такой исход можно и вовсе не добавлять в таблицу.

Функция  $\log$  в написанных формулах – это логарифм по произвольному основанию, большему единицы. Основание несущественно, ведь по своей сути изменение основания логарифма отвечает изменению единиц измерения величины  $H$  (мы же будем использовать в дальнейшем в качестве основания двойку, но об этом позже). Для удобства, объединим все сказанное в одном определении.

**Определение 2.2.2** Пусть эксперимент  $\Omega$  описывается таблицей

$\Omega$	$\omega_1$	$\omega_2$	$\dots$	$\omega_n$
$P$	$P_1$	$P_2$	$\dots$	$P_n$

Энтропией (или мерой неопределенности)  $H(\Omega)$  эксперимента  $\Omega$  называется величина

$$H(\Omega) = - \sum_{i=1}^n P_i \log P_i,$$

где  $\log$  – логарифм по произвольному основанию, большему единицы, а выражения вида  $0 \log 0$  считаются равными нулю.

Давайте сразу посмотрим, что же получится в уже рассмотренных ранее примерах. Итак, в примере с правильной монеткой эксперимент  $\Omega = \{\text{Орел}, \text{Решка}\}$  описывается таблицей

$\Omega$	Орел	Решка
$P$	$\frac{1}{2}$	$\frac{1}{2}$

Энтропия этого эксперимента вычисляется следующим образом:

$$H(\Omega) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = -\log \frac{1}{2}.$$

Эксперимент с фальшивой монеткой описывается таблицей

$\Omega$	Орел	Решка
P	$\frac{99}{100}$	$\frac{1}{100}$

а значит энтропия этого эксперимента вычисляется, как:

$$H(\Omega) = -\frac{99}{100} \log \frac{99}{100} - \frac{1}{100} \log \frac{1}{100}.$$

Итак, как мы уже упоминали ранее, будем использовать логарифм по основанию 2. Объясним же свой выбор мы чуть позже. Выбрав значение основания логарифма, мы можем вычислить только что написанные выражения: для правильной монетки энтропия будет равна

$$-\log_2 \frac{1}{2} = 1,$$

а для фальшивой

$$-\frac{99}{100} \log_2 \frac{99}{100} - \frac{1}{100} \log_2 \frac{1}{100} \approx 0.081.$$

Смотрите, результаты более чем соответствуют интуитивным представлениям. Там, где интуитивная неопределенность больше, больше и энтропия, и наоборот.

Давайте теперь проверим, выполняются ли наши ожидания не на конкретном примере, а в общем случае. Во-первых, как оказывается, энтропия – величина неотрицательная.

### Теорема 2.2.1

$$H(\Omega) \geq 0.$$

**Доказательство.** Так как  $P_i \in [0, 1]$ , то  $P_i \log P_i \leq 0$ , ведь основание логарифма (по определению энтропии) больше единицы. Тогда и

$$\sum_{i=1}^n P_i \log P_i \leq 0,$$

как сумма неположительных слагаемых. А тогда

$$H(\Omega) = -\sum_{i=1}^n P_i \log P_i \geq 0.$$

□

Итак, мера неопределенности неотрицательна. Видимо, она должна быть равна нулю в том и только том случае, когда неопределенности нет, то есть, когда какой-то исход происходит с вероятностью 1. Это тоже выполняется.

**Теорема 2.2.2** Энтропия равна нулю тогда и только тогда, когда какое-то значение  $P_i$  равно единице, то есть:

$$H(\Omega) = 0 \Leftrightarrow \exists i \in \{1, 2, \dots, n\} : P_i = 1.$$

**Доказательство.** В одну сторону доказательство следует из того, что если  $P_i = 1$ , то все остальные  $P_k$ , при  $k \neq i$  равны нулю, значит и слагаемые, им отвечающие, равны нулю (так как, по соглашению,  $0 \cdot \log 0 = 0$ ). Само же слагаемое с номером  $i$  равно нулю, так как  $1 \cdot \log 1 = 0$ .

Доказательство в обратную сторону проводится тоже не сложно. Достаточно заметить, что функция  $x \log x$  на отрезке  $[0, 1]$  равна нулю только на его концах (в нуле, опять же, по установленному соглашению, а в единице, так как  $\log 1 = 0$ ), а во всех точках интервала  $(0, 1)$  она отрицательна. Значит, чтобы энтропия была равна нулю,  $P_i$  должны принимать значения 0 или 1, но так как  $\sum_{i=1}^n P_i = 1$ , то существует только одно значение  $i$  такое, что  $P_i = 1$ . □

Какими еще свойствами должна обладать энтропия? Исходя из рассмотренных примеров, судя по всему, энтропия должна быть наибольшей в случае, когда все исходы равновероятны (в наших примерах такая ситуация возникала в экспериментах с правильной монеткой и правильным кубиком). Кроме того, чем больше равновероятных исходов, тем энтропия тоже должна быть больше, что мы опять же поняли интуитивно на разобранных примерах.

Оказывается, что введенная функция тоже удовлетворяет этому требованию, итак.

**Теорема 2.2.3** Энтропия  $H(\Omega)$  максимальна в случае, когда все исходы эксперимента равновозможны, то есть когда эксперимент описывается таблицей вида

$\Omega$	$\omega_1$	$\omega_2$	$\dots$	$\omega_n$
$P$	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

В этом случае энтропия равна

$$H(\Omega) = \log n.$$

**Доказательство.** Ясно, что для эксперимента, описываемого таблицей

$$\frac{\Omega}{P} \left| \begin{array}{c|c|c|c|c} \omega_1 & \omega_2 & \dots & \omega_n \\ \hline \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{array} \right.$$

равенство выполнено, ведь

$$H(\Omega) = - \sum_{i=1}^n \frac{1}{n} \log \frac{1}{n} = - \log \frac{1}{n} = \log n.$$

Осталось показать, что это значение действительно максимально. Для этого воспользуемся неравенством Йенсена для выпуклых вниз функций, которое утверждает, что для чисел  $p_1, p_2, \dots, p_n > 0$ , таких, что  $p_1 + p_2 + \dots + p_n = 1$  и любых  $x_1, x_2, \dots, x_n$  из интервала выпуклости функции справедливо:

$$f(p_1x_1 + p_2x_2 + \dots + p_nx_n) \leq p_1f(x_1) + p_2f(x_2) + \dots + p_nf(x_n),$$

или

$$f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i).$$

Рассмотрим функцию  $f(x) = x \log x$ . Так как основание логарифма, согласно договоренности, больше единицы, то эта функция выпукла вниз. Положим в неравенстве Йенсена

$$x_i = P_i, \quad p_i = \frac{1}{n},$$

тогда

$$f\left(\sum_{i=1}^n p_i x_i\right) = f\left(\frac{1}{n} \sum_{i=1}^n P_i\right) = f\left(\frac{1}{n}\right) = -\frac{1}{n} \log n.$$

Кроме того,

$$\sum_{i=1}^n p_i f(x_i) = \frac{1}{n} \sum_{i=1}^n P_i \log P_i.$$

Тогда, в силу неравенства Йенсена,

$$-\frac{1}{n} \log n \leq \frac{1}{n} \sum_{i=1}^n P_i \log P_i \Leftrightarrow -\sum_{i=1}^n P_i \log P_i \leq \log n \Leftrightarrow H(\Omega) \leq \log n.$$

что и требовалось доказать.  $\square$

Оказывается, при некоторых дополнительных предположениях, которые мы оставим за кадром, можно доказать, что представление, предложенное Шенноном, единственno с точностью до положительного сомножителя. Иными словами, если функция удовлетворяет описанным выше трем свойствам и еще кое-чему, чего мы касаться не будем, то она равна

$$-\alpha \sum_{i=1}^n P_i \log P_i$$

для некоторого  $\alpha > 0$ . По сути, это и означает, что основание логарифма в выражении для функции может быть любым, большим единицы.

## 3 Условная энтропия

### 3.1 Наводящие соображения

Как вы, наверное, уже поняли, вероятности исходов экспериментов, которые встречаются в реальной жизни, зачастую неизвестны. Однако, проводя эксперимент и собирая его результаты, у нас накапливается некоторая статистика, которая каждому исходу сопоставляет какое-то, вообще говоря, количественное значение. Например, предположим, что вы проводите опрос, пойдет или нет ваш друг играть в футбол. Данные опроса вы можете наблюдать в таблице:

	Да	Нет
	9	5

По данным из таблицы понятно, что всего было опрошено 14 человек, причем 9 из них приняли приглашение, а пятеро отказались. Тогда имеет смысл оценить вероятность каждого исхода эксперимента, используя, как обычно, частоту. В предложенном примере оценки вероятностей исходов «Да» (играть в футбол) и «Нет» (не играть в футбол) будут равны

$$P(\text{Да}) = \frac{9}{14}, \quad P(\text{Нет}) = \frac{5}{14},$$

а энтропия эксперимента, который описывается таблицей

$\Omega$	Да	Нет
$P$	$\frac{9}{14}$	$\frac{5}{14}$

составленной на основе вычисленных частот, равна

$$H(\Omega) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} \approx 0.94.$$

Энтропия близка к единице, что является максимумом для эксперимента, когда исходов всего два, а основание логарифма равно двум ( $\log_2 2 = 1$ ), а значит ничего определенного об опросе мы сказать не можем.

В рассмотренном эксперименте мы не учитывали ничего, кроме ответа случайно опрошенного человека. Но что, если учитывать какие-то дополнительные факторы? Скажем, составляя таблицу частот, мы теперь будем не просто опрашивать друзей на предмет: пойдет ли он играть в футбол или нет,

но и будем обращать свое внимание на погоду в текущий момент времени. Тогда таблица будет более сложной и, например, может выглядеть следующим образом:

Погода \ Играть в футбол	Да	Нет
Солнечно	6	0
Пасмурно	2	2
Дождь	1	3

Какова теперь энтропия такой системы? Стала ли она меньше за счет новой информации? И вообще, а как вычислить энтропию в такой ситуации? Видимо, разработанного нами ранее аппарата для одного эксперимента уже не хватает. Давайте исправим эту ситуацию.

### 3.2 Определение условной энтропии

Итак, давайте рассмотрим два эксперимента  $\Omega$  и  $\Theta$ , первый из которых состоит из исходов  $\omega_i$ ,  $i \in \{1, 2, \dots, m\}$ , а второй – из исходов  $\theta_j$ ,  $j \in \{1, 2, \dots, n\}$ . Рассматривая пару экспериментов  $(\Omega, \Theta)$ , логично и паре исходов  $(\omega_i, \theta_j)$ ,  $i \in \{1, 2, \dots, m\}$ ,  $j \in \{1, 2, \dots, n\}$  сопоставить вероятность  $P_{ij} \geq 0$  так, чтобы

$$\sum_{i=1}^m \sum_{j=1}^n P_{ij} = 1,$$

снова потому, что никаких других вариантов быть не может (может возникнуть лишь какой-то исход первого эксперимента, и какой-то второго). Оказывается разумным ввести следующее определение.

**Определение 3.2.1** Экспериментом  $(\Omega, \Theta)$  назовем произвольное множество пар исходов  $(\omega_i, \theta_j)$ ,  $i \in \{1, 2, \dots, m\}$ ,  $j \in \{1, 2, \dots, n\}$ , каждой из которых сопоставлено число  $P_{ij} \geq 0$ , называемое вероятностью исхода  $(\omega_i, \theta_j)$ , такое, что

$$\sum_{i=1}^m \sum_{j=1}^n P_{ij} = 1,$$

Ясно, что эксперимент  $(\Omega, \Theta)$  может быть описан (и отождествлен) со следующей таблицей:

$(\Omega, \Theta)$	$\theta_1$	$\theta_2$	$\dots$	$\theta_n$
$\omega_1$	$P_{11}$	$P_{12}$	$\dots$	$P_{1n}$
$\omega_2$	$P_{21}$	$P_{22}$	$\dots$	$P_{2n}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\omega_m$	$P_{m1}$	$P_{m2}$	$\dots$	$P_{mn}$

Так как перед нами тоже эксперимент (просто имеющий  $m \cdot n$  исходов), то энтропия такого эксперимента записывается, как

$$H((\Omega, \Theta)) = - \sum_{i=1}^m \sum_{j=1}^n P_{ij} \log P_{ij}.$$

По написанной таблице понятно, как восстановить эксперименты  $\Omega$  и  $\Theta$  по отдельности. Так, чтобы найти вероятность того, что произошел исход  $\omega_i$ , достаточно сложить все вероятности в таблице, стоящие в  $i$ -ой строке:

$$P(\omega_i) = \sum_{j=1}^n P_{ij}, \quad i \in \{1, 2, \dots, m\},$$

а чтобы найти вероятность, что произошел исход  $\theta_j$ , нужно сложить все вероятности в таблице, стоящие в  $j$ -ом столбце:

$$P(\theta_j) = \sum_{i=1}^m P_{ij}, \quad j \in \{1, 2, \dots, n\}.$$

Итак, мы приходим к следующей теореме.

**Теорема 3.2.1** *Пусть эксперимент  $(\Omega, \Theta)$  задается таблицей:*

$(\Omega, \Theta)$	$\theta_1$	$\theta_2$	$\dots$	$\theta_n$
$\omega_1$	$P_{11}$	$P_{12}$	$\dots$	$P_{1n}$
$\omega_2$	$P_{21}$	$P_{22}$	$\dots$	$P_{2n}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\omega_m$	$P_{m1}$	$P_{m2}$	$\dots$	$P_{mn}$

Тогда эксперименты  $\Omega$  и  $\Theta$  могут быть восстановлены с использованием следующих соотношений:

$$P(\omega_i) = \sum_{j=1}^n P_{ij}, \quad i \in \{1, 2, \dots, m\},$$

$$P(\theta_j) = \sum_{i=1}^m P_{ij}, \quad j \in \{1, 2, \dots, n\}.$$

Давайте сразу рассмотрим пример. Пусть эксперимент  $(\Omega, \Theta)$  заключается в подбрасывании некоторой монеты одним из двух мальчиков: Петей или Степой. Эксперимент  $(\Omega, \Theta)$  описывается следующей таблицей:

$(\Omega, \Theta)$	Петя	Степа
Орел	0.4	0.275
Решка	0.1	0.225

Энтропия этого эксперимента равна

$$H((\Omega, \Theta)) = -0.4 \log_2 0.4 - 0.275 \log_2 0.275 - 0.1 \log_2 0.1 - 0.225 \log_2 0.225 \approx 1.86.$$

Так как максимальная энтропия в данном эксперименте может быть равна  $\log_2 4 = 2$ , то можно заключить, что эксперимент  $(\Omega, \Theta)$  обладает достаточно большой неопределенностью.

Ясно, что как эксперимент  $\Omega$ , так и эксперимент  $\Theta$  состоят из двух исходов:  $\Omega = \{\text{Орел}, \text{Решка}\}$ ,  $\Theta = \{\text{Петя}, \text{Степа}\}$ . Просуммировав значения, стоящие в строках, придем к эксперименту  $\Omega$ , который может быть записан следующей таблицей:

$\Omega$	Орел	Решка
P	0.675	0.325

Можно сделать вывод, что монетка явно не является правильной, так как выпадает орлом со значительно большей вероятностью, чем решкой. Сложив значения, стоящие в столбцах, придем к эксперименту  $\Theta$ , описываемому таблицей:

$\Theta$	Петя	Степа
P	0.5	0.5

Из полученной таблицы можно сделать следующий вывод: Петя и Степа подбрасывают монетку или нет с одинаковыми (равными) вероятностями.

Может оказаться известным, что в эксперименте  $\Theta$  произошло событие  $\theta_j$ , тогда вероятности исходов эксперимента  $\Omega$  меняются, причем понятно каким образом – согласно формуле условной вероятности, которую мы уже встречали в байесовском классификаторе. Вычисляются эти вероятности следующим образом:

$$P(\omega_i | \theta_j) = \frac{P(\omega_i \cap \theta_j)}{P(\theta_j)} = \frac{P_{ij}}{P(\theta_j)}, \quad i \in \{1, 2, \dots, m\}, j \in \{1, 2, \dots, n\},$$

Может быть и противоположная ситуация, когда известно, что произошел исход  $\omega_i$  эксперимента  $\Omega$ . Тогда вероятности исходов эксперимента  $\Theta$  вычисляются так:

$$P(\theta_j | \omega_i) = \frac{P(\omega_i \cap \theta_j)}{P(\omega_i)} = \frac{P_{ij}}{P(\omega_i)}, \quad i \in \{1, 2, \dots, m\}, j \in \{1, 2, \dots, n\}.$$

Понятно, что для экспериментов с новыми вероятностями можно тоже вычислить энтропию. Например,

$$H(\Omega|\theta_j) = - \sum_{i=1}^m P(\omega_i|\theta_j) \log P(\omega_i|\theta_j) = - \sum_{i=1}^m \frac{P_{ij}}{\sum_{i=1}^m P_{ij}} \log \frac{P_{ij}}{\sum_{i=1}^m P_{ij}}$$

и

$$H(\Theta|\omega_i) = - \sum_{j=1}^n P(\theta_j|\omega_i) \log P(\theta_j|\omega_i) = - \sum_{j=1}^n \frac{P_{ij}}{\sum_{j=1}^n P_{ij}} \log \frac{P_{ij}}{\sum_{j=1}^n P_{ij}}.$$

Эти энтропии называются *условными энтропиями*.

**Определение 3.2.2** Условной энтропией эксперимента  $\Omega$  при условии, что эксперимент  $\Theta$  оказался в состоянии  $\theta_j$  (то есть если произошел исход  $\theta_j$ ),  $j \in \{1, 2, \dots, n\}$ , называется величина

$$H(\Omega|\theta_j) = - \sum_{i=1}^m P(\omega_i|\theta_j) \log P(\omega_i|\theta_j).$$

**Замечание 3.2.1** Ясно, что совершенно аналогичным образом (просто ввиду симметрии) определяется условная энтропия эксперимента  $\Theta$  при условии, что эксперимент  $\Omega$  оказался в состоянии  $\omega_i$ .

Итак, давайте вернемся к примеру с монетой и посмотрим, как изменится энтропия. Напомним таблицу, которой описывается эксперимент  $(\Omega, \Theta)$ :

$(\Omega, \Theta)$	Петя	Степа
Орел	0.4	0.275
Решка	0.1	0.225

Пусть известно, что монету бросал Петя, тогда

$$P(\text{Орел}|\text{Петя}) = \frac{P(\text{Орел} \cap \text{Петя})}{P(\text{Петя})} = \frac{0.4}{0.5} = 0.8,$$

и

$$P(\text{Решка}|\text{Петя}) = \frac{P(\text{Решка} \cap \text{Петя})}{P(\text{Петя})} = \frac{0.1}{0.5} = 0.2.$$

Запишем полученные результаты в виде таблицы:

$$\begin{array}{c|c|c} (\Omega, \theta_1) = (\Omega|\text{Петя}) & (\text{Орел}|\text{Петя}) & (\text{Решка}|\text{Петя}) \\ \hline P & 0.8 & 0.2 \end{array}.$$

Условная энтропия эксперимента  $\Omega$  при условии, что бросал монету Петя, составит:

$$H(\Omega|\theta_1) = H(\Omega|\text{Петя}) = -0.8 \log_2 0.8 - 0.2 \log_2 0.2 \approx 0.72.$$

Проделав аналогичные расчеты в ситуации, когда бросает монету Степа, приходим к следующей таблице:

$(\Omega, \theta_2) = (\Omega \text{Степа})$	$(\text{Орел} \text{Степа})$	$(\text{Решка} \text{Степа})$
P	0.55	0.45

Условная энтропия же в этом случае равна

$$H(\Omega|\theta_2) = H(\Omega|\text{Степа}) = -0.55 \log_2 0.55 - 0.45 \log_2 0.45 \approx 0.99.$$

Итак, условная энтропия в случае, когда монетку бросает Петя меньше, чем в случае, когда ее бросает Степа. Полученные значения, опять-таки, подтверждают наши интуитивные ожидания – достаточно сравнить таблицы

$(\Omega, \theta_1) = (\Omega \text{Петя})$	$(\text{Орел} \text{Петя})$	$(\text{Решка} \text{Петя})$
P	0.8	0.2

$(\Omega, \theta_2) = (\Omega \text{Степа})$	$(\text{Орел} \text{Степа})$	$(\text{Решка} \text{Степа})$
P	0.55	0.45

Эксперимент в случае, когда монетку бросает Петя, имеет много меньшую неопределенность (вероятности исходов сильно отличаются), а значит факт того, что монету бросает Петя, дает нам несколько больше информации.

Остался последний рывок. По большому счету, так как каждый исход эксперимента  $\Omega$  и  $\Theta$  происходит с какой-то вероятностью, то и условная энтропия принимает свои значения с некоторой вероятностью. Иными словами, значение  $H(\Omega|\theta_1)$  принимается с вероятностью  $P(\theta_1)$ , значение  $H(\Omega|\theta_2)$  принимается с вероятностью  $P(\theta_2)$ , и так далее. Так как  $\Theta$  – эксперимент, то сумма вероятностей элементарных исходов равна единице, а значит условная энтропия  $H(\Omega|\theta_j)$  является случайной величиной с рядом распределения

$H(\Omega \theta_j)$	$H(\Omega \theta_1)$	$\dots$	$H(\Omega \theta_n)$
P	$P(\theta_1)$	$\dots$	$P(\theta_n)$

**Определение 3.2.3** Случайная величина  $H(\Omega|\theta_j)$ , заданная выше, называется условной энтропией эксперимента  $\Omega$  при условии, что произошел эксперимент  $\Theta$ .

Оказывается, всю систему хорошо характеризует так называемая полная условная энтропия.

**Определение 3.2.4** Полной условной энтропией эксперимента  $\Omega$  при условии, что произошел эксперимент  $\Theta$ , называется величина

$$H(\Omega|\Theta) = E(H(\Omega|\theta_j)) = \sum_{j=1}^n P(\theta_j)H(\Omega|\theta_j).$$

Снова вернемся к примеру с монетой, описанному нами ранее. Вычисленные значения условной энтропии события  $(\Omega, \theta)$  с соответствующими вероятностями запишем в виде ряда распределения

$H(\Omega \theta_j)$	$H(\Omega \text{Петя})$	$H(\Omega \text{Степа})$
$P$	$P(\text{Петя})$	$P(\text{Степа})$

В итоге приходим к следующей таблице:

$H(\Omega \theta_j)$	0.72	0.99
$P$	0.5	0.5

В результате, полная условная энтропия составит

$$H(\Omega|\Theta) = 0.72 \cdot 0.5 + 0.99 \cdot 0.5 = 0.855.$$

Давайте теперь посмотрим, что произошло с энтропией эксперимента  $\Omega$ . В первоначальной, самой общей ситуации, так как эксперимент  $\Omega$  описывается таблицей вида

$\Omega$	Орел	Решка
$P$	0.675	0.325

то его энтропия равнялась  $H(\Omega) = 0.910$ . В то же время, знание того, что происходит с экспериментом  $\Theta$  уменьшило энтропию, так как полная условная энтропия стала равна  $H(\Omega|\Theta) = 0.855$ .

### 3.3 Пример на вычисление условной энтропии

Давайте посмотрим, что же получится в уже озвученном ранее примере. Напомним, что друзьям задавался вопрос: «пойдешь играть в футбол или нет?». При этом фиксировался не только ответ конкретного человека, но и погода на момент опроса. Результаты представлены в следующей таблице:

Погода \ Играть в футбол (Ответ)	Да	Нет
Солнечно	6	0
Пасмурно	2	2
Дождь	1	3

Давайте найдем полную условную энтропию такой системы, а именно  $H(\text{Ответ}|\text{Погода})$ , но для начала, по заполненной таблице составим таблицу вероятностей эксперимента (Погода, Ответ), используя, как обычно, частоту

(Погода, Ответ)	Да	Нет
Солнечно	$\frac{6}{14}$	0
Пасмурно	$\frac{2}{14}$	$\frac{2}{14}$
Дождь	$\frac{1}{14}$	$\frac{3}{14}$

Каждое значение в представленной таблице получено, как отношение количества появлений конкретного исхода, к общему числу исходов. Например, вероятность события, что «друг пойдет играть в футбол, а одновременно идет дождь» составит:

$$P(\text{Ответ} = \text{Да} \cap \text{Погода} = \text{Дождь}) = \frac{1}{14}.$$

То есть мы просто посмотрели, сколько друзей согласилось играть в футбол в дождь, и разделили на общее число исходов.

Перейдем к вычислениям условной энтропии, а для этого зафиксируем некоторое состояние события погода. Пусть, например, идет дождь. Легко найти вероятность того, что идет дождь. Она равна

$$P(\text{Погода} = \text{Дождь}) = \frac{4}{14},$$

так как дождь шел в 4 случаях из 14. Можно рассуждать и с помощью разработанной нами теории:

$$\begin{aligned} P(\text{Погода} = \text{Дождь}) &= P(\text{Погода} = \text{Дождь} | \text{Ответ} = \text{Да}) + \\ &+ P(\text{Погода} = \text{Дождь} | \text{Ответ} = \text{Нет}) = \frac{1}{14} + \frac{3}{14} = \frac{4}{14}. \end{aligned}$$

Теперь можно вычислить условную вероятность события, что человек согласится играть в футбол при условии, что идет дождь. Эта вероятность равна

$$\begin{aligned} P(\text{Ответ} = \text{Да} | \text{Погода} = \text{Дождь}) &= \\ &= \frac{P(\text{Ответ} = \text{Да} \cap \text{Погода} = \text{Дождь})}{P(\text{Погода} = \text{Дождь})} = \frac{\frac{1}{14}}{\frac{4}{14}} = \frac{1}{4} = 0.25, \end{aligned}$$

а так как исхода всего два, то

$$\begin{aligned} P(\text{Ответ} = \text{Нет} | \text{Погода} = \text{Дождь}) &= \\ &= 1 - P(\text{Ответ} = \text{Да} | \text{Погода} = \text{Дождь}) = 0.75. \end{aligned}$$

На основании полученных условных вероятностей, мы теперь можем найти условную энтропию эксперимента «Ответ» при условии, что эксперимент «Погода» находится в состоянии «Дождь», а именно

$$H(\text{Ответ}|\text{Погода} = \text{Дождь}) = -0.25 \log_2 0.25 - 0.75 \log_2 0.75 \approx 0.81.$$

Действуя аналогичным образом, найдем условные энтропии эксперимента «Ответ» в случаях, когда эксперимент «Погода» оказывается в состояниях «Солнечно» и «Пасмурно»:

$$H(\text{Ответ}|\text{Погода} = \text{Солнечно}) = -1 \log_2 1 - 0 \log_2 0 = 0,$$

$$H(\text{Ответ}|\text{Погода} = \text{Пасмурно}) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1.$$

Итак, мы нашли условные энтропии, а значит можем составить следующий ряд распределения для условной энтропии

$H(\text{Ответ} \text{Погода})$	0	0.81	1
P	$\frac{6}{14}$	$\frac{4}{14}$	$\frac{4}{14}$

Математическое ожидание полученной случайной величины характеризует полную условную энтропию эксперимента «Ответ» относительно эксперимента «Погода». Согласно определению, получаем

$$E(H(\text{Ответ}|\text{Погода})) = 0 \cdot \frac{6}{14} + 0.81 \cdot \frac{4}{14} + 1 \cdot \frac{4}{14} \approx 0.517,$$

что говорит нам о «средней» неопределенности системы. Этот результат, в общем-то, понятен и интуитивно, если взглянуть на исходную таблицу

Погода \ Играть в футбол (Ответ)	Да	Нет
Солнечно	6	0
Пасмурно	2	2
Дождь	1	3

Смотрите, если солнечно, никто не отказывается поиграть в футбол, а если дождливо, большая часть друзей останется дома. В пасмурную же погоду ответ друзей разделился поровну.

### 3.4 Энтропия и прирост информации

Итак, что же мы узнали? Мы узнали, что энтропия показывает степень неопределенности состояния некоторого эксперимента или системы. Чем

больше мы знаем о системе, тем ее состояние становится менее неопределенным. Именно по этой причине удобно измерять изменение наших знаний о системе, отслеживая изменение энтропии.

За счет чего возможно уменьшить энтропию системы или, что то же самое, получить прирост информации? Конечно, за счет дополнительных знаний о системе. Достаточно, чтобы произошло некое событие, предшествующее проведению эксперимента, событие, которое в некотором смысле доопределяет эксперимент.

На самом деле мы уже проделали все необходимые нам вычисления и выкладки, только не записали итоговое выражение. Давайте вспомним, изначально мы рассмотрели эксперимент, заданный таблицей

	Да	Нет
9		5

и, по сути, ничего не знали о системе, кроме ответов на вопрос. В таком эксперименте мы получили энтропию, равную

$$H(\text{Ответ}) \approx 0.94.$$

Потом же, исходя из второго события – погоды, мы получили новые сведения, и полная условная энтропия эксперимента «Ответ» относительно эксперимента «Погода» стала равна

$$H(\text{Ответ}|\text{Погода}) \approx 0.52.$$

Какой же прирост информации мы получили? Думаем вы догадались, что прирост логично определить следующим образом:

$$H(\text{Ответ}) - H(\text{Ответ}|\text{Погода}) = 0.42.$$

Резюмируя, введем определение.

**Определение 3.4.1** Приростом информации (*information gain*) называется величина

$$IG(\Omega|\Theta) = H(\Omega) - H(\Omega|\Theta).$$

Отметим, что прирост информации всегда неотрицателен. Это следует и из здравого смысла, и из того, что, как легко проверить,

$$H(\Omega) \geq H(\Omega|\Theta)$$

при любом эксперименте  $\Theta$ .

### 3.5 Вычисление прироста информации на конкретном примере

Рассмотрим еще один пример, когда экспериментов (или событий) не два, а больше, и найдем прирост информации в зависимости от различных событий. Итак, представьте себе выставку кошек, которых оценивают по ряду критериев, и в результате оценивания, кошкам, которые проходят по критериям, дают медаль с надписью «Чемпион». Перед вами таблица, содержащая информацию о породе кошек, цвете шерсти и росте в холке, которые влияют на привлекательность.

Порода	Шерсть	Рост	Привлекательность
Британец	Белый	Высокий	Нет
Британец	Серый	Высокий	Да
Британец	Белый	Низкий	Да
Мейн-кун	Белый	Высокий	Нет
Мейн-кун	Коричневый	Высокий	Да
Мейн-кун	Коричневый	Низкий	Нет
Рэгдолл	Серый	Высокий	Да
Мейн-кун	Серый	Низкий	Да

Интересно, какой из критериев является самым информативным. На что больше всего обращали внимания судьи?

Представьте для начала, что вы, как посетитель выставки, видите лишь результаты колонки привлекательность, в которой говорится, какая кошка привлекательна, а какая – нет. Рассчитаем энтропию  $H$  эксперимента (события)  $\Omega$  «Привлекательность», а для этого перенесем данные в привычную таблицу:

$$\begin{array}{c|c} \text{Да} & \text{Нет} \\ \hline 5 & 3 \end{array}.$$

Составим таблицу эксперимента  $\Omega$ , она имеет следующий вид:

$$\begin{array}{c|c|c} \Omega & \text{Да} & \text{Нет} \\ \hline P & \frac{5}{8} & \frac{3}{8} \end{array}.$$

Энтропия этого эксперимента легко вычисляется и равна

$$H(\Omega) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} \approx 0.954.$$

Так как максимальная энтропия рассмотренного эксперимента может быть равна  $\log_2 2 = 1$ , то вывод очевиден: сплошная неопределенность. Оказывается, что рассмотрение только финальных оценок судей – дело совершенно неинформативное.

Пусть теперь вы видите еще и породу кошек (эксперимент  $\Theta$ ), тогда таблица имеет следующий вид

Порода \ Привлекательность	Да	Нет
Британец	2	1
Мейн-кун	2	2
Рэгдолл	1	0

Попробуем понять, насколько для нас информативен признак «Порода», а для этого вычислим полную условную энтропию и прирост информации. Сначала рассчитаем условные вероятности, они равны:

$$\mathsf{P}(\text{Привлекательность} = \text{Да} | \text{Порода} = \text{Британец}) = \frac{2}{3},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Нет} | \text{Порода} = \text{Британец}) = \frac{1}{3},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Да} | \text{Порода} = \text{Мейн-кун}) = \frac{1}{2},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Нет} | \text{Порода} = \text{Мейн-кун}) = \frac{1}{2},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Да} | \text{Порода} = \text{Рэгдолл}) = 1$$

$$\mathsf{P}(\text{Привлекательность} = \text{Нет} | \text{Порода} = \text{Рэгдолл}) = 0.$$

Теперь вычислим условные энтропии.

$$\mathsf{H}(\text{Привлекательность} | \text{Порода} = \text{Британец}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \approx 0.918,$$

$$\mathsf{H}(\text{Привлекательность} | \text{Порода} = \text{Мейн-кун}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1,$$

$$\mathsf{H}(\text{Привлекательность} | \text{Порода} = \text{Рэгдолл}) = -\frac{1}{1} \log_2 \frac{1}{1} = 0.$$

Тогда полная условная энтропия равна

$$\mathsf{H}(\text{Привлекательность} | \text{Порода}) = \frac{3}{8} \cdot 0.918 + \frac{4}{8} \cdot 1 + \frac{1}{8} \cdot 0 = 0.844.$$

Таким образом, зная породу, мы получаем прирост информации, равный

$$|G(\text{Привлекательность} | \text{Порода})| = 0.954 - 0.844 = 0.110.$$

Много это или мало – познается в сравнении. Предлагаем вам посчитать самостоятельно значения прироста информации, если вместо породы будут

данные о шерсти или росте в холке кошек. Результаты должны получиться следующие:

$$\text{IG}(\text{Привлекательность}|\text{Шерсть}) \approx 0.360,$$

$$\text{IG}(\text{Привлекательность}|\text{Рост}) \approx 0.003.$$

Полученные значения прироста информации показывают нам, что критерий «Шерсть» является самым информативным. Иначе говоря, результаты строгого жюри, выдавшего кошкам медали, лучше всего соотносятся с тем, какого цвета у кошек шерсть. Приведенный пример, конечно, весьма игрушечный, но надеемся, что идея ясна.

Кроме того, надеемся ясна и идея того, зачем мы так подробно разобрали понятие энтропии и прироста информации. Перейдем теперь к нашей цели – построению деревьев принятия решений.

## 4 Деревья принятия решений

Итак, мы рассмотрели способ измерения прироста информации с использованием понятия энтропии. Теперь мы готовы перейти к построению дерева принятия решений, однако для начала неплохо бы определиться: а что это такое?

### 4.1 Немного о самих деревьях

Если вы знакомы с теорией графов, то определение дерева принятия решений может быть дано следующим образом.

**Определение 4.1.1** *Дерево принятия решений – это дерево, то есть ациклический связный граф, имеющее следующие метки:*

- *узлы дерева, не являющиеся листьями – это атрибуты.*
- *в листьях дерева находятся отклики – результаты классификации.*
- *на ребрах находится правило – значение узла (атрибута), из которого исходит ребро.*

Даже если вы не знакомы с теорией графов, то введенное понятие прекрасно иллюстрируются на уже знакомом нам дереве принятия решений, цель которого – классифицировать предложения о работе на два типа: те, которые стоит принять, и те, которые стоит отклонить. Как видно из рисунка, из верхнего узла дерева, отвечающего за предиктор (или атрибут) «размер зарплаты», выходят три ребра. Если значение узла меньше, чем 50000, то мы сразу попадаем в лист: предложение предлагается отклонить. Если значение узла больше или равно, чем 100000, то мы снова попадаем в лист, но уже

с другим откликом: предложение предлагается принять. Если же значение узла находится в диапазоне от 50000 включительно до 100000 не включительно, то ребро нас ведет в следующий узел, который отвечает за время, требующееся на дорогу до работы. В зависимости от значения этого узла, возникают дополнительные ветвления, или ветки дерева, которые вы легко отследите самостоятельно. Обратите внимание, любой путь по построенному дереву заканчивается листом, и в каждый лист можно попасть из начальной (корневой) вершины дерева – это и есть требование связности в определении. Ацикличность же, или отсутствие циклов, обеспечивается тем, что как в каждый узел (кроме изначального), так и в каждый лист входит лишь одно ребро.

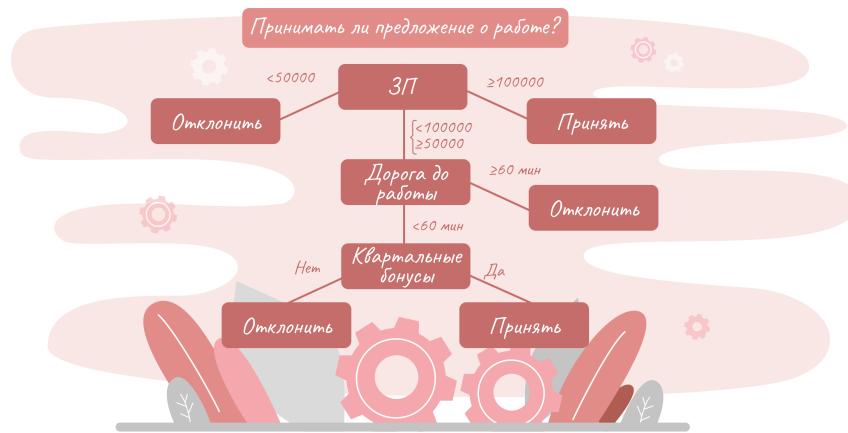


Рис. 2: Пример дерева решений.

**Определение 4.1.2** Глубиной конкретного листа (или узла) в дереве называется количество ребер, которые соединяют его с корневой вершиной. Глубиной дерева называют максимум из глубин его листов.

В нашем примере всего 5 листов. Глубина каждого из них равна (слева-направо): 1, 3, 3, 2 и 1, соответственно, а значит глубина дерева равна 3.

Разобравшись с тем, а что такое дерево принятия решений, а также как измерять информативность признаков, перейдем к рассмотрению конкретных алгоритмов построения деревьев принятия решений.

## 4.2 Алгоритм построения дерева и пример с кошками

Составим алгоритм построения дерева принятия решений на тренировочном наборе данных  $x_1, x_2, \dots, x_n$ , состоящем из  $n$  элементов с  $p$  предикторами  $X_1, X_2, \dots, X_p$  каждый, и откликом  $Y$ .

1. Пусть на вход подается множество объектов  $X$ . Среди  $p$  предикторов выбрать тот, для которого прирост информации максимальен. Итак, решается задача

$$\arg \max_{Q \in \{X_1, X_2, \dots, X_p\}} I(G(Y|Q))$$

2. Пусть выбран предиктор  $X_i$ , принимающий на наборе данных  $X$  ровно  $t$  уникальных значений. Выполнить разделение набора данных  $X$  на подмножества  $S_1, \dots, S_t$  по уникальным значениям предиктора  $X_i$ .
3. Для каждого множества  $S_i$ ,  $i \in \{1, 2, \dots, t\}$ , если энтропия по отклику не равна нулю, повторить шаги 1 и 2.

Если использовать введенную выше терминологию деревьев принятия решений, то выбранный на первом шаге алгоритма предиктор  $X_i$  – это узел дерева, а разделение множества объектов на  $t$  подмножеств по уникальному значению предиктора – это установление  $t$  ребер из узла  $X_i$ . Равенство же нулю энтропии для некоторого множества  $S_i$  означает, что установленное ребро, в результате которого получилось множество  $S_i$ , указывает на лист, которому нужно присвоить значение откликов объектов из  $S_i$  (ведь раз энтропия ноль, то все объекты имеют одинаковый отклик).

Опробуем алгоритм на уже озвученном примере с кошками. Итак, таблица исходных данных в наших обозначениях такова:

№	Порода ( $X_1$ )	Шерсть ( $X_2$ )	Рост ( $X_3$ )	Привлекательность ( $Y$ )
$x_1$	Британец	Белый	Высокий	Нет
$x_2$	Британец	Серый	Высокий	Да
$x_3$	Британец	Белый	Низкий	Да
$x_4$	Мейн-кун	Белый	Высокий	Нет
$x_5$	Мейн-кун	Коричневый	Высокий	Да
$x_6$	Мейн-кун	Коричневый	Низкий	Нет
$x_7$	Рэгдолл	Серый	Высокий	Да
$x_8$	Мейн-кун	Серый	Низкий	Да

Иными словами, каждая строчка таблицы – это тренировочный объект, обладающий тремя предикторами  $X_1, X_2, X_3$  – «Порода», «Шерсть» и «Рост», соответственно, и откликом  $Y$  – «Привлекательность».

Первый шаг алгоритма нами уже был проделан ранее. Согласно разобранному примеру, самым информативным оказался критерий «Шерсть», или предиктор  $X_2$ . Так как  $X_2$  принимает три уникальных значения: «Белый», «Серый» и «Коричневый», то, согласно второму шагу алгоритма, весь набор тренировочных данных нужно разделить на 3 группы (по значению

$X_2$ ). Пусть первая группа  $S_{\text{Коричневый}}$  – группа кошек с коричневой шерстью, тогда в нее входят  $x_5, x_6$ . Во вторую группу  $S_{\text{Белый}}$ , с белой шерстью, входят  $x_1, x_3, x_4$ . В последнюю же группу  $S_{\text{Серый}}$ , с серой шерстью, входят все остальные, то есть  $x_2, x_7, x_8$ . Давайте это визуализируем.

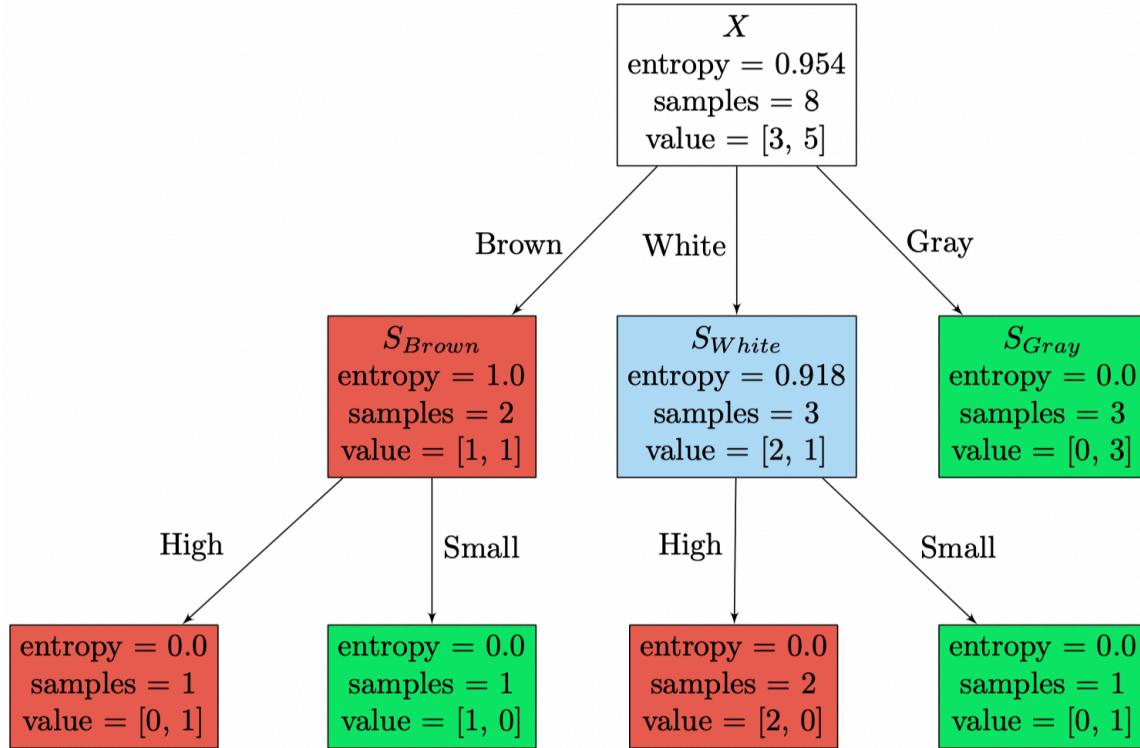


Рис. 3: Дерево принятия решений для задачи о кошках.

Посмотрите на рисунок 3 (он несколько отличается по визуализации и информативности от ранее описанных деревьев). В верхнем белом блоке (корневом узле дерева) содержится информация об исходном наборе данных  $X = \{x_1, x_2, \dots, x_n\}$ . Во второй строке этого блока указана энтропия – это энтропия, вычисленная по отклику  $Y$  в исходном наборе данных. Она, как нами было показано, примерно равна 0.954. В третьей строке указано количество элементов исследуемого набора данных, а в четвертой – соотношение между привлекательными и не привлекательными кошками в исходном наборе данных (непривлекательных 3, привлекательных 5).

Так как «Шерсть» – самый информативный признак на первом этапе, принимающий 3 различных значения, то из корневого узла выходит три ребра, и появляется второй уровень дерева, состоящий из трех групп:  $S_{\text{Коричневый}}$ ,  $S_{\text{Белый}}$ ,  $S_{\text{Серый}}$ , отвечающих трем различным узлам. Так как среди кошек с коричневой шерстью 1 привлекательная и 1 нет, то энтропия по отклику «Привлекательность» в наборе данных  $S_{\text{Коричневый}}$  равна 1. Аналогично, так как среди кошек с белой шерстью 2 непривлекательных и 1 привлекательная,

то энтропия по отклику «Привлекательность» в наборе данных  $S_{\text{Белый}}$  равна

$$-\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} \approx 0.918.$$

Среди кошек с серой шерстью все три оказались привлекательными, а потому энтропия по отклику «Привлекательность» в группе  $S_{\text{Серый}}$  равна 0, и дальнейшее разбиение на подмножества группы  $S_{\text{Серый}}$ , в отличие от остальных двух групп, не требуется, а узел, соответствующий этому блоку, становится листом с откликом «Да».

В других же подмножествах остается неопределенность, а значит, в соответствии с алгоритмом, нужно определить новые критерии разбиения для каждого из них. Если внимательно посмотреть на подмножества  $S_{\text{Коричневый}}$  и  $S_{\text{Белый}}$  (ну или если просто посчитать), то становится понятно, что наиболее информативным критерием в каждой группе будет рост в холке. Отсюда и разделение на третьем уровне дерева. Легко видеть, что теперь каждое образовавшееся подмножество (их всего 4) имеет нулевую энтропию, значит алгоритм завершен, а все узлы последнего уровня становятся листьями с откликами (слева-направо) «Да», «Нет», «Нет» и «Да», соответственно.

Классификация новых объектов, имея построенное дерево, выполняется крайне просто. Новый объект сначала проходит классификацию по верхнему уровню (в нашем примере – по цвету шерсти). Если цвет шерсти серый, то дерево решений говорит, что кошка точно понравится судьям. Если же цвет коричневый или белый, то далее нужно смотреть на рост кошки (на следующий уровень). Согласно дереву, коричневая высокая или белая низкая кошка понравится судьям, а коричневая низкая или белая высокая – не понравится.

**Замечание 4.2.1** Последний пункт описанного в начале алгоритма – это лишь одно из возможных условий остановки процесса деления набора тренировочных данных на группы (и, как следствие, одно из возможных условий остановки построения ДПР). Ситуация, когда энтропия каждой получившейся группы равна нулю означает, что в каждой группе все элементы имеют один и тот же отклик, а значит понятно и как эту группу «охарактеризовать» – по отклику. В то же время, особенно на больших объемах данных, разделение до нулевой энтропии – дело достаточно трудоемкое, зачастую приводящее к переобучению. Поэтому часто используют следующие критерии остановки:

- ограничение глубины дерева. Узлы, имеющие максимальную установленную глубину, становятся листами;
- ограничение минимального количества элементов в группе. Если при разделении рассматриваемого множества на подмножества по уни-

кальным значениям предиктора, получающиеся группы тренировочных данных содержат меньше элементов, чем задано исследователем, то деление останавливается. В терминах ДПР это означает, что текущий узел становится листом;

- достижение в группе какого-то заданного критерия: например, значения неопределенности. В терминах ДПР это опять же означает, что текущий узел становится листом.

В каждом из описанных случаев, однако, остается открытым вопрос: какое же значение приписать листу? Обычно ему приписывают тот отклик, чьих представителей в соответствующей группе тренировочных данных большинство. Подробнее об этом мы поговорим чуть позже.

### 4.3 Бинарное дерево решений

Рассмотренный алгоритм построения дерева принятия решения удобен в том случае, когда мы работаем с предикторами, все возможные значения которых присутствуют в тренировочных данных. Если тестовый объект обладает значением предиктора, которого не было в тренировочных данных, то описанный алгоритм классификации, очевидно, не работает. Скажем, пусть нас интересует следующий вопрос: понравится ли судьям черная невысокая кошка? Взглянув на построенное ранее дерево (рисунок 3), мы понимаем, что уже на первом уровне наша классификация ломается: по какой ветке идти? Значение признака «Шерсть» не подходит ни под один из приведенных вариантов.

Озвученная проблема является, на самом деле, весьма серьезной. Дело в том, что нередко мы не знаем все возможные значения, которые может принимать атрибут рассматриваемых объектов, а часто бывает и так, что этих значений и вообще – бесконечное число (особенно, если значения атрибутов – числовые). Примерами числовых атрибутов, принимающих бесконечное число значений, могут служить: рост человека, текущее время, дата, масса товара на весах на рынке и многое-многое другое. Если же атрибут не числовой, он редко может принимать бесконечное число значений, но может принимать «очень много» значений. Примером могут служить названия городов по всему миру.

Кроме того, чем больше значений атрибутов мы имеем, тем дольше строится и само решающее дерево. К тому же, оно становится очень «ветвистым», что мешает как интерпретации модели в целом, так и способствует ее, модели, переобучению. Обозначенных причин, наверное, достаточно, чтобы понять, что требуется разработать какой-то новый подход к построению решающих деревьев, обладающий следующими преимуществами:

- возможность классификации объектов с ранее (то есть при обучении) не встречавшимися значениями предикторов;
- возможность объединять (группировать, укрупнять) множество значений признаков.

Выполнение последнего требования, как мы уже сказали, способствует уменьшению количества ветвлений и, как следствие, более простой интерпретируемости и более высокой скорости работы модели.

Все эти размышления приводят нас к так называемым бинарным деревьям принятия решений, или к бинарным решающим деревьям.

**Определение 4.3.1** *Бинарное дерево принятия решений – это дерево принятия решений, из каждого узла которого выходит ровно два ребра.*

Итак, бинарные деревья принятия решений – это все те же деревья принятия решений, главная особенность которых состоит в том, что разделение осуществляется всегда на два подмножества, в каждом из которых заново рассчитывается энтропия и по проделанным расчетам определяется наиболее информативный критерий дальнейшего разбиения (если это разбиение требуется). Разделение же осуществляется по принципу: если значения атрибута удовлетворяют «установленному» условию, то объект относится к первому множеству, а если не удовлетворяют – ко второму. Таким образом, объекты, значения признаков которых ранее не встречались, все равно будут отнесены к одному из двух возможных подмножеств, а значит в итоге будут классифицированы. Но как проводить разделение? Что это за «установленные условия»?

## 4.4 Типы признаков и их группировка

Начнем с некоторых важных определений. Еще раз отметим, что мы работаем с тренировочным набором данных  $x_1, x_2, \dots, x_n$  объема  $n$ , каждый объект которого обладает  $p$  атрибутами

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}).$$

**Определение 4.4.1** *Признак объекта называется категориальным, если он принимает конечное число значений.*

Категориальные признаки могут быть как числовыми, так и не числовыми. В рассмотренном примере с кошками все признаки – категориальные. Признак «Шерсть» принимает какое-то значение из конечного множества {Белый, Серый, Коричневый}, а «Рост» – из конечного множества {Высокий, Низкий}. Кроме категориальных, оказывается удобным выделять и некатегориальные признаки.

**Определение 4.4.2** Если признак объекта не является категориальным, то мы будем называть его некатегориальным.

Примерами некатегориальных признаков, как мы уже упоминали ранее, являются: рост, текущее время, масса товара на весах в магазине и многие другие. Некатегориальные признаки чаще всего являются числовыми.

Теперь давайте решим, как можно объединять признаки в группы. Для некатегориального числового признака  $X_i$  часто поступают вот таким образом: множество его значений делят на два множества:

первое множество:  $X_i \leq C$ , второе множество:  $X_i > C$ ,

причем число  $C$  (или так называемое пороговое значение) подбирается отдельно.

**Замечание 4.4.1** Давайте чуть подробнее поговорим про выбор числа  $C$ . Как этот выбор осуществляется? Тут возможно много вариантов, мы же приведем лишь несколько.

1. *Первый вариант такой. Для начала определяют множество значений (обычно отрезок)  $[a, b]$  некатегориального числового признака  $X_i$  либо исходя из каких-то теоретических соображений, либо исходя из значений этого признака у тренировочных данных. В последнем случае в качестве  $a$  часто берут наименьшее значение рассматриваемого признака среди тренировочных данных, а в качестве  $b$  – наибольшее. Следующий этап – это задание шага  $h > 0$  изменения значения  $C$ . При выбранном шаге  $h$ , в качестве значений  $C$  разумно рассмотреть следующие:*

$$\{a, a + h, \dots, a + (k - 1)h, a + kh\}, \quad k \in \mathbb{Z},$$

*где  $a + (k - 1)h < b$ , но  $a + kh \geq b$ . Для каждого из выбранных значений вычисляют прирост информации, отвечающий полученному разделению на два класса:  $X_i \leq C$  и  $X_i > C$ .*

2. *Второй вариант похож на первый. Значения  $C$  выбираются таким образом, чтобы при каждом следующем значении  $C$  к множеству  $X_i \leq C$  добавлялось ровно одно новое значение предиктора  $X_i$ . Формально это можно описать следующим образом. Пусть  $\{x_{1i}, x_{2i}, \dots, x_{ti}\}$ ,  $1 \leq t \leq n$  – множество всех возможных различных значений предиктора  $X_i$  на тренировочных данных объема  $n$ . Предположим также, что они упорядочены по возрастанию, то есть*

$$x_{1i} < x_{2i} < \dots < x_{ti}.$$

Тогда в качестве значений  $C$  разумно взять следующие:

$$\{C_1, C_2, \dots, C_t\},$$

где  $C_t \geq x_{ti}$ ,  $x_{ji} \leq C_j < x_{(j+1)i}$ ,  $j \in \{1, 2, \dots, (t-1)\}$ .

В итоге выбирают то разделение, при котором прирост информации наибольший.

Если же признак  $X_i$  является категориальным и принимает значения из множества  $M = \{x_{1i}, x_{2i}, \dots, x_{ti}\}$ ,  $t \in \{1, 2, \dots, n\}$ , то его множество значений можно разделить на две части, например, так:

первое множество:  $X_i \in M_1$ , второе множество:  $X_i \in M_2$ ,

где  $M_1 \cap M_2 = \emptyset$ ,  $M_1 \cup M_2 = M$ ,  $M_j \neq \emptyset$ ,  $j \in \{1, 2\}$ . Иными словами, все множество  $M$  уникальных значений признака  $X_i$  разбивается на 2 непустые непересекающиеся части.

**Замечание 4.4.2** В библиотеках, реализованных в различных инструментах, чаще всего множество  $M_1$  составляется ровно из одного значения признака  $X_i$ , тогда как  $M_2$  составляется из всех остальных значений. Тем самым, если  $M = \{x_{1i}, x_{2i}, \dots, x_{ti}\}$ ,  $t \in \{1, 2, \dots, n\}$ , то

$$M_1 = \{C\}, \quad M_2 = M \setminus M_1,$$

где на каждой итерации  $C$  принимает по очереди значения каждого элемента множества  $M$ . В итоге выбирают то разделение, при котором прирост информации наибольший.

#### 4.4.1 Алгоритм построения бинарного дерева решений и пример с кошками

Составим алгоритм построения бинарного дерева для тренировочного набора данных  $X = \{x_1, x_2, \dots, x_n\}$  объема  $n$  с  $p$  предикторами  $X_1, X_2, \dots, X_p$  и откликом  $Y$ .

- Пусть на вход подается множество объектов  $X$ . Среди  $p$  предикторов выбрать тот, для фиксированного разбиения значений которого на подмножества достигается наибольший прирост информации. Итак, решается задача

$$\arg \max_{Q \in \{X_1, X_2, \dots, X_p\}} \text{IG}(Y|Q(C)),$$

где  $C$  определяется в зависимости от типа признака, согласно описанным выше правилам.

2. Пусть выбран предиктор  $X_i$  и найдено  $C$ . Выполнить разделение по выбранному признаку на два подмножества  $S_C$  и  $\bar{S}_C$ . В первое подмножество входят объекты, значения  $i$ -ого атрибута которых удовлетворяют критерию разделения, во второе – все остальные.
3. Повторить шаги 1 и 2 для множеств  $S_C$  и  $\bar{S}_C$  либо согласно ограничением числа уровней (согласно замечанию 4.2.1), либо до нулевой энтропии в каждой из групп, если ограничение не задано.

Опробуем алгоритм все на тех же кошках. Так, согласно выполненным ранее вычислениям, предиктор «Шерсть» со значением «Серый» обеспечивал нулевую энтропию (то есть максимальный прирост информации). Этот критерий и будет первым при формировании двух подмножеств  $S_{\text{Серый}}$  и  $\bar{S}_{\text{Серый}}$ . В первое множество войдут все кошки с серой шерстью, во второе – все остальные. Энтропия во втором множестве составит

$$-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \approx 0.971.$$

В первом подмножестве неопределенности нет, а вот во втором неопределенность сохраняется. Согласно третьему шагу алгоритма это означает, для этого подмножества необходимо повторить первые два шага алгоритма. Дальнейшее разбиение подмножества и критерии можно увидеть на рисунке 4. Оказывается, на втором уровне самым информативным снова является разделение кошек по цвету, но теперь уже на «Белых» и «Не белых». После этого, на последнем уровне, кошки делятся в зависимости от размера. Легко видеть, что на последнем уровне дерева все подмножества имеют нулевую энтропию, что означает, что построение завершено.

Бинарное дерево, как мы уже отмечали, позволяет проводить классификацию новых объектов даже в случае, когда значение какого-то предиктора оказалось уникальным (или новым). Так, классификация невысокой черной кошки в случае дерева принятия решений, построенного ранее, была невозможна. В то же время, если рассмотреть только что построенное бинарное дерево, классификация возможна. Так как черная кошка – не серая, то идем по левой ветке. Опять, так как черная кошка – не белая, снова идем по левой ветке. Ну и, наконец, так как рассматриваемая кошка – не высокая, то идем по правой ветке и делаем вывод, что кошка не понравится судьям.

#### 4.4.2 Синтетический пример

Опробуем алгоритм на синтетических данных с двумя предикторами  $X_1$  и  $X_2$ , причем на таких, которые интуитивно разбиваются прямой на два подмножества. Предикторы, отвечающие рассматриваемым данным, принимают лишь числовые значения.

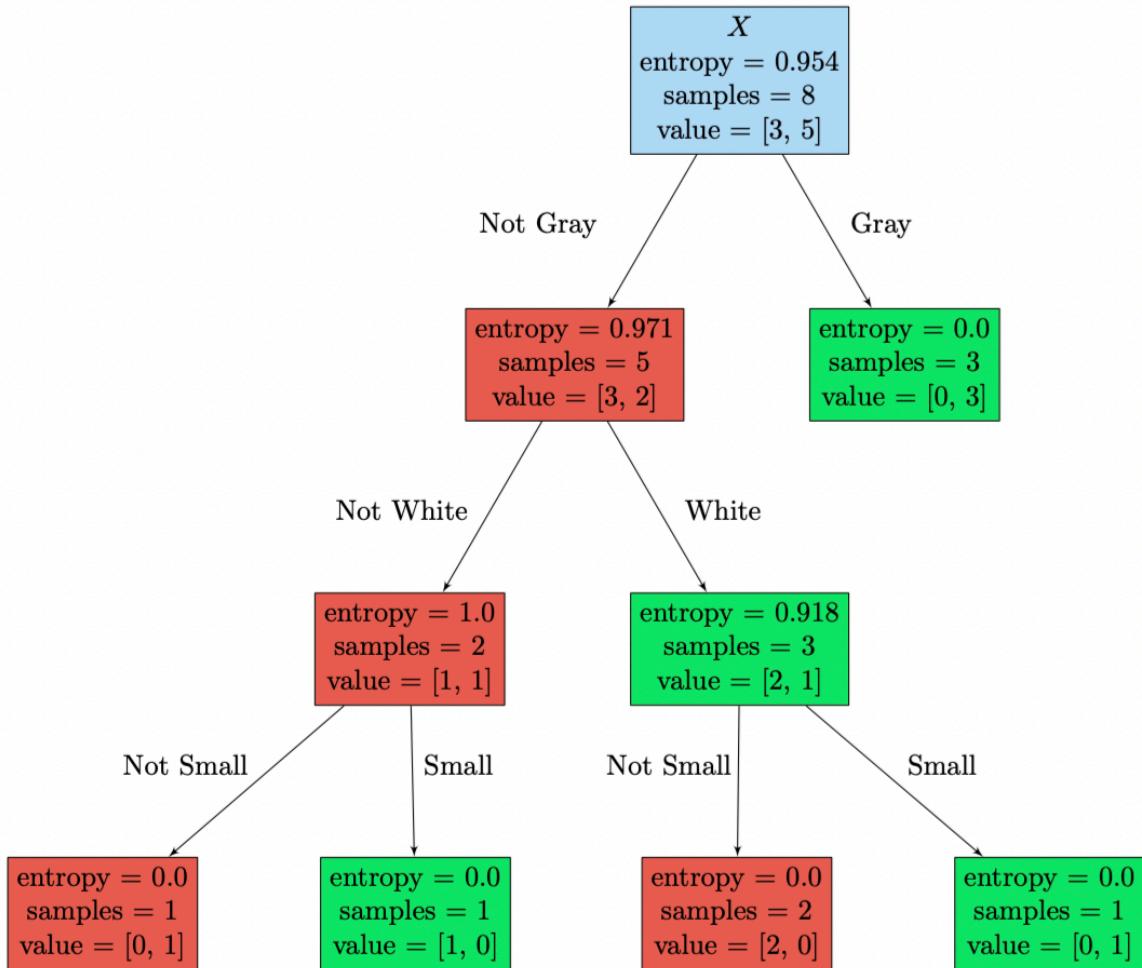


Рис. 4: Пример бинарного дерева решений.

Задача все та же – провести двухклассовую классификацию. У данных с откликом 0 (зеленые) предикторы независимы и имеют распределение  $N_{0,1}$ . Таких точек сгенерировано 100 штук. У данных с откликом 1 (желтые) предикторы тоже независимы, но имеют распределение  $N_{2,1}$ . Всего сгенерировано 200 значений, они изображены на рисунке 5.

Построим бинарное решающее дерево, ограничившись лишь двумя уровнями (деревом глубины 2). Начальная энтропия рассматриваемого набора данных, конечно, максимальна и равна единице, так как у нас поровну единиц и нулей в отклике. Расчет в системе моделирования привел к тому, что для начального набора разделение будет оптимальным по признаку  $X_2 \leq 0.77$ . Таким образом, выполняются действия, аналогичные рассмотренному примеру с породами кошек, однако заметно возрастает количество вычислений.

На рисунке 6 отражено сказанное нами ранее: исходный набор из 200 элементов делится на два по признаку  $X_2 \leq 0.77$ . При этом к зеленому блоку относятся те данные, которые удовлетворяют условию разделению, то есть данные, второй признак которых не превосходит 0.77, а к красному – все

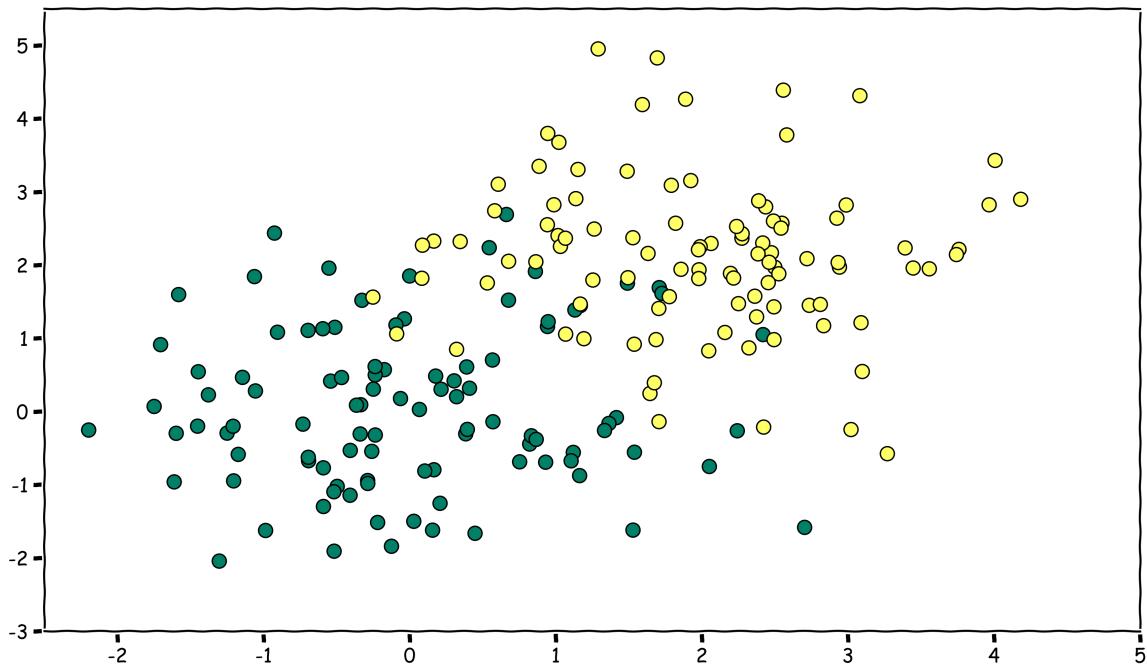


Рис. 5: Объекты интуитивно разделимы.

остальные. Так как энтропия в каждой группе отлична от нуля, то алгоритм продолжает свою работу. Видно, что в обеих группах следующее разделение идет по признаку  $X_1$ . Левая группа делится по условию  $X_1 \leq 1.59$ , а правая – по условию  $X_1 \leq 0.04$ .

В данном примере глубина ограничена 2 уровнями, но, как мы видим, даже у такого «ограниченного» дерева есть группа с нулевой энтропией (самая левая), что, конечно, хорошо. Согласно сформулированному алгоритму, все узлы на втором уровне дерева становятся листами с откликами «Зеленый», «Желтый», «Зеленый», «Желтый», соответственно. Например, почему второй лист отвечает отклику «Желтый»? Как видно, из всех тренировочных данных в этот блок попадает 3 зеленых и 7 желтых объектов. Отклик же устанавливается по отклику наибольшего числа входящих объектов. Так как объектов с откликом «Желтый» большинство, то таков отклик и у листа.

Классификацию (а также сформулированные условия разделения) при помощи ДПР в нашем примере можно визуализировать непосредственно на плоскости. Для этого на первом этапе всю плоскость достаточно разделить прямой  $X_2 = 0.77$ . Мы видим на рисунке 7, что плоскость разделилась на две части: нижняя, отвечающая левому блоку в дереве ( $X_2 \leq 0.77$ ) и верхняя, отвечающая правому блоку ( $X_2 > 0.77$ ).

Далее, условие следующего разделения для нижней части плоскости (левого блока в дереве) – это условие  $X_1 \leq 1.59$ , а для верхней (или правого блока) – условие  $X_1 \leq 0.04$ . Визуализацию разделения вы видите на рисунке 8.

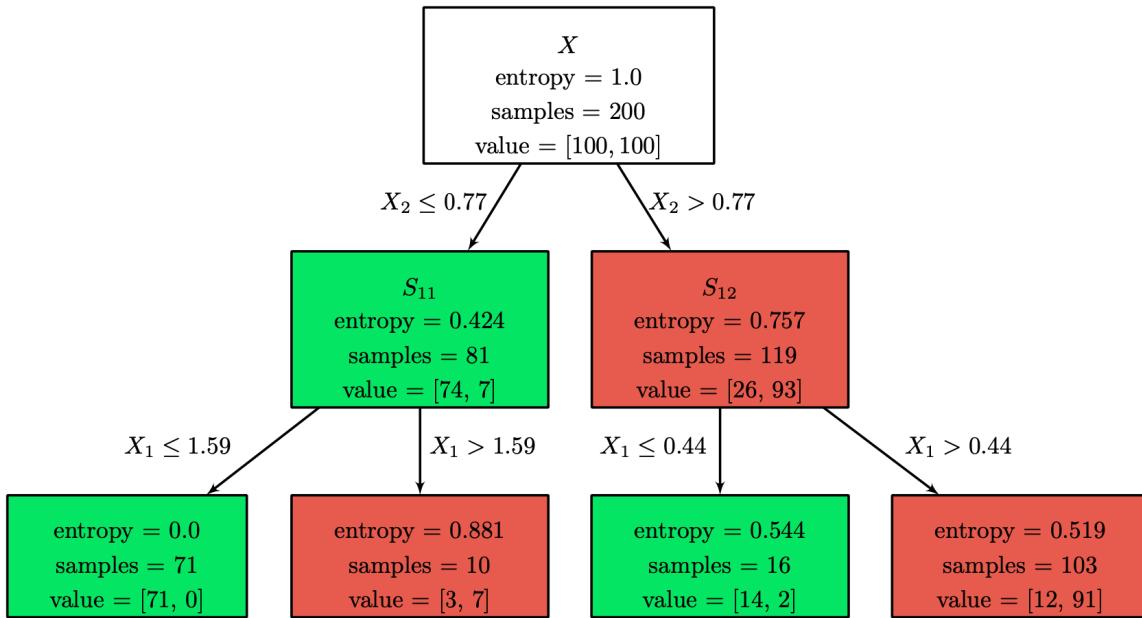


Рис. 6: Дерево решений.

На этом этапе дерево считается обученным. Мы видим, что оно допускает ошибки и на тренировочных данных – это ситуации, когда желтые точки оказываются в зеленой области, или, наоборот, когда зеленые точки попадают в желтую область; всего таких ошибок – 17. Посмотрим, как наше дерево будет работать на новой выборке из нормального распределения с такими же параметрами. Как видно, в целом, число ошибок сохранилось: на этапе обучения их было 17, а на этапе тестирования – 21. На интуитивном уровне, дерево ошибается примерно в 11% случаев, что не так уж и плохо, а значит дерева с двумя уровнями вполне достаточно. Но что произойдет, если увеличить количество уровней, например, до шести? Конечно, дерево станет заметно больше, но и число групп с нулевой энтропией возрастет. На рисунке изображены области классификации, отвечающие условиям деления на группы при построении дерева на все тех же исходных данных, как и при построении дерева с двумя уровнями.

Похоже, исходя из наглядного представления дерева, что мы получили переобучение: и число областей, и их причудливая форма кажутся совершенно неестественными. И это действительно так. Протестировав наше дерево на тестовом наборе, мы видим, что имеются как пустые области, которые не используются вовсе (в них не попадает ни одно из тестовых данных), так и области, в которые теперь попадает много точек из другого класса – ошибки классификации.

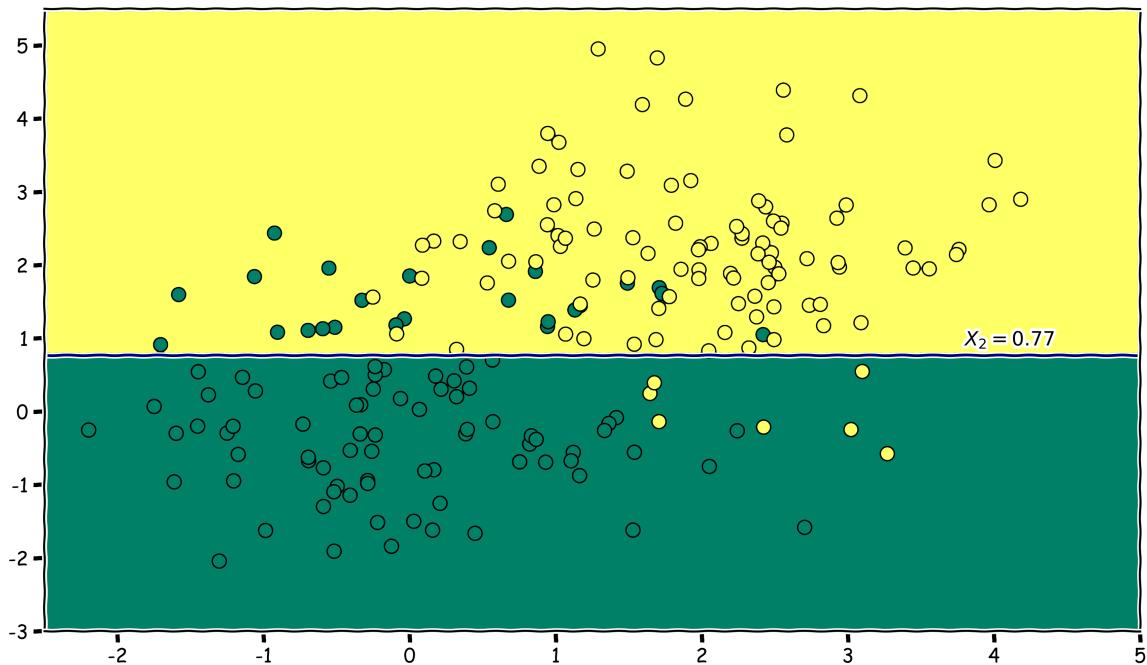


Рис. 7: Классификация на основе дерева решений (первое разделение).

## 5 Неопределенность Джини

### 5.1 Определение и свойства

При построении деревьев принятия решений используют и другие критерии информативности. Среди них можно выделить особо так называемую неопределенность Джини (Gini impurity), которая является мерой неправильной классификации.

Что значит неправильная классификация? Для демонстрации происходящего, снова вернемся к эксперименту с опросом друзей. В нем мы получили, что из 14 опрошенных друзей 9 согласились пойти играть в футбол, а 5 нет. Тогда эксперимент задается следующей таблицей

$\Omega$	Да	Нет
$P$	$\frac{9}{14}$	$\frac{5}{14}$

Предположим, что мы берем случайного друга. Какова вероятность события, что мы его неверно классифицируем? Это событие распадается на два: он либо хочет играть в футбол, а мы его отнесем к группе, которая не хочет, либо наоборот. Ясно, что вероятность такого события может быть вычислена, как

$$P(\text{друг неверно классифицирован}) = \frac{9}{14} \cdot \frac{5}{14} + \frac{5}{14} \cdot \frac{9}{14} = \frac{90}{196}.$$

Эта величина и есть неопределенность Джини или мера неправильной классификации. Итак, введем формальное определение.

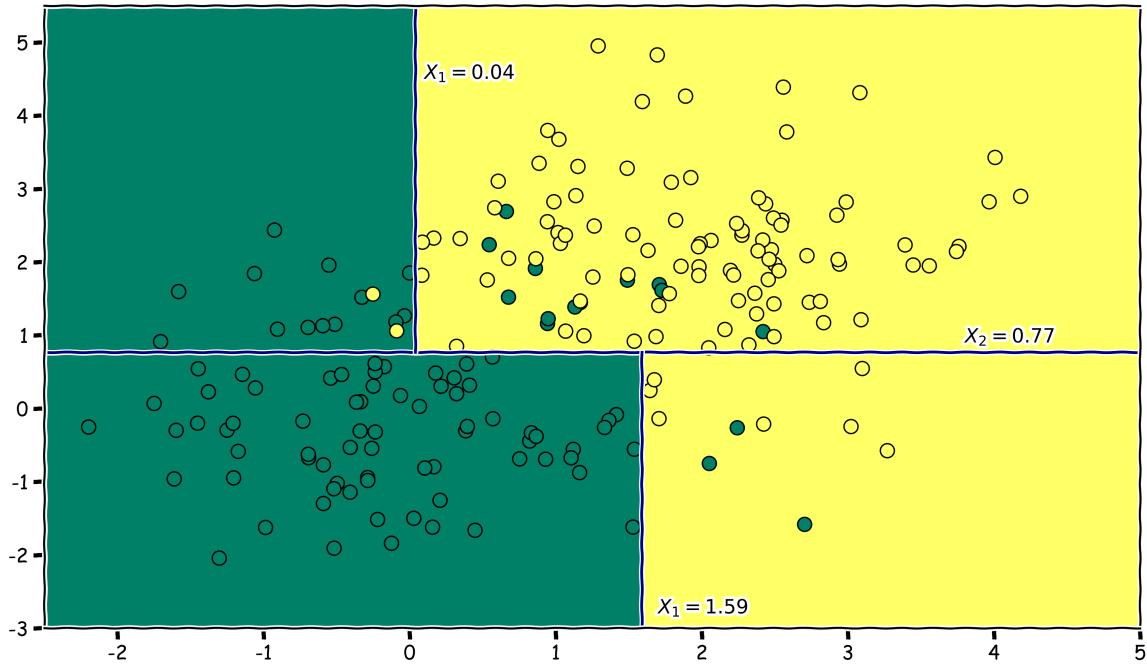


Рис. 8: Классификация на основе дерева решений (итоговое разделение).

**Определение 5.1.1** Пусть эксперимент  $\Omega$  описывается таблицей

$$\begin{array}{c|c|c|c|c} \Omega & \omega_1 & \omega_2 & \dots & \omega_n \\ \hline P & P_1 & P_2 & \dots & P_n \end{array}.$$

Неопределенностью Джини  $G(\Omega)$  эксперимента  $\Omega$  называется величина

$$G(\Omega) = \sum_{i=1}^n P_i(1 - P_i).$$

**Замечание 5.1.1** Полезно заметить, что выражения для неопределенности Джини эквивалентным образом может быть переписано и вот так:

$$G(\Omega) = 1 - \sum_{i=1}^n P_i^2.$$

Это представление легко получается из следующих выкладок:

$$G(\Omega) = \sum_{i=1}^n P_i(1 - P_i) = \sum_{i=1}^n (P_i - P_i^2) = \sum_{i=1}^n P_i - \sum_{i=1}^n P_i^2 = 1 - \sum_{i=1}^n P_i^2.$$

Как же связаны неопределенность Джини и энтропия? Оказывается, неопределенность Джини обладает схожими свойствами, что и энтропия.

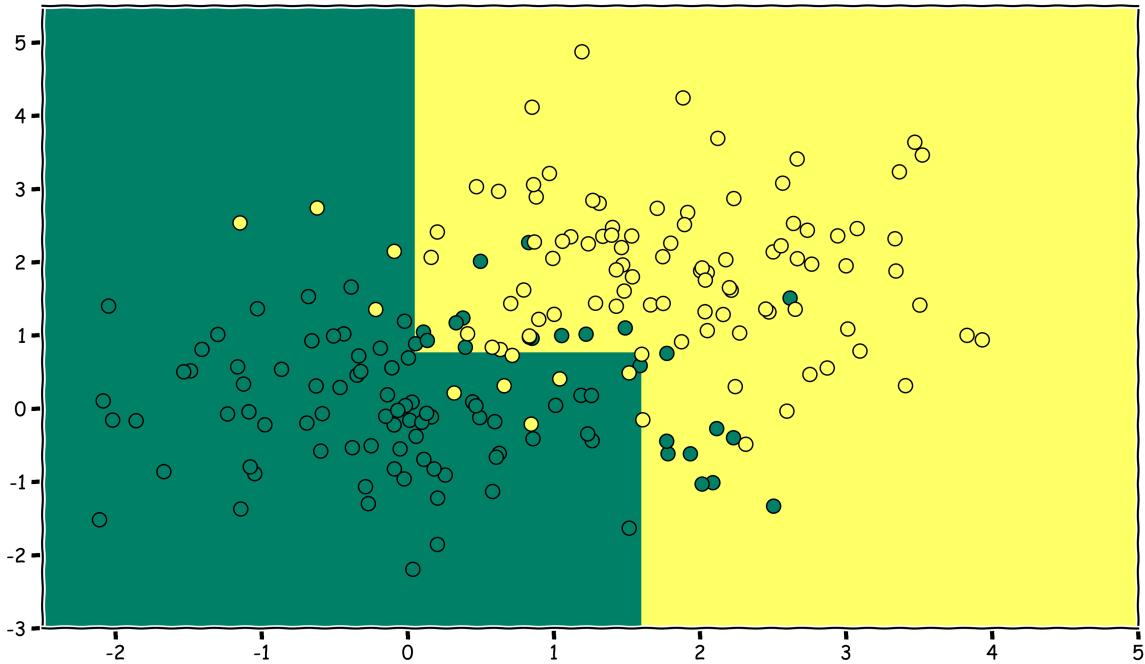


Рис. 9: Классификация новых объектов на основе обученной модели.

**Теорема 5.1.1** Пусть рассматривается эксперимент  $\Omega$ . Тогда

$$G(\Omega) = 0 \Leftrightarrow H(\Omega) = 0.$$

Так как энтропия равна нулю тогда и только тогда, когда у эксперимента ровно один исход имеет вероятность 1 (то есть нет неопределенности), то тоже самое можно сказать и про неопределенность Джини.

**Доказательство.** Если  $H(\Omega) = 0$ , то существует исход эксперимента  $\omega_i$  такой, что его вероятность  $P_i = 1$ . Тогда вероятности всех остальных исходов (если они есть) равны 0, а значит

$$G(\Omega) = 1 - P_i^2 = 1 - 1 = 0.$$

Обратно, если  $G(\Omega) = 0$ , то

$$\sum_{i=1}^n P_i(1 - P_i) = 0.$$

Так как все слагаемые неотрицательны, то каждое из них равно 0, а значит  $P_i(1 - P_i) = 0$  при всех  $i \in \{1, 2, \dots, n\}$ . В итоге, для каждого  $P_i$  выполнено одно из двух: оно либо равно 0, либо 1. Но так как сумма всех  $P_i$  равна 1, то лишь одно  $P_i = 1$ , а остальные равны нулю. Значит, и энтропия такого эксперимента равна нулю.  $\square$

Аналогично энтропии, неопределенность Джини максимальна в случае, когда все исходы эксперимента  $\Omega$  равновозможны, а именно справедливо следующее наблюдение.

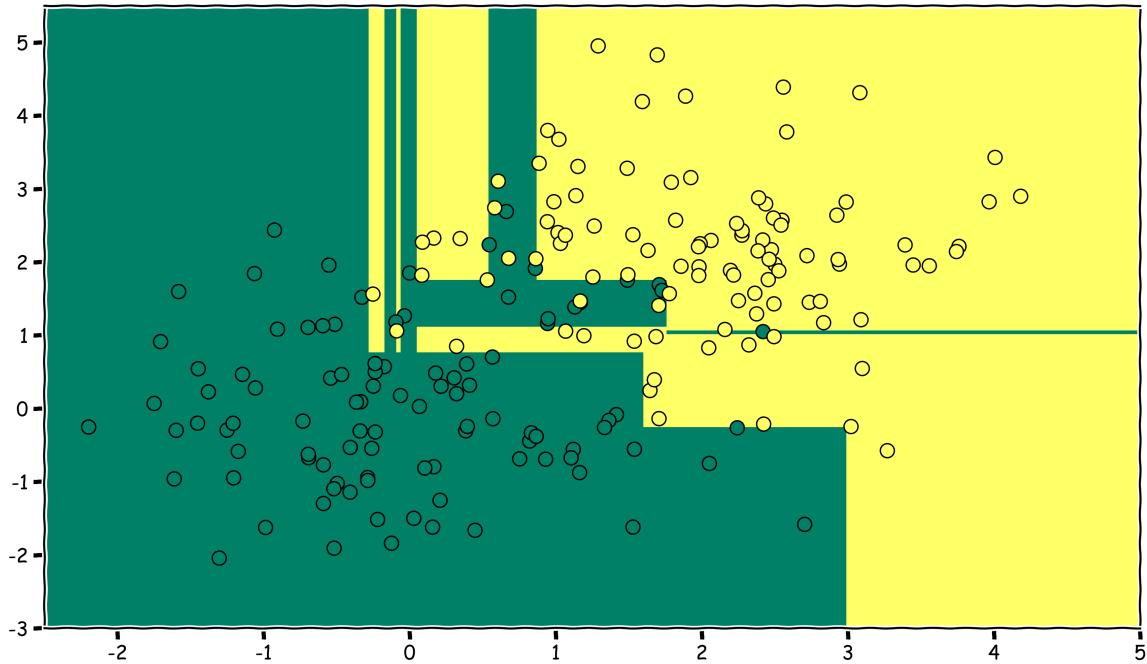


Рис. 10: Классификация на основе дерева решений (6 уровней).

**Теорема 5.1.2** Неопределенность Джини  $G$  максимальна в случае, когда все исходы эксперимента равновозможны, то есть, когда эксперимент описывается таблицей вида

$\Omega$	$\omega_1$	$\omega_2$	$\dots$	$\omega_n$
$P$	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

В этом случае неопределенность Джини равна

$$G(\Omega) = 1 - \frac{1}{n},$$

**Доказательство.** Ясно, что на эксперименте, заданным таблицей

$\Omega$	$\omega_1$	$\omega_2$	$\dots$	$\omega_n$
$P$	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

получаем, что

$$G(\Omega) = 1 - \sum_{i=1}^n \frac{1}{n^2} = 1 - \frac{1}{n}.$$

Докажем, что это значение максимально. Так как функция  $\frac{1}{x}$  выпукла вниз, то для нее справедливо неравенство Йенсена: для любых чисел  $p_1, p_2, \dots, p_n > 0$  таких, что  $p_1 + p_2 + \dots + p_n = 1$  и любых  $x_1, x_2, \dots, x_n$  из интервала выпуклости функции справедливо:

$$f(p_1 x_1 + p_2 x_2 + \dots + p_n x_n) \leq p_1 f(x_1) + p_2 f(x_2) + \dots + p_n f(x_n)$$

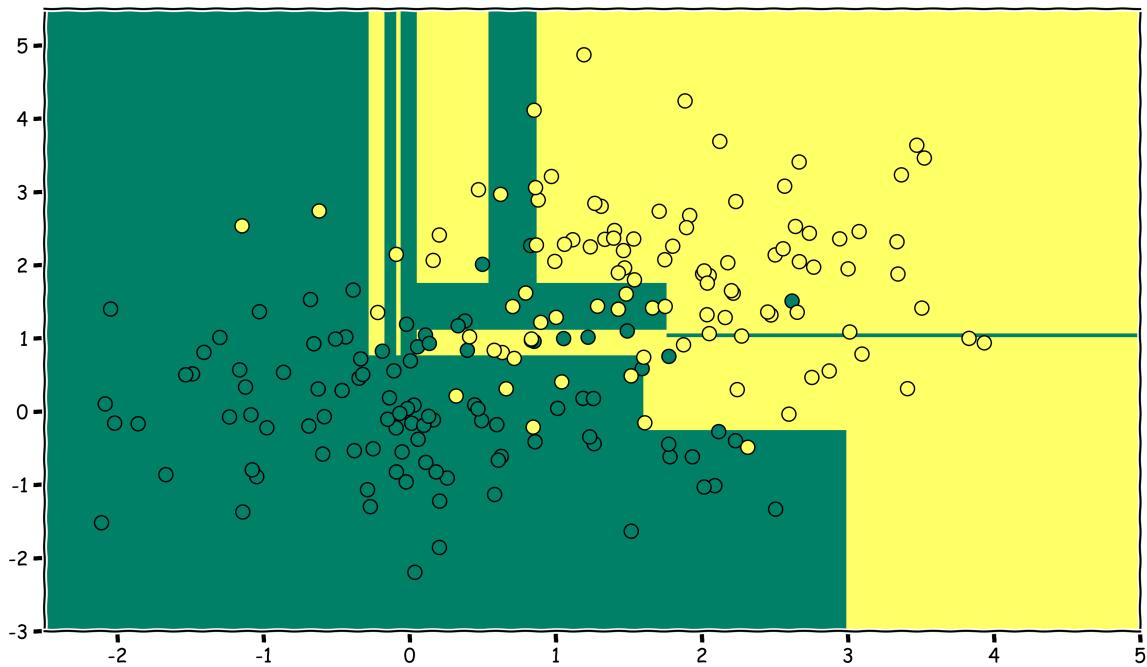


Рис. 11: Классификация новых объектов на основе обученной модели.

или

$$f \left( \sum_{i=1}^n p_i x_i \right) \leq \sum_{i=1}^n p_i f(x_i).$$

Положим  $p_i = P_i$  и  $x_i = \frac{1}{P_i}$ , получим

$$f \left( P_1 \cdot \frac{1}{P_1} + \dots + P_n \cdot \frac{1}{P_n} \right) = f(n) = \frac{1}{n} \leq \sum_{i=1}^n P_i \frac{1}{P_i} = \sum_{i=1}^n P_i^2,$$

откуда

$$1 - \sum_{i=1}^n P_i^2 \leq 1 - \frac{1}{n}.$$

□

Легко видеть, что, как и энтропия, неопределенность Джини растет с ростом количества равновозможных исходов эксперимента  $\Omega$ .**Следствие 5.1.3** Из последней теоремы легко получить важное следствие: неопределенность Джини  $G(\Omega)$  любого эксперимента  $\Omega$  всегда находится в следующих пределах:

$$0 \leq G(\Omega) < 1.$$

Как и ранее, резонно задаться вопросом: а как определяется неопределенность Джини в случае эксперимента  $(\Omega, \Theta)$ ? Напомним, что экспериментом  $(\Omega, \Theta)$  называется произвольное множество пар исходов  $(\omega_i, \theta_j)$ ,

$i \in \{1, 2, \dots, m\}$ ,  $j \in \{1, 2, \dots, n\}$ , каждой из которых сопоставлено число  $P_{ij} \geq 0$ , называемое вероятностью исхода  $(\omega_i, \theta_j)$  такое, что

$$\sum_{i=1}^m \sum_{j=1}^n P_{ij} = 1.$$

При этом эксперимент  $(\Omega, \Theta)$  может быть описан (и отождествлен) следующей таблицей:

$(\Omega, \Theta)$	$\theta_1$	$\theta_2$	$\dots$	$\theta_n$
$\omega_1$	$P_{11}$	$P_{12}$	$\dots$	$P_{1n}$
$\omega_2$	$P_{21}$	$P_{22}$	$\dots$	$P_{2n}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\omega_m$	$P_{m1}$	$P_{m2}$	$\dots$	$P_{mn}$

А тогда, для эксперимента  $\Omega$  при условии, что эксперимент  $\Theta$  оказался в состоянии  $\theta_j$ , где  $j \in \{1, 2, \dots, n\}$  могут быть составлены следующие  $n$  таблиц условных вероятностей:

$$\frac{(\Omega, \theta_j)}{P} \mid \frac{(\omega_1 | \theta_j)}{P(\omega_1 | \theta_j)} \mid \frac{(\omega_2 | \theta_j)}{P(\omega_2 | \theta_j)} \mid \dots \mid \frac{(\omega_m | \theta_j)}{P(\omega_m | \theta_j)}.$$

Понятно, что каждая из таблиц описывает эксперимент, для которого можно вычислить неопределенность Джини. Как и в случае с энтропией, введем следующее определение.

**Определение 5.1.2** Условной неопределенностью Джини эксперимента  $\Omega$  при условии, что эксперимент  $\Theta$  оказался в состоянии  $\theta_j$ ,  $j \in \{1, 2, \dots, n\}$  называется число

$$G(\Omega | \theta_j) = 1 - \sum_{i=1}^m P^2(\omega_i | \theta_j).$$

В результате, для каждого состояния  $\theta_j$ , которое происходит с некоторой вероятностью  $P(\theta_j)$ , неопределенность Джини принимает значения  $G(\Omega | \theta_j)$ , а значит является случайной величиной с рядом распределения:

$$\frac{G(\Omega | \theta_j)}{P} \mid \frac{G(\Omega | \theta_1)}{P(\theta_1)} \mid \dots \mid \frac{G(\Omega | \theta_n)}{P(\theta_n)}.$$

Так, полученную систему хорошо характеризует так называемая взвешенная неопределенность Джини.

**Определение 5.1.3** Взвешенной неопределенностью Джини эксперимента  $\Omega$  при условии, что произошел эксперимент  $\Theta$ , называется величина

$$G(\Omega | \Theta) = E(G(\Omega | \theta_j)) = \sum_{j=1}^n P(\theta_j) G(\Omega | \theta_j).$$

Прирост информации, основывающийся на понятии неопределенности Джини, вводится следующим образом.

**Определение 5.1.4** Приростом Джини (*Gini Gain*) называется величина

$$\text{GG}(\Omega|\Theta) = \text{G}(\Omega) - \text{G}(\Omega|\Theta).$$

Значение прироста Джини используют при построении деревьев принятия решений аналогично тому, как мы использовали прирост информации: каждый раз на первом шаге алгоритма выбирается наибольший прирост Джини.

## 5.2 Небольшое сравнение прироста Джини и энтропии

На практике рассмотренные критерии прироста информативности дают почти одинаковый результат. В этом легко убедиться, построив, для примера, графики функций энтропии и удвоенной неопределенности Джини в случае двух исходов:

$$I_{\text{Entropy}} = -p \log_2 p - (1-p) \log_2(1-p),$$
$$I_{\text{Gini}} = 2 \cdot (p(1-p) + (1-p)(1-(1-p))).$$

Удвоение неопределенности Джини позволяет нормировать значения функции  $I_{\text{Gini}}$  до единицы, что, условно, «уравнивает единицы измерения». Как видно рисунка 12, неопределенность системы с двумя исходами для различных значений вероятности сопоставимы.

Таким образом, нет явных предпосылок использовать конкретный метод измерения информативности, так как они оба справляются со своей задачей и дают почти одинаковый результат. Однако вычисление энтропии – задача более затратная (с точки зрения вычислений) из-за присутствующего в определении логарифма. Именно по этой причине часто, по умолчанию, в инструментах используется неопределенность Джини.

## 5.3 Прирост Джини на данных

Вернемся к примеру с кошками. Перед вами таблица, содержащая информацию о породе, цвете шерсти и росте в холке, которые влияют на привлекательность конкретной кошки, участвующей в соревнованиях.

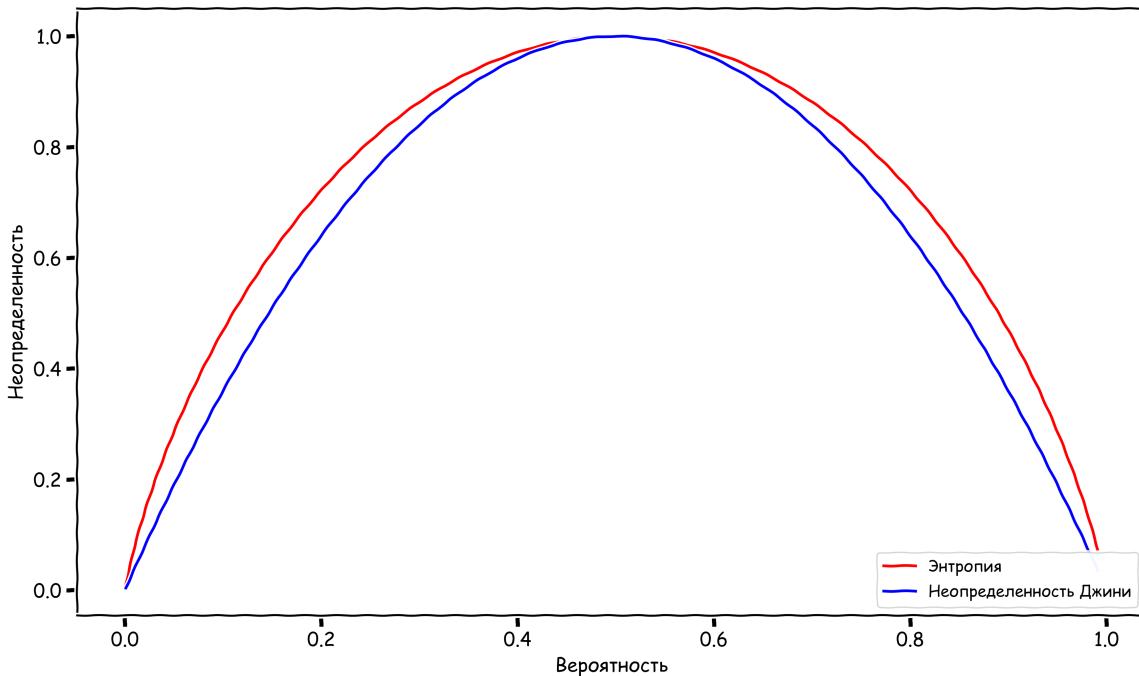


Рис. 12: Сравнение энтропии и неопределенности Джини.

Порода	Шерсть	Рост	Привлекательность
Британец	Белый	Высокий	Нет
Британец	Серый	Высокий	Да
Британец	Белый	Низкий	Да
Мейн-кун	Белый	Высокий	Нет
Мейн-кун	Коричневый	Высокий	Да
Мейн-кун	Коричневый	Низкий	Нет
Рэгдолл	Серый	Высокий	Да
Мейн-кун	Серый	Низкий	Да

Условные вероятности отнесения к целевому классу «Привлекательность» в зависимости от породы были найдены ранее, они равны:

$$\mathsf{P}(\text{Привлекательность} = \text{Да} | \text{Порода} = \text{Британец}) = \frac{2}{3},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Нет} | \text{Порода} = \text{Британец}) = \frac{1}{3},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Да} | \text{Порода} = \text{Мейн-кун}) = \frac{1}{2},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Нет} | \text{Порода} = \text{Мейн-кун}) = \frac{1}{2},$$

$$\mathsf{P}(\text{Привлекательность} = \text{Да} | \text{Порода} = \text{Рэгдолл}) = 1$$

$$\mathsf{P}(\text{Привлекательность} = \text{Нет} | \text{Порода} = \text{Рэгдолл}) = 0.$$

Тогда неопределенность Джини для каждой породы будет принимать следующие значения:

$$G(\text{Привлекательность} | \text{Порода} = \text{Британец}) = 1 - \left( \left( \frac{2}{3} \right)^2 + \left( \frac{1}{3} \right)^2 \right) = \frac{4}{9},$$

$$G(\text{Привлекательность} | \text{Порода} = \text{Мейн-кун}) = 1 - \left( \left( \frac{1}{2} \right)^2 + \left( \frac{1}{2} \right)^2 \right) = \frac{1}{2},$$

$$G(\text{Привлекательность} | \text{Порода} = \text{Рэгдолл}) = 1 - (1^2 + 0^2) = 0.$$

При этом состояния, то есть выбор конкретной породы, определяются с некоторой вероятностью, а именно:  $\frac{3}{8}$ ,  $\frac{4}{8}$  и  $\frac{1}{8}$  соответственно. Запишем полученные значения в таблицу:

$G(\text{Привлекательность}   \text{Порода})$	$\frac{4}{9}$	$\frac{1}{2}$	0
P	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{1}{8}$

Окончательно, взвешенная неопределенность Джини составит:

$$G(\text{Привлекательность} | \text{Порода}) = \frac{3}{8} \cdot \frac{4}{9} + \frac{4}{8} \cdot \frac{1}{2} + \frac{1}{8} \cdot 0 = \frac{5}{12}.$$

При этом неопределенность Джини, эксперимента «Привлекательность» равна:

$$G(\text{Привлекательность}) = 1 - \left( \left( \frac{5}{8} \right)^2 + \left( \frac{3}{8} \right)^2 \right) = \frac{15}{32},$$

а значит прирост Джини:

$$GG(\text{Привлекательность} | \text{Порода}) = \frac{15}{32} - \frac{5}{12} = \frac{5}{96} \approx 0.052.$$

Аналогичным образом рассчитываются прирост Джини для всех атрибутов. Конечно, атрибут – шерсть, в такой эвристике тоже лучше всего характеризует прирост информации, как и в случае с энтропией.

## 6 Деревья принятия решений на реальном примере

Одним из примеров применения деревьев принятия решений являются системы скоринга, используемые банками для оценки клиентов. Например, некий банк имеет клиентскую базу, и хочет понять, какие клиенты согласятся на оформление потребительского кредита в будущем, а какие нет.

Рассмотрим данные (рисунок 13) о текущих клиентах банка<sup>1</sup> – всего 12 предикторов, среди которых: возраст, опыт работы, доход, семейное положение, образование и проч., и отклик – *PersonalLoan*, принимающий значений 0, если клиент отказался оформить потребительский кредит, и 1 – если согласился. Исходный объем данных содержит 5000 объектов. Для дальнейшей оценки качества модели, мы его разделили на тренировочные и тестовые данные в отношении 70% на 30%. Таким образом, обучение модели осуществляется на 3500 объектах (3158 из которых относятся к классу 0 – клиенты отказавшиеся оформлять потребительский кредит, остальные к классу 1 – согласились). Отметим, что все предикторы принимают числовые значения,

	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
ID													
1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1

Рис. 13: Пример тренировочных данных.

то есть являются некатегориальными. Будем строить бинарное дерево решений. В качестве критерия разделения будем использовать критерий  $X_i \leq C$ , где  $C$  – пороговый параметр из диапазона значений, принимаемых предиктором  $X_i$ . Так, на первом уровне, самым информативным является разделение по доходу (предиктор *Income*) с пороговым значением 92.5 тысячи долларов в год. Такое разделение приводит к достаточно информативному множеству  $S_{11}$  с близкой к нулю энтропией, второе же множество –  $S_{12}$  обладает высокой степенью неопределенности.

Теперь построим второй уровень дерева. Для  $S_{11}$  самым информативным будет разделение по предиктору *CCAvg* – расходы по кредитным картам в месяц. Для  $S_{12}$  – *Education* – уровень образования (он принимает значения 1 – для студентов, 2 – для выпускников и 3 – для имеющих профессиональное образование (степень)). В результате, уже на втором уровне мы получаем подмножество из 2393 клиентов, которые отнесены к классу 0 (левый блок с нулевой энтропией) – это клиенты, которые не соглашались на оформление потребительского кредита, имели доход не более 92.5 тысяч долларов в год и незначительные расходы по кредитной карте – не более 2.95 тысячи долларов в месяц.

Третий уровень дерева построен для оставшихся множеств с ненулевой энтропией – множеств  $S_{22}, S_{23}, S_{24}$ . На этом уровне тоже получаем множество с нулевой энтропией (правый нижний блок на рисунке 14). На этот раз блок

<sup>1</sup><https://www.kaggle.com/itsmesunil/bank-loan-modelling>

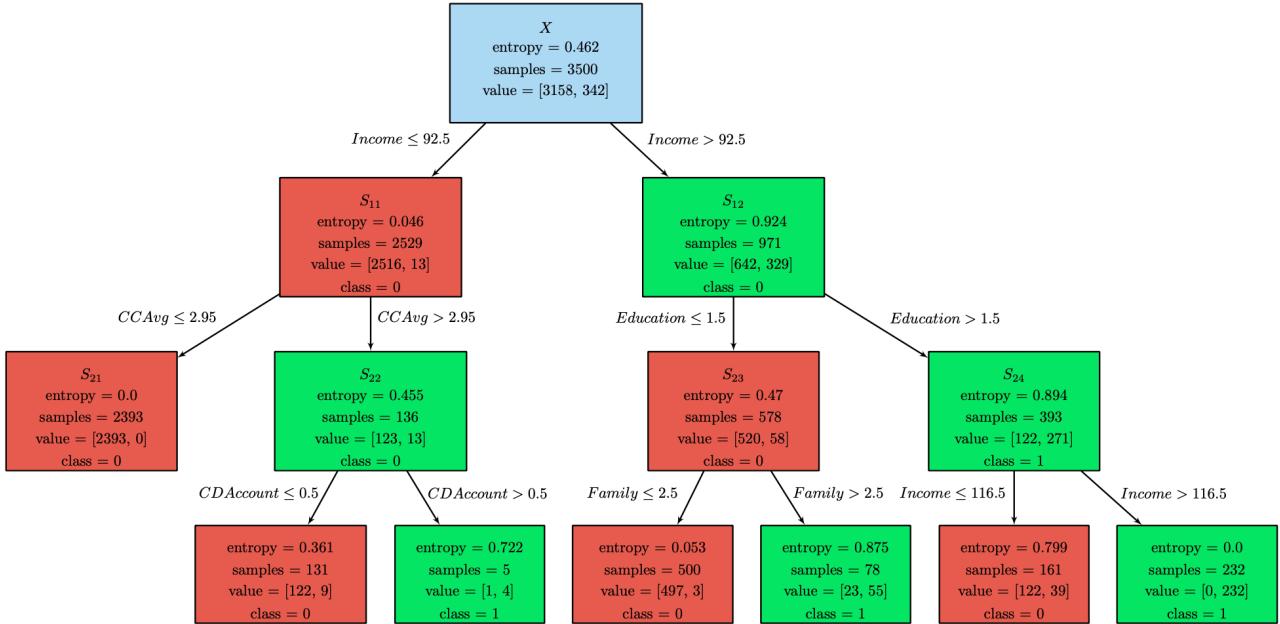


Рис. 14: Дерево принятия решений (энтропия).

отвечает клиентам, согласившимся оформить потребительский кредит. Что это за клиенты, согласно построенному дереву? Это клиенты с высоким доходом, более 116.5 тысяч долларов в год и либо выпускники, либо имеющие дополнительный уровень образования (степень).

В остальных подмножествах присутствуют разные клиенты, и мы можем либо продолжить построение дерева, либо закончить на трех уровнях. В таком случае класс назначается, как обычно, исходя из «большинства».

На основании классификации тестовых данных, по построенному дереву может быть составлена матрица ошибок: Полученные результаты свидетель-

Матрица ошибок		Исходный класс	
		+	-
Прогноз	+	TP=108	FP=7
	-	FN=30	TN=1355

ствуют о высокой точности модели

$$Precision = \frac{TP}{TP + FP} = \frac{108}{108 + 7} \approx 0.9391,$$

а также полноте

$$Recall = \frac{TP}{TP + FN} = \frac{108}{108 + 30} \approx 0.7826.$$

Обучение на аналогичных данных, но с использованием неопределенности Джини, приводит к результату, отображеному на рисунке 15. Можно

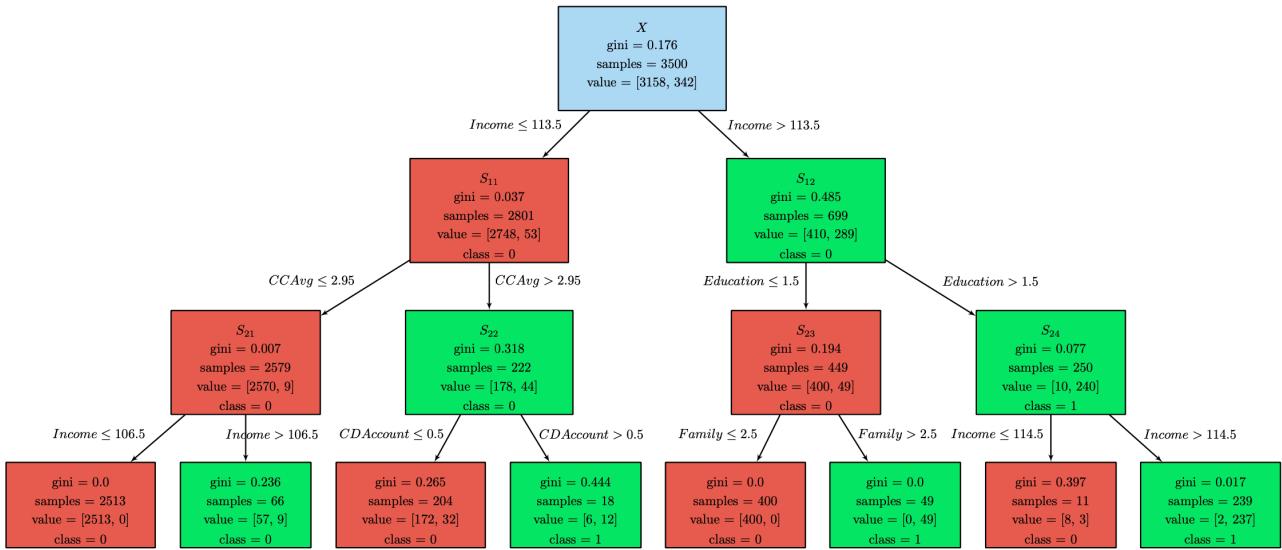


Рис. 15: Дерево принятия решений (Джини).

заметить, что многие критерии схожи с ранее построенным деревом. Кроме того, классификация тестовых данных дает тоже схожий результат: Точность

Матрица ошибок		Исходный класс	
		+	-
Прогноз	+	TP=113	FP=4
	-	FN=25	TN=1358

и полнота немного, но повысились:

$$Precision = \frac{TP}{TP + FP} = \frac{113}{113 + 4} \approx 0.9658,$$

$$Recall = \frac{TP}{TP + FN} = \frac{113}{113 + 25} \approx 0.8188.$$

## 7 Заключение

Итак, в этой лекции мы осветили еще один подход к классификации – деревья принятия решений, основная сфера применения которых – поддержка процессов принятия управленческих решений. Метод не наделен такими тонкостями, как подбор гиперпараметров модели, и позволяет быстро обучить модель. Выбор той или иной меры неопределенности существенно не сказывается на результатах классификации, а ограничения по глубине дерева или количеству элементов в каждом узле дерева являются эвристическими подходами, т.е. не гарантируют лучшего результата или вообще работают

только в каких-то частных случаях. Кроме того, обоснованных рекомендаций по тому, какой метод лучше работает, в настоящее время не существует, и окончательное решение остается на откуп исследователю.