Машинное обучение, ФКН ВШЭ
Семинар №7

1 Критерии информативности

При построении дерева необходимо задать *функционал* $Q(X, j, s)$, на основе которого осуществляется разбиение выборки на каждом шаге. Рассмотрим различные способы задания таких функционалов в задачах классификации.

Введем обозначения, которыми будем пользоваться. Для вершины m обозначим

- R_m — множество объектов, попавших в эту вершину;
- $N_m = |R_m|$ — количество таких объектов;
- p_{mk} — доля объектов класса k в вершине m, если решается задача классификации:
$$p_{mk} = \frac{1}{N_m} \sum_{(x_i, y_i) \in R_m} [y_i = k];$$
- Через k_m обозначим класс, чьих представителей оказалось больше всего среди объектов, попавших в вершину m:
$$k_m = \arg\max_k p_{mk}.$$

Критерий информативности (impurity criteria, критерий «нечистоты») носит смысл «насколько сильно отличаются целевые переменные объектов из вершины». Например, рассмотрим критерий информативности, который является долей объектов из R_m, которые были бы неправильно классифицированы, если бы вершина m была листовой и относила все объекты к классу k_m:

$$H_E(R_m) = \frac{1}{N_m} \sum_{(x_i, y_i) \in R_m} [y_i \neq k_m].$$

Функционал ошибки, соответствующий критерию информативности H, при ветвлении вершины m обычно определяется как

$$Q(R_m, j, s) = H(R_m) - \frac{N_\ell}{N_m} H(R_\ell) - \frac{N_r}{N_m} H(R_r),$$

где ℓ и r - индексы левой и правой дочерних вершин. Данный функционал необходимо максимизировать.

Задача 1.1. *Покажите, что критерий информативности H_E также можно записать в виде*

$$H_E(R_m) = 1 - p_{m,k_m}$$

Решение. Заметим, что

$$1 = \frac{1}{N_m} \sum_{(x_i,\, y_i) \in R_m} [y_i \neq k_m] + \frac{1}{N_m} \sum_{(x_i,\, y_i) \in R_m} [y_i = k_m]$$

Откуда сразу получаем

$$H_E(R_m) = \frac{1}{N_m} \sum_{(x_i,\, y_i) \in R_m} [y_i \neq k_m] = 1 - p_{m,k_m}$$

■

Критерий H_E является достаточно грубым, поскольку учитывает частоту p_{m,k_m} лишь одного класса. Обычно используют индекс Джини или энтропийный критерий.

2 Индекс Джини

Критерий информативности в этом случае имеет вид

$$H_G(R_m) = \sum_{k \neq k'} p_{mk} p_{mk'}.$$

Задача 2.1. *Покажите, что индекс Джини $H_G(R_m)$ также можно записать в виде:*

$$H_G(R_m) = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) = 1 - \sum_{k=1}^{K} p_{mk}^2.$$

Решение.

$$\sum_{k \neq k'} p_{mk} p_{mk'} = \sum_{k=1}^{K} p_{mk} \sum_{k' \neq k} p_{mk'} = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) = \sum_{k=1}^{K} p_{mk} - \sum_{k=1}^{K} p_{mk}^2 = 1 - \sum_{k=1}^{K} p_{mk}^2.$$

■

Задача 2.2. *Рассмотрим вершину m и объекты R_m, попавшие в нее. Сопоставим в соответствие вершине m алгоритм $a(x)$, который выбирает класс случайно, причем класс k выбирается с вероятностью p_{mk}. Покажите, что матожидание частоты ошибок этого алгоритма на объектах из R_m равно индексу Джини.*

Решение.

$$\mathbb{E} \frac{1}{N_m} \sum_{x_i \in R_m} [y_i \neq a(x_i)] = \frac{1}{N_m} \sum_{(x_i,\, y_i) \in R_m} \mathbb{E}[y_i \neq a(x_i)] = \frac{1}{N_m} \sum_{(x_i,\, y_i) \in R_m} (1 - p_{m,y_i}) =$$

$$= \sum_{k=1}^{K} \frac{\sum_{(x_i,\, y_i) \in R_m} [y_i = k]}{N_m} (1 - p_{mk}) = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}).$$

■

Выясним теперь, какой смысл имеет максимизация функционала, соответствующего критерию информативности Джини. Сразу выбросим из функционала $H_G(R_m)$, поскольку данная величина не зависит от j и s. Преобразуем критерий:

$$
\begin{aligned}
&-\frac{N_\ell}{N_m}H_G(R_\ell) - \frac{N_r}{N_m}H_G(R_r) = -\frac{1}{N_m}\left(N_\ell - \sum_{k=1}^{K} p_{\ell k}^2 N_\ell + N_r - \sum_{k=1}^{K} p_{rk}^2 N_r\right) = \\
&= \frac{1}{N_m}\left(\sum_{k=1}^{K} p_{\ell k}^2 N_\ell + \sum_{k=1}^{K} p_{rk}^2 N_r - N_m\right) = \{N_m \text{ не зависит от } j \text{ и } s\} = \\
&= \sum_{k=1}^{K} p_{\ell k}^2 N_\ell + \sum_{k=1}^{K} p_{rk}^2 N_r.
\end{aligned}
$$

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 $p_{\ell k}N_\ell$; соответственно, число пар объектов с одинаковыми метками, попавших в левое поддерево, равно $\sum_{k=1}^{K} p_{\ell k}^2 N_\ell^2$. Интересующая нас величина равна

$$
\sum_{k=1}^{K} p_{\ell k}^2 N_\ell^2 + \sum_{k=1}^{K} p_{rk}^2 N_r^2.
$$

Заметим, что данная величина очень похожа на полученное выше представление для критерия Джини. Таким образом, максимизацию функционала Джини можно условно интерпретировать как максимизацию числа пар объектов одного класса, оказавшихся в одном поддереве.

3 Энтропийный критерий

Рассмотрим дискретную случайную величину, принимающую K значений с вероятностями $p_1, \dots, p_K$ соответственно. **Энтропия** этой случайной величины определяется как $H(p) = -\sum_{k=1}^{K} p_k \log_2 p_k$.

Задача 3.1. *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 (логарифм 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:

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

∎

Энтропия ограничена снизу нулем, причем минимум достигается на вырожденных распределениях ($p_i = 1$, $p_j = 0$ для $i \neq j$).

Энтропийный функционал определяется как

$$Q_H(R_m, j, s) = H(p_m) - \frac{N_\ell}{N_m} H(p_\ell) - \frac{N_r}{N_m} H(p_r),$$

где $p_i = (p_{i1}, \dots, p_{iK})$ - распределение классов в i-й вершине. Видно, что данный критерий отдает предпочтение более «вырожденным» распределениям классов.

4 Критерии в задачах регрессии

В задачах регрессии, как правило, в качестве критерия выбирают дисперсию ответов в листе:

$$H_R(R_m) = \frac{1}{N_m} \sum_{(x_i, y_i) \in R_m} \left(y_i - \frac{1}{N_m} \sum_{(x_i, y_i) \in R_m} y_j \right)^2.$$

Можно использовать и другие критерии — например, среднее абсолютное отклонение от медианы.