Машинное обучение 1
Семинар 7
Решающие деревья

1 Выбор предикатов в вершинах

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. Далее мы обсудим, какие именно критерии информативности часто используют в задачах регрессии и классификации.

§1.1 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.

§1.2 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: $k_* = \arg\max_k p_k$.

1.2.1 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.

1.2.2 Критерий Джини

Рассмотрим ситуацию, в которой мы выдаём в вершине не один класс, а распределение на всех классах $c = (c_1, \dots, c_K)$, $\sum_{k=1}^K c_k = 1$. 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:

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

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 1.1. *Иногда критерий Джини записывают в виде*

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

Покажите, что эта запись эквивалентна нашему определению.

Решение.

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

■

Задача 1.2. *Рассмотрим вершину m и объекты R, попавшие в нее. Поставим в соответствие вершине m алгоритм $a(x)$, который выбирает класс случайно, причем класс k выбирается с вероятностью p_k. Покажите, что матожидание частоты ошибок этого алгоритма на объектах из R_m равно индексу Джини.*

Решение.

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

■

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

$$\sum_{k=1}^{K} p_{\ell k}^2 |R_\ell|^2 + \sum_{k=1}^{K} p_{rk}^2 |R_r|^2. \tag{1.1}$$

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 (1.1).

1.2.3 Энтропийный критерий

Мы уже знакомы с более популярным способом оценивания качества вероятностей — логарифмическими потерями, или логарифмом правдоподобия:

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

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

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:

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

∎

1.2.4 Выбор критерия

Рассмотрим простой пример с двумя классами. Пусть в текущую вершину попало 400 объектов первого класса и 400 объектов второго класса. Допустим, нужно сделать выбор между двумя разбиениями, одно из которых генерирует поддеревья с числом объектов $(300, 100)$ и $(100, 300)$ (первое число в паре — число объектов первого класса в подвыборке, второе —число объектов второго класса), а другое — с числом объектов $(200, 400)$ и $(200, 0)$. Оба разбиения дают ошибку классификации 0.25, но критерий Джини и энтропийный критерий отдадут предпочтение второму разбиению, что логично, поскольку правая вершина окажется листовой и сложность дерева окажется меньше.

В заключение отметим, что нет никаких четких правил для выбора функционала качества, и на практике лучше всего выбирать его с помощью кросс-валидации. Более того, это далеко не самый важный гиперпараметр — так, между критерием Джини и энтропийным критерием нет очень большой разницы с точки зрения результатов.