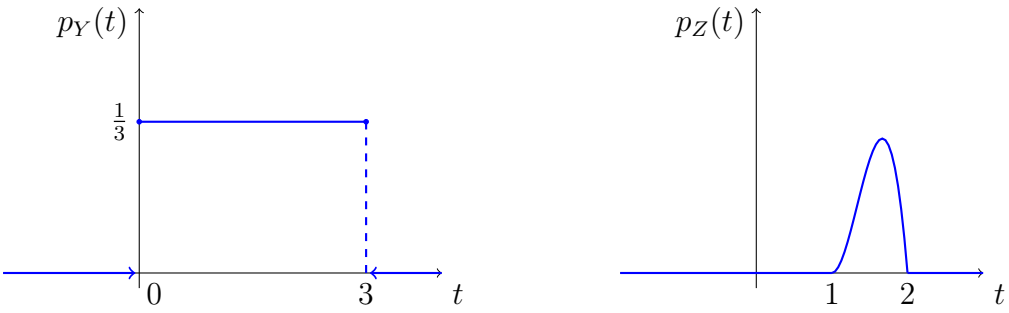
Машинное обучение, ФКН ВШЭ
Семинар №7

На этом семинаре мы поговорим про обучение деревьев. Чаще всего при их обучении для задачи классификации используют энтропию и критерий Джини. Сначала мы обсудим смысл этих критериев разбиения, затем научимся использовать их для строительства деревьев и заодно поговорим про интересную интерпретацию дивергенции Кульбака-Лейблера.

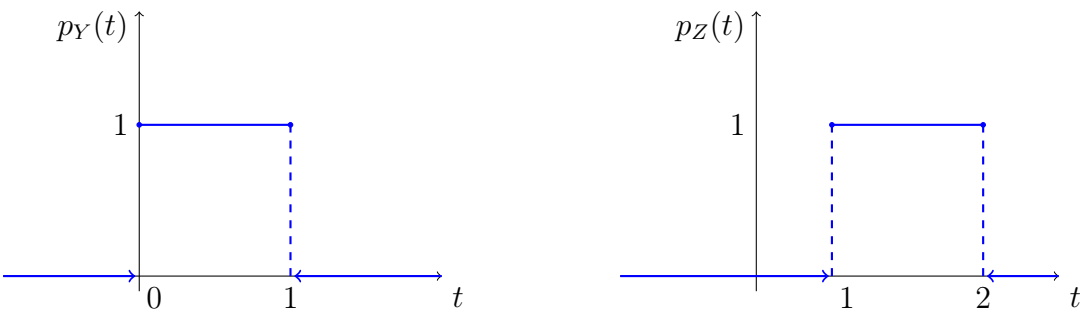
1 Энтропия

Давайте посмотрим на распределение двух случайных величин. Какая из них более предсказуемая, левая или правая?



Случайная величина Y (левая) сконцентрирована на широком отрезке. Она равновероятно может выскочить из любой его части. Плотность случайной величины Z (правая) имеет пикообразную форму. Её вероятностная масса сконцентрирована на узком отрезке, из-за этого её проще прогнозировать. Ошибка, которую мы будем допускать, окажется меньше из-за формы распределения.

Какая из следующих двух случайных величин более предсказуема, левая или правая?



Случайные величины Y и Z отличаются друг от друга только отрезком. Одна распределена на $[0;1]$, вторая на $[1;2]$. Их форма одинакова. Они принимают разные значения, но одинаково непредсказуемо. Их одинаково сложно прогнозировать.

Энтропия – это мера непредсказуемости случайной величины Y, это то количество информации, которое мы получаем, наблюдая её. Она никак не опирается на те значения, которые принимает случайная величина и для дискретного случая определяется как

$$H(X) = \mathbb{E}(-\log \mathbb{P}(X)).$$

Для непрерывного случая энтропия определяется как

$$H(X) = \mathbb{E}(-\log p_X(t)).$$

В качества основания логарифма берут либо 2 либо натуральный логарифм. Если A – событие, $\mathbb{P}(A)$ – вероятность события A, тогда величину $S(A) = -\log \mathbb{P}(A)$ можно интерпретировать как *«удивление» (surprise)* от того, что событие A произошло. Чем ниже вероятность события, тем сильнее мы удивляемся. Если $\mathbb{P}(A) = 0.5$, то $S(A) = 1$. Если $\mathbb{P}(A) = 0.5^{10}$, то $S(A) = 10$.

Энтропия довольно часто используется в машинном обучении. Например, с помощью неё обучают деревья. Кроме того, на ней базируется понятие спутанности (perplexity), которое определяется как

$$\text{Perplexity}(X) = e^{H(X)}$$

Со спутанностью мы встретимся, когда будем говорить про TSNE (метод с помощью которого высоко-размерные данные можно визуализировать на плоскости).

Задача 1.1. *Посчитайте энтропию для случайных величин:*

а)

Y	1	17	26
$\mathbb{P}(Y=k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

б) $Y \sim U[0;a]$

в) $Y \sim N(0,\sigma^2)$

Решение. Энтропия никак не смотрит на то, какие именно значения принимает случайная величина. Её интересует только то, как вероятность размазана по этим значениям:

$$H(Y) = -\frac{1}{3}\cdot\ln\left(\frac{1}{3}\right) - \frac{1}{3}\cdot\ln\left(\frac{1}{3}\right) - \frac{1}{3}\cdot\ln\left(\frac{1}{3}\right) = \ln 3.$$

Для случайной величины, принимающей 4 значения с вероятностями $\frac{1}{4}$ энтропия будет равна $\ln 4$, а в общем случае для $Y \sim U[0;a]$ энтропия составит

$$H(Y) = \int_0^a \frac{1}{a}\cdot\left(-\ln\left(\frac{1}{a}\right)\right)dt = \ln a.$$

Чем больше значений принимает равномерная случайная величина, тем она непредсказуемее. Для вырожденного распределения энтропия окажется нулевой. Для нормального распределения

$$H(Y) = \mathbb{E}(-\ln(p_Y(t))) = \int_{-\infty}^{+\infty} p_Y(t) \cdot \ln p_Y(t) dt =$$
$$= \mathbb{E}\left(\frac{1}{2}\ln(2\pi\sigma^2) + \frac{Y^2}{2\cdot\sigma^2}\right) = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}.$$

Если попробовать подставить в формулу разные значения σ, то можно получить следущую примерную табличку:

σ	1	10	100
$H(Y)$	$\ln 4.13$	$\ln 41.3$	$\ln 413$

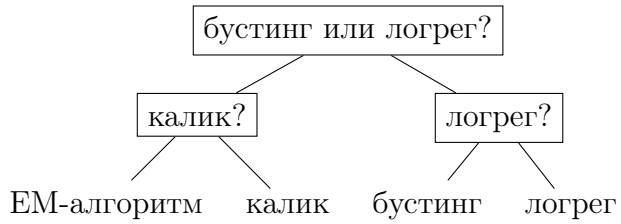
Выходит, что случайные величины $X \sim U[0;4]$ и $Y \sim N(0,1)$ в плане непредсказуемости очень похожи.

■

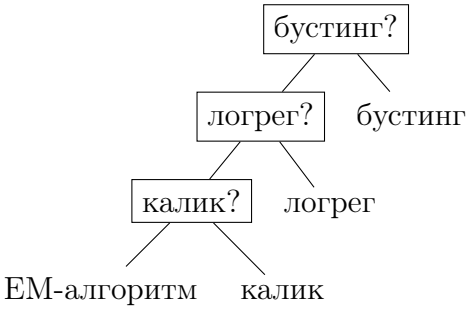
Давайте обсудим смысл энтропии и поймём, почему именно она служит хорошим критерием разбиения в деревьях. Пусть Женя загадал какое-то X. Мы знаем, что женя обычно загадывает что-то из четырёх вариантов с конкретными вероятностями

X	бустинг	логрег	EM-алгоритм	калик
$\mathbb{P}(X)$	$^1/_2$	$^1/_4$	$^1/_8$	$^1/_8$

Наша задача отгадать X за наименьшее количество-вопросов. Стратегию по отгадыванию можно строить по-разному. Например, следующая стратегия позволит нам гарантированно отгадать X за два вопроса. В дереве ниже при ответе да мы движемся в правый лист.



Если мы постоянно играем с Женей в эту игру, мы можем попробовать угадывать X за наименьшее число вопросов в среднем. Тогда первым вопросом надо задавать самый вероятный вариант, а дальше спрашивать по мере убывания вероятности. Оптимальная стратегия будет выглядеть следующим образом



Величина энтропии $H(X)$ будет тогда интерпретироваться как минимальное количество вопросов, которое мы в среднем зададим Жене. С вероятностью $\frac{1}{2}$ мы зададим только 1 вопрос, с вероятностью $\frac{1}{4}$ два и так далее

$$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{2}{8} \cdot 3 = -\left[\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{4} \cdot \log_2 \frac{1}{4} + \frac{2}{8} \cdot \log_2 \frac{1}{8}\right] = \frac{7}{4} < 2.$$

Когда дерево решает задачу бинарной классификации, ему на каждом объекте надо отгадывать класс. Кажется, что при разбиении, логично оптимизировать энтропию.

2 Критерий Джини

Пусть случайная величина X принимает K значений. Тогда *критерием (индексом) Джини* называют величину

$$J(X) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 = \sum_{k=1}^{K} p_k^2.$$

Задача 2.1. *Шаман Одэхингум*[1] *по прошлым наблюдениям знает, что большая охота на мамонта оказывается удачной с вероятностью* 0.3. *Если племя ждёт от него прогноз охоты, то Одэхингум, поплясав вокруг костра (10 минут) и постуча в бубен (42 раза) прогнозирует удачную охоту с вероятностью* 0.3 *и неудачную с вероятностью* 0.7.

Конкурирующий шаман Пэпина[2] *всегда прогнозирует неудачную охоту, как более вероятную. Когда шаман даёт неверный прогноз, его бьют палками.*

1. *Какова вероятность того, что Одэхингум ошибётся?*
2. *Кто чаще бывает бит палками?*
3. *Чему равен индекс Джини для случайной величины равной удаче с вероятностью* 0.3 *и неудаче с вероятностью* 0.7*?*

Решение. Ошибку Одэхингума мы можем найти по формуле полной вероятности как $0.3 \cdot 0.7 + 0.7 \cdot 0.3 = 0.42$ В первом слагаемом шаман сказал, что мамонта поймают, но его не поймали. Во втором наоборот. Выходит, что после охоты его будут бить палкой с вероятностью 0.42.

Шаман Пэпина всегда говорит, что охота будет неудачной. То есть её побьют палками с вероятностью 0.3.

Найдём значение индекса Джини для случайной величины

X	1	0
$\mathbb{P}(X = k)$	0.3	0.7.

[1] легкое колебание воды

[2] виноградная лоза, растущая вокруг дуба

$$J(X) = \sum_{k=1}^{K} p_k(1 - p_k) = 0.3 \cdot 0.7 + 0.7 \cdot 0.3 = 0.42.$$

Выходит, что индекс Джини измеряет, как часто шамана Одэхингума будут бить палками и отражает неопределённость, заложенную в случайную величину.

Мы будем использовать индекс Джини для обучения деревьев. Шаманами будут предикаты внутри вершин. Предикаты в них будут пытаться раздробить выборку на две части так, чтобы их пореже били палками. При этом, деревья будут бинарными. То есть мы не будем позволять вершинам скатываться в наивные прогнозы.

■

Задача 2.2. *Предположим, что мы оказались в конкретной вершине дерева m. В ней лежат объекты R. Поставим в соответствие вершине m алгоритм $a(x)$, который выбирает класс случайно, причем класс k выбирается с вероятностью p_k, где p_k – это доля объектов класса k в вершине. Покажите, что матожидание частоты ошибок этого алгоритма равно индексу Джини.*

Решение.

$$\mathbb{E}\left(\frac{1}{|R|}\sum_{(x_i,y_i)\in R}[y_i \neq a(x_i)]\right) = \frac{1}{|R|}\sum_{(x_i,y_i)\in R}\mathbb{E}[y_i \neq a(x_i)] = \frac{1}{|R|}\sum_{(x_i,y_i)\in R}(1 - p_{y_i}) =$$

$$= \sum_{k=1}^{K}\frac{\sum_{(x_i,y_i)\in R}[y_i = k]}{|R|}(1 - p_k) = \sum_{k=1}^{K} p_k(1 - p_k).$$

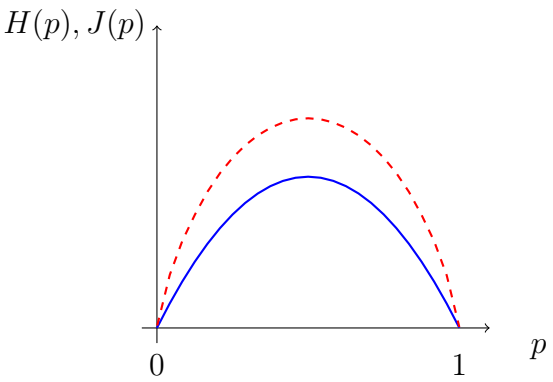
■

Задача 2.3. *Случайная величина X принимает значение 1 с вероятностью p и значение 0 с вероятностью $1 - p$. Постройте график зависимости индекса Джини и энтропии от p.*

Решение. Найдём требуемые величины

$$J(X) = p(1 - p) + (1 - p)p = 2(p - p^2)$$
$$H(X) = -(p \cdot \ln p + (1 - p) \cdot \ln(1 - p)).$$

На графике ниже пунктирной линией изображена энтропия. Непрерывная линия – критерий Джини.



Видно, что поведение этих функций очень похоже. Наибольшая неопределённость достигается в точке $p = 0.5$. На практике, обычно, нет особой разницы, от того, какой из двух критериев использовать для обучения дерева.

■

3 Связь критериев разбиения с функциями потерь

При построении решающего дерева необходимо задать *функционал качества*, на основе которого осуществляется разбиение выборки на каждом шаге. Этот функционал определяет, какой именно предикат лучше всего выбрать для данной внутренней вершины. Обозначим через R_m множество объектов, попавших в вершину, разбиваемую на данном шаге, а через R_ℓ и R_r — объекты, попадающие в левое и правое поддерево соответственно при заданном предикате. Мы будем использовать функционалы следующего вида:

$$Q(R_m, j, s) = H(R_m) - \frac{|R_\ell|}{|R_m|} H(R_\ell) - \frac{|R_r|}{|R_m|} H(R_r).$$

Здесь $H(R)$ — это *критерий информативности* (impurity criterion), который оценивает качество распределения целевой переменной среди объектов множества R. Чем меньше разнообразие целевой переменной, тем меньше должно быть значение критерия информативности — и, соответственно, мы будем пытаться минимизировать его значение. Функционал качества $Q(R_m, j, s)$ мы при этом будем максимизировать.

В каждом листе дерево будет выдавать константу — вещественное число, вероятность или класс. Исходя из этого, можно предложить оценивать качество множества объектов R тем, насколько хорошо их целевые переменные предсказываются константой (при оптимальном выборе этой константы):

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c),$$

где $L(y, c)$ — некоторая функция потерь.

§3.1 Регрессия

Как обычно, в регрессии выберем квадрат отклонения в качестве функции потерь. В этом случае критерий информативности будет выглядеть как

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - c)^2.$$

Как известно, минимум в этом выражении будет достигаться на среднем значении целевой переменной. Значит, критерий можно переписать в следующем виде:

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \left(y_i - \frac{1}{|R|} \sum_{(x_j, y_j) \in R} y_j \right)^2.$$

Мы получили, что информативность вершины измеряется её дисперсией — чем ниже разброс целевой переменной, тем лучше вершина. Разумеется, можно использовать и другие функции ошибки L — например, при выборе абсолютного отклонения мы получим в качестве критерия среднее абсолютное отклонение от медианы.

§3.2 Классификация

Обозначим через p_k долю объектов класса k ($k \in \{1, \dots, K\}$), попавших в вершину R:

$$p_k = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i = k].$$

Через k_* обозначим класс, чьих представителей оказалось больше всего среди объектов, попавших в данную вершину: $k_* = \arg\max\limits_k p_k$.

3.2.1 Ошибка классификации

Рассмотрим индикатор ошибки как функцию потерь:

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq c].$$

Легко видеть, что оптимальным предсказанием тут будет наиболее популярный класс k_* — значит, критерий будет равен следующей доле ошибок:

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq k_*] = 1 - p_{k_*}.$$

Данный критерий является достаточно грубым, поскольку учитывает частоту p_{k_*} лишь одного класса.

3.2.2 Энтропийный критерий

Для классификации в качестве функции потерь мы обычно использовали логистическую функцию потерь. Давайте выведем для неё оптимальный константный прогноз:

$$H(R) = \min_{\sum_k c_k = 1} \left(-\frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} [y_i = k] \log c_k \right).$$

Для вывода оптимальных значений c_k вспомним, что все значения c_k должны суммироваться в единицу. Как известного из методов оптимизации, для учёта этого ограничения необходимо искать минимум лагранжиана:

$$L(c, \lambda) = -\frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} [y_i = k] \log c_k + \lambda \sum_{k=1}^{K} c_k \to \min_{c_k}$$

Дифференцируя, получаем:

$$\frac{\partial}{\partial c_k} L(c, \lambda) = -\frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i = k] \frac{1}{c_k} + \lambda = -\frac{p_k}{c_k} + \lambda = 0,$$

откуда выражаем $c_k = p_k/\lambda$. Суммируя эти равенства по k, получим

$$1 = \sum_{k=1}^{K} c_k = \frac{1}{\lambda}\sum_{k=1}^{K} p_k = \frac{1}{\lambda},$$

откуда $\lambda = 1$. Значит, минимум достигается при $c_k = p_k$, как и в предыдущем случае. Подставляя эти выражения в критерий, получим, что он будет представлять собой энтропию распределения классов:

$$H(R) = -\sum_{k=1}^{K} p_k \log p_k.$$

Из теории вероятностей известно, что энтропия ограничена снизу нулем, причем минимум достигается на вырожденных распределениях ($p_i = 1$, $p_j = 0$ для $i \neq j$). Максимальное же значение энтропия принимает для равномерного распределения. Отсюда видно, что энтропийный критерий отдает предпочтение более «вырожденным» распределениям классов в вершине.

На всякий случай докажем утверждение про максимум энтропии.

Задача 3.1. *Покажите, что энтропия ограничена сверху и достигает своего максимума на равномерном распределении* $p_1 = \cdots = p_K = 1/K$.

Решение. Нам понадобится неравенство Йенсена: для любой вогнутой функции f выполнено

$$f\left(\sum_{i=1}^{n} a_i x_i\right) \geqslant \sum_{i=1}^{n} a_i f(x_i),$$

если $\sum_{i=1}^{n} a_i = 1$.

Применим его к логарифму в определении энтропии (он является вогнутой функцией):

$$H(p) = \sum_{k=1}^{K} p_k \log_2 \frac{1}{p_k} \leqslant \log_2\left(\sum_{k=1}^{K} p_i \frac{1}{p_i}\right) = \log_2 K.$$

Наконец, найдем энтропию равномерного распределения:

$$-\sum_{k=1}^{K} \frac{1}{K}\log_2\frac{1}{K} = -K\frac{1}{K}\log_2\frac{1}{K} = \log_2 K.$$

■

3.2.3 Критерий Джини

Рассмотрим ситуацию, в которой мы выдаём в вершине не один класс, а распределение на всех классах $c = (c_1, \ldots, c_K)$, $\sum_{k=1}^{K} c_k = 1$. Качество такого распределения можно измерять, например, с помощью критерия Бриера (Brier score):

$$H(R) = \min_{\sum_k c_k = 1} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} (c_k - [y_i = k])^2.$$

Легко заметить, что здесь мы, по сути, ищем каждый c_k как оптимальную с точки зрения MSE константу, приближающую индикаторы попадания объектов выборки в класс k. Это означает, что оптимальный вектор вероятностей состоит из долей классов p_k:

$$c_* = (p_1, \ldots, p_K)$$

Если подставить эти вероятности в исходный критерий информативности и провести ряд преобразований, то мы получим критерий Джини:

$$H(R) = \sum_{k=1}^{K} p_k(1 - p_k).$$

Задача 3.2. *Иногда критерий Джини записывают в виде*

$$H(R) = \sum_{k \neq k'} p_k p_{k'}.$$

Покажите, что эта запись эквивалентна нашему определению.

Решение.

$$\sum_{k \neq k'} p_k p_{k'} = \sum_{k=1}^{K} p_k \sum_{k' \neq k} p_{k'} = \sum_{k=1}^{K} p_k(1 - p_k).$$

■

Выясним теперь, какой смысл имеет максимизация функционала качества, основанного на критерии Джини. Сразу выбросим из критерия $H(R_m)$, поскольку данная величина не зависит от j и s. Обозначим долю объектов класса k в вершине m через p_{mk}. Преобразуем критерий:

$$- \frac{|R_\ell|}{|R_m|} H(R_\ell) - \frac{|R_r|}{|R_m|} H(R_r) = - \frac{1}{|R_m|} \left(|R_\ell| - \sum_{k=1}^{K} p_{\ell k}^2 |R_\ell| + |R_r| - \sum_{k=1}^{K} p_{rk}^2 |R_r| \right) =$$

$$= \frac{1}{|R_m|} \left(\sum_{k=1}^{K} p_{\ell k}^2 |R_\ell| + \sum_{k=1}^{K} p_{rk}^2 |R_r| - |R_m| \right) = \{|R_m| \text{ не зависит от } j \text{ и } s\} =$$

$$= \sum_{k=1}^{K} p_{\ell k}^2 |R_\ell| + \sum_{k=1}^{K} p_{rk}^2 |R_r|.$$

Запишем теперь в наших обозначениях число таких пар объектов (x_i, x_j), что оба объекта попадают в одно и то же поддерево, и при этом $y_i = y_j$. Число объектов класса k, попавших в поддерево ℓ, равно $p_{\ell k}|R_\ell|$; соответственно, число пар объектов с одинаковыми метками, попавших в левое поддерево, равно $\sum_{k=1}^{K} p_{\ell k}^2 |R_\ell|^2$. Интересующая нас величина равна

$$\sum_{k=1}^{K} p_{\ell k}^2 |R_\ell|^2 + \sum_{k=1}^{K} p_{rk}^2 |R_r|^2. \tag{3.1}$$

Заметим, что данная величина очень похожа на полученное выше представление для функционала Джини. Таким образом, максимизацию критерия Джини можно условно интерпретировать как максимизацию числа пар объектов одного класса, оказавшихся в одном поддереве. Более того, иногда функционал Джини определяют именно через выражение (3.1).

4 Обучаем дерево

Задача 4.1. *Винни-Пух регулярно ходит в гости к кролику и ест мёд. Иногда он съедает слишком много и застревает в двери. Кролик собрал выборку из числа горшков съеденного мёда, x и фактов застревания в двери, y.*

x_i	1	4	2	3	3	1
y_i	0	1	1	0	1	0

Помогите Кролику построить дерево для классификации. В качестве критерия информативности используется критерий Джини. Дерево строится до глубины равной трём.

Решение. Функционал для оптимизации вышлядит как

$$Q(R_m, j, s) = H(R_m) - \frac{|R_\ell|}{|R_m|}H(R_\ell) - \frac{|R_r|}{|R_m|}H(R_r) \to \max_{j,s}.$$

Можно переписать его для рассматриваемой пары листьев как

$$\frac{|R_\ell|}{|R_m|}H(R_\ell) + \frac{|R_r|}{|R_m|}H(R_r) \to \min_{j,s}.$$

В качестве $H(R)$ мы берём критерий Джини. Посчитаем неопределённость в корне до первого разбиения

$$H(R_0) = J(R_0) = 1 - \left(\frac{1}{2^2} + \frac{1}{2^2}\right) = 0.5$$

Обычно в качестве точек для разбиения рассматривают величины $\gamma_k = \frac{x_k + x_{k+1}}{2}$. У нас есть три точки для разбиения.

x_i	1	1	2	3	3	4
y_i	0	0	1	0	1	1

x_i	1	1	2	3	3	4
y_i	0	0	1	0	1	1

x_i	1	1	2	3	3	4
y_i	0	0	1	0	1	1

Посчитаем значение функционала качества для каждого из разбиений. Если разбиение будет сделано по порогу $\frac{1+2}{2} = 1.5$, в первую вершину попадут объекты только нулевого класса. В левой вершине не будет никакой неопределенности, критерий Джини будет равен нулю. В правой вершине четверть объектов будет нулевого класса. Чтобы посчитать информативность разбиения, нам надо на последнем шаге взвесить оба листа на размеры их подвыборок

$$J_1(R_L) = 1 - (1^2) = 0$$
$$J_1(R_R) = 1 - ((3/4)^2 + (1/4)^2) = 0.375$$
$$J(R_1) = 2/6 \cdot 0 + 4/6 \cdot 0.375 = 0.25.$$

По аналогии посчитаем информативность разбиения, сделанного по порогу $\frac{2+3}{2} = 2.5$

$$J_1(R_L) = 1 - (1/3^2 + (2/3)^2) = 0.44$$
$$J_1(R_R) = 1 - (1/3^2 + (2/3)^2) = 0.44$$
$$J(R_1) = 1/2 \cdot 0.44 + 1/2 \cdot 0.44 = 0.44.$$

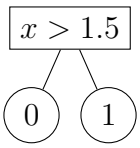
Остаётся порог $\frac{3+4}{2} = 3.5$

$$J_1(R_L) = 1 - ((2/5)^2 + (3/5)^2) = 12/25$$
$$J_1(R_R) = 0$$
$$J(R_1) = 5/6 \cdot 12/25 + 1/6 \cdot 0 = 0.4.$$

Самое маленькое значение $J(R_1)$ нам даёт первое разбиение. Оно же даст нам самое большое значение

$$J(R_0) - J(R_1) = J(R_0) - \frac{|R_\ell|}{|R_0|}H(R_\ell) - \frac{|R_r|}{|R_0|}H(R_r).$$

Выходит, что если $x > 1.5$, мы прогнозируем 1, если меньше, тогда 0.

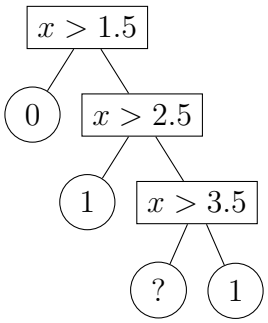


В правом листе все объекты нулевого класса. Продолжить растить дерево мы можем только из левого листа. В рамках него мы можем сделать два разбиения

x_i	2	3	3	4
y_i	1	0	1	1

x_i	2	3	3	4
y_i	1	0	1	1

Они оба обладают одинаковой информативностью, поэтому выберем их в любом порядке. Итоговое дерево имеет вид



В нашем примере не очень понятно, что прогнозировать если мы оказываем в листе, соответствующему $x = 3$. Можно попробовать прогнозировать 1 с вероятностью 0.5.

■

5 Дивергенция Кульбака-Лейблера

Дивергенция Кульбака-Лейблера тесно связана с энтропией. Давайте проинтерпретируем её в том же контексте, что и энтропию. Пусть распределение p – это то, как Женя загадывает величину X, а распределение q – это моё мнение о нём.

X	бустинг	логрег	EM-алгоритм	калик
p	0.5	0.25	0.125	0.125
q	0.125	0.125	0.25	0.5

Отгадывающий не знает распределения p и задаёт вопросы по q.

Кросс-энтропией $CE(p||q)$ называют среднее количество вопросов, которое необходимо задать, если я использую для угадывания распределение q, а в реальность варианты загадываются по распределению p.

$$CE(p||q) = -\mathbb{E}_p(\ln q(v)) = -\sum_v p(v) \cdot \ln q(v)$$

Дивергенцией Кульбака-Лейблера $KL(p||q)$ называются количество лишних вопросов, которые мне пришлось задать Жене

$$KL(p||q) = CE(p||q) - H(p) = \mathbb{E}_p(-\ln q(v)) - \mathbb{E}_p(-\ln p(v)) = \sum_v p(v) \cdot \ln \frac{p(v)}{q(v)}$$