

Деревья решений

Воробьёва Мария

- maria.vorobyova.ser@gmail.com
- @SparrowMaria

Критерии качества. Обобщение

Запишем задачу формально: мы находимся в узле R_m и наша задача оптимально разбить на два подмножества:

$$H(R_m) - \frac{|R_l|}{|R_m|} H(R_l) - \frac{|R_r|}{|R_m|} H(R_r) \rightarrow \max$$

Что означает, что $G(Q_m, \theta)$ должен быть минимальным:

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta))$$

Критерии качества. Обобщение

Разбиение будет “хорошим”, если $G(Q_m, \theta)$,будет минимальным:

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta))$$

Для задачи классификации:

Gini:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

Log Loss or Entropy:

$$H(Q_m) = - \sum_k p_{mk} \log(p_{mk})$$

Критерии качества. Обобщение

Для задачи регрессии:

Mean Squared Error:

$$\bar{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y$$
$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2$$

Half Poisson deviance:

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} \left(y \log \frac{y}{\bar{y}_m} - y + \bar{y}_m \right)$$

Mean Absolute Error:

$$\text{median}(y)_m = \text{median}(y)_{y \in Q_m}$$
$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} |y - \text{median}(y)_m|$$

Алгоритмы построения дерева решений.

История

ID3 (Iterative Dichotomiser 3) (Разработан Джоном Р. Квинланом):

- Работает только с дискретной целевой переменной.
- Построенные деревья являются квалифицирующими, то есть каждый лист соответствует конкретному значению целевой переменной.
- Число потомков в узле не ограничено, и дерево может иметь различную структуру.
- Алгоритм не поддерживает обработку пропущенных данных, что требует предварительной обработки данных перед использованием ID3.

C4.5:

- Является продвинутой версией алгоритма ID3.
- Работает с дискретными и непрерывными значениями атрибутов.
- Позволяет обрабатывать пропущенные значения атрибутов.
- Деревья решений, построенные с помощью C4.5, могут быть квалифицирующими, но также и квантифицирующими (допускающими промежуточные значения на листьях).

CART (Classification and Regression Tree) (Разработан Leo Breiman):

- Поддерживает как задачи классификации, так и регрессии.
- Может работать как с дискретными, так и с непрерывными целевыми переменными.
- Деревья, построенные на основе CART, имеют только два потомка для каждого узла, что делает их бинарными.

Как бороться с переобучением

Ранняя остановка:

- ограничение глубины дерева (max_depth)
- ограничение минимального количества наблюдений во внутреннем узле (min_samples_split)
- ограничение минимального количества наблюдений в листе (min_samples_leaf)
- ограничение min_weight_fraction_leaf, узел будет разделен на дочерние узлы только если min_weight_fraction_leaf (взвешенная доля суммарного веса примеров в узле) превысит min_weight_fraction_leaf % от общего суммарного веса всех примеров.
- ограничение на количество листиков - max_leaf_nodes
- минимальный прирост : min_impurity_decrease

Как бороться с переобучением

Стрижка дерева или Pruning

- Вводим штраф за сложность csr_alpha $R_\alpha(T) = R(T) + \alpha|\tilde{T}|$
- $R(T)$ - ошибки дерева, T - количество узлов
- наша задача минимизировать $R_\alpha(T)$

Деревья решений. Плюсы

- Интерпретация
- Универсальность (классификация и регрессия)
- Быстрая обучаемость и предсказания
- Устойчивость к выбросам
- Не требуют масштабирования данных
- Эффективны для больших наборов данных
- Выявление важности признаков
- Легкая обработка пропущенных значений
- Возможность комбинирования в ансамблях

Деревья решений. Минусы

- Склонность к переобучению на сложных данных
- Неустойчивость к небольшим изменениям данных
- Сложность построения оптимальных деревьев с большим количеством признаков
- Не всегда гарантировано нахождение глобально оптимального решения
- Могут создавать слишком сложные модели, которые трудно интерпретировать
- Подвержены проблемам с несбалансированными данными в задачах классификации



УНИВЕРСИТЕТ
ИННОПОЛИС

ВОПРОСЫ И ОТВЕТЫ