Семинары по решающим деревьям

Евгений Соколов
sokolov.evg@gmail.com

28 сентября 2014 г.

1 Основы

§1.1 Определение

Рассмотрим бинарное дерево, в котором:

- каждой внутренней вершине v приписана функция $\beta_v : \mathbb{X} \to \{0, 1\}$;
- каждой листовой вершине v приписана метка класса $c_v \in Y$.

Рассмотрим теперь алгоритм $a(x)$, который стартует из корневой вершины v_0 и вычисляет значение функции β_{v_0}. Если оно равно нулю, то алгоритм переходит в левую дочернюю вершину, иначе в правую, вычисляет значение предиката в новой вершине и делает переход или влево, или вправо. Процесс продолжается, пока не будет достигнута листовая вершина; алгоритм возвращает тот класс, который приписан этой вершине. Такой алгоритм называется *бинарным решающим деревом*.

§1.2 Построение деревьев

Опишем простейший алгоритм построения бинарного решающего дерева. Начнем со всей обучающей выборки X^ℓ и найдем наилучшее ее разбиение на две части $R_1(j, s) = \{x \,|\, x_j \leqslant s\}$ и $R_2(j, s) = \{x \,|\, x_j > s\}$ с точки зрения заранее заданного критерия $Q(X, j, s)$. Найдя наилучшие значения j и s, создадим корневую вершину дерева, поставив ей в соответствие функцию $[x_j \leqslant s]$. Объекты разобьются на две части — одни попадут в левое поддерево, другие в правое. Для каждой из этих подвыборок повторим процедуру, построив дочерние вершины для корневой, и так далее. Если после очередного разбиения выборки на две части в одной из половин окажутся объекты лишь одного класса, то создадим листовую вершину, которой будет соответствовать класс попавших в нее объектов.

§1.3 Критерии информативности

При построении дерева необходимо задать *критерий информативности* $Q(X, j, s)$, на основе которого осуществляется разбиение выборки на каждом шаге. Рассмотрим различный варианты таких критериев.

Пусть R_m — множество объектов обучающей выборки, попавших в вершину m. Обозначим через p_{mk} долю объектов класса k ($k \in \{1, \ldots, K\}$), попавших в вершину m:

$$p_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} [y_i = k],$$

где $N_m = |R_m|$. Через k_m обозначим класс, чьих представителей оказалось больше всего среди объектов, попавших в вершину m: $k_m = \arg\max\limits_k p_{mk}$.

1.3.1 Ошибка классификации

Вычислим долю объектов из R_m, которые были бы неправильно классифицированы, если бы вершина m была листовой и относила все объекты к классу k_m:

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

Критерий информативности при ветвлении вершины m определяется как

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

где ℓ и r — индексы левой и правой дочерних вершин.

Задача 1.1. *Покажите, что ошибку классификации также можно записать в виде* $F_E(R_m) = 1 - p_{m,k_m}$.

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

1.3.2 Индекс Джини

Функционал имеет вид

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

Критерий информативности определяется так же, как и в предыдущем случае:

$$Q_G(R_m, j, s) = F_G(R_m) - \frac{N_\ell}{N_m} F_G(R_\ell) - \frac{N_r}{N_m} F_G(R_r).$$

Задача 1.2. *Покажите, что индекс Джини* $F_G(R_m)$ *также можно записать в виде* $F_G(R_m) = \sum_{k=1}^K p_{mk}(1 - p_{mk}) = 1 - \sum_{k=1}^K p_{mk}^2$.

Решение.

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

■

Задача 1.3. *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.

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

■

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

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

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

$$\sum_{k=1}^{K}p_{\ell k}^2 N_\ell^2 + \sum_{k=1}^{K}p_{rk}^2 N_r^2. \tag{1.1}$$

Заметим, что данная величина очень похожа на полученное выше представление для критерия Джини. Таким образом, максимизацию критерия Джини можно условно интерпретировать как максимизацию числа пар объектов одного класса, оказавшихся в одном поддереве. Более того, иногда индекс Джини определяют именно через выражение (1.1).

1.3.3 Энтропийный критерий

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

Задача 1.4. *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:

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

■

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

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

1.3.4 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.