Деревья решений

Гребенюк А.С.

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Кафедра Статистического Моделирования

СПб, 2021

План доклада

Дерево решений

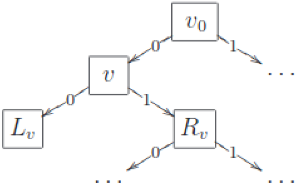

Рис. . построение бинарного дерева решений

Дано

Набор данных

$$\mathbf{X} \in \mathbb{R}^{n\times p}.$$

Зависимые переменные

$$\mathbf{y} \in \mathbb{R}^{n}.$$

$\mathbf{x}_i \in \mathbb{R}^p$ — вектор–строки $\mathbf{X}$.
$X_j \in \mathbb{R}^n$ — вектор–столбцы $\mathbf{X}$.
$y \in \{1, \cdots, K\}$ — задача классификации.
$y \in \mathbb{R}$ — задача регрессии.

Генеральная постановка

Предполагаем, что η и $\boldsymbol{\xi}$ функционально зависимы:

$$\eta = \varphi(\boldsymbol{\xi}) + \varepsilon,$$

φ — неизвестная функция.
$\eta \in \mathbb{R}$ — случайная величина, зависимая переменная.
$\boldsymbol{\xi} \in \mathbb{R}^p$ — случайный вектор, признаки.
$\varepsilon \in \mathbb{R}$ — случайная величина, ошибка.

Выборочная постановка

$$y_i = \varphi(\mathbf{x}_i) + \varepsilon_i,$$

φ — неизвестная функция.
y_i — реализация случайной величины η, зависимая переменная.
$\mathbf{x}_i$ — реализация случайного вектора $\boldsymbol{\xi}$, признаки.
$\varepsilon_i \in \mathbb{R}$ — реализация случайной величины ε, ошибка.

Выбор модели

$$\varphi(\mathbf{x}_i, \mathbf{\Theta}) = \sum_{j=1}^{J} c_j \mathbb{I}_{(\mathbf{x}_i \in R_j)}.$$

Выбор функции потерь

$$\mathrm{RSS} = \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in R_j} (y_i - \varphi(\mathbf{x}_i, \mathbf{\Theta}))^2.$$

Задача оптимизации

$$\mathrm{RSS} = \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in R_j} (y_i - \varphi(\mathbf{x}_i, \mathbf{\Theta}))^2 \to \min_{R_1, \cdots, R_J}.$$

$$\hat{c}_j = \frac{1}{|R_j|} \sum_{\mathbf{x}_i \in R_j} y_i.$$

Пример регрессии

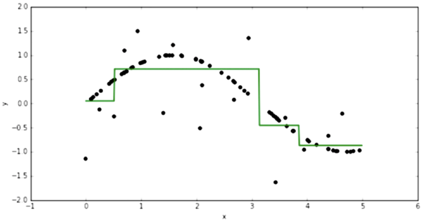

Рис. . регрессионная модель

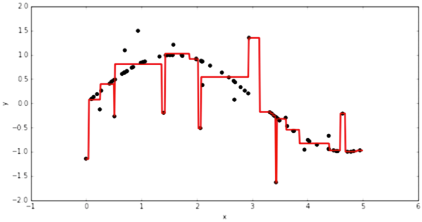

Рис. . переобученная регрессионная модель

Пример регрессии

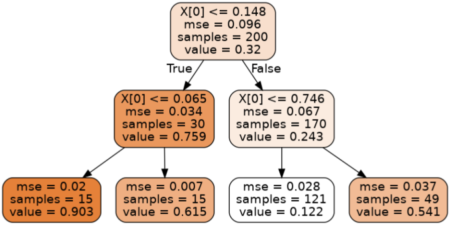

Рис. . регрессионное дерево решений

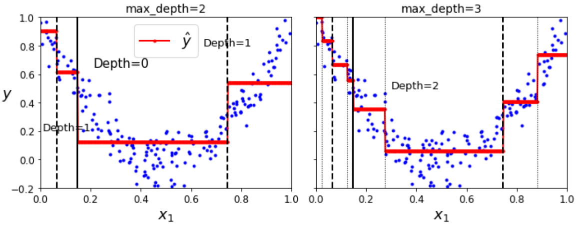

Рис. . модели регрессионного дерева решений

Выбор модели

$$\varphi(\mathbf{x}_i, \mathbf{\Theta}) = \sum_{j=1}^{J} c_j \mathbb{I}_{(\mathbf{x}_i \in R_j)}.$$

Выбор функции потерь

$$p_{jk} = \frac{1}{|R_j|} \sum_{\mathbf{x}_i \in R_j} \mathbb{I}_{(y_i = k)}$$

p_{jk} — доля объектов класса $k \in \{1, \cdots K\}$ в области R_j.

Частота ошибок классификации

$$E = \frac{1}{|R_j|} \sum_{\mathbf{x}_i \in R_j} \mathbb{I}_{(y_i \neq k)}.$$

Индекс Джинни

$$G = \sum_{k=1}^{K} p_{jk}(1 - p_{jk}),$$

$$G = 1 - \sum_{k=1}^{K} p_{jk}^2.$$

Кросс-энтропия

$$CI = -\sum_{k=1}^{K} p_{jk} \log p_{jk}.$$

Индекс Джини

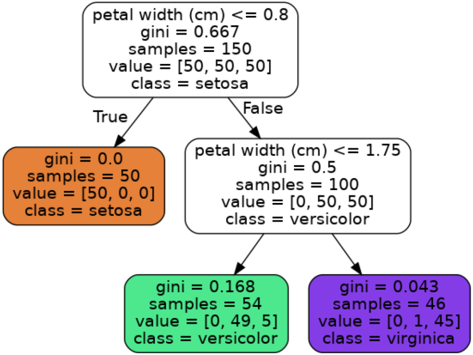

Рис. . дерево решений

Узел на глубине 0:

$$1-\left(\frac{50}{150}\right)^2-\left(\frac{50}{150}\right)^2-\left(\frac{50}{150}\right)^2=1-3\cdot\left(\frac{1}{3}\right)^2=1-\frac{1}{3}=0.666.$$

Узел на глубине 1 слева:

$$1-\left(\frac{50}{50}\right)^2-0-0=0.$$

Узел на глубине 1 справа:

$$1-0-\left(\frac{50}{100}\right)^2-\left(\frac{50}{100}\right)^2=0.5.$$

Узел на глубине 2 слева:

$$1-0-\left(\frac{49}{54}\right)^2-\left(\frac{5}{54}\right)^2=0.168.$$

Узел на глубине 2 справа:

$$1-0-\left(\frac{1}{46}\right)^2-\left(\frac{45}{46}\right)^2=0.043.$$

Impurity

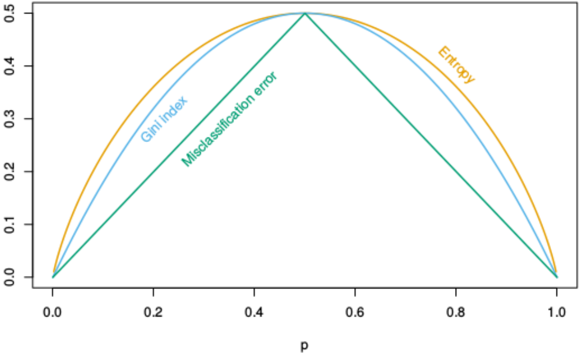

Рис. . Загрязненность *impurity* узла для двухклассовой классификации измеряется как доля индивидов p, отнесенных ко второму классу

CART

Разбиваем данные на две части

$$R_1(j,s) = \{\mathbf{x}_i \in \mathbf{X} | X_j < s\}$$

и

$$R_2(j,s) = \{\mathbf{x}_i \in \mathbf{X} | X_j \geq s\}.$$

Оптимизационная задача

$$\sum_{i:\mathbf{x}_i \in R_1(j,s)} (y_i - \hat{c}_1) + \sum_{i:\mathbf{x}_i \in R_2(j,s)} (y_i - \hat{c}_2) \to \min_{j,s}.$$

где где оценка коэффициента

$\hat{c}_j = \dfrac{1}{|R_j|} \sum\limits_{\mathbf{x}_i \in R_j(j,s)} y_i, \quad j = 1, 2.$

Алгоритм CART — жадный.

ID3

1. $\mathbf{X}$ — обучающая выборка, $\mathbf{y} \in \{1, \cdots, k\}$.
2. Если все $\mathbf{x}_i$ имеют класс k, ставим метку 1 в корень и выходим из цикла.
3. Если ни один $\mathbf{x}_i$ не имеет класс k, ставим метку 0 в корень и выходим из цикла.
4. Предикат $R(\mathbf{x}_i) := \{\mathbf{x}_i | X_j \lesseqgtr s_j\}$ для которого информационная выгода наибольшая.
5. Разбиваем $\mathbf{X}$ на $\mathbf{X}_0$ и $\mathbf{X}_1$ по предикату R

$$\mathbf{X}_0 := \{\mathbf{x}_i \in \mathbf{X} : R(\mathbf{x}_i) = 0\},$$

$$\mathbf{X}_1 := \{\mathbf{x}_i \in \mathbf{X} : R(\mathbf{x}_i) = 1\}.$$

6. Если $\mathbf{X}_0 = \varnothing$ или $\mathbf{X}_1 = \varnothing$, создаем новый лист v, k_v — класс, в котором находится большинство элементов $\mathbf{x}_i$.
7. Иначе создаем внутреннюю вершину v:
   7.1 $R_v = R$;
   7.2 L_v;
   7.3 R_v.

Описанный выше процесс может дать хорошие прогнозы на обучающем наборе, но, вероятно, *переобучится*, что приведет к плохим результатам на тестовых наборах. Почему?

$$\sum_{j=1}^{|T|} \sum_{\mathsf{x}_i \in R_j} (y_i - \hat{y}_{R_j})^2 + \alpha |T|.$$

Критерий остановки:

1. Ограничение макс. глубины дерева;
2. Ограничение мин. числа объектов в листе n_{min};
3. Ограничение макс. количества листьев в дереве;
4. Остановка в случае, если все объекты в листе относятся к одному классу.

Рассмотрим данные о зарплате в бейсболе. Заработная плата имеет цветовую маркировку от низкой (синий, зеленый) до высокой (желтый, красный).

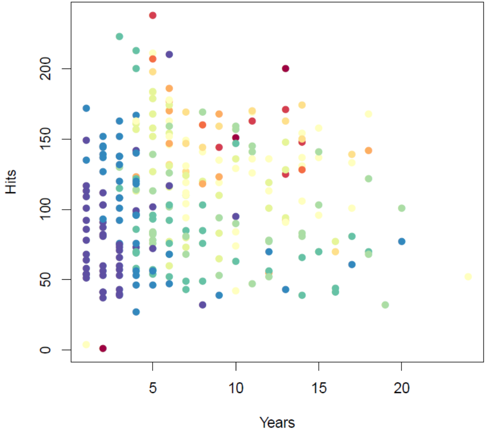

Рис. . данные о бейсболе

Пример классификации: бейсбол

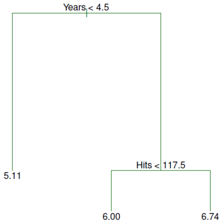

Рис. . классификационная модель

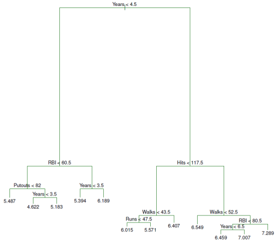

Рис. . переобученная классификационная модель

Пример классификации: бейсбол

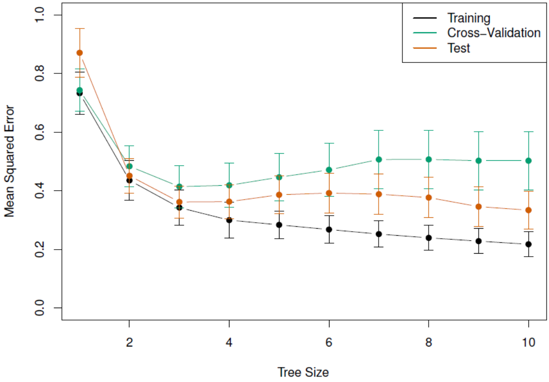

Рис. . среднеквадратическая ошибка в зависимости от размера дерева

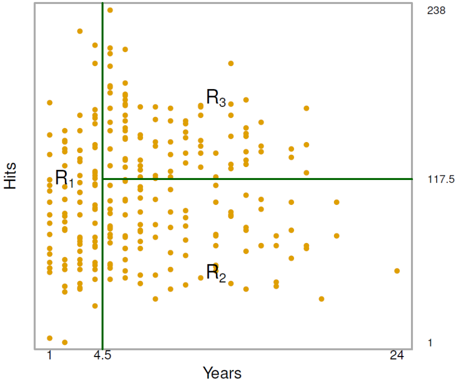

Рис. . результат классификации

$$R_1 = \{X | Years < 4.5\},$$
$$R_2 = \{X | Years \geq 4.5, Hits < 117.5\},$$
$$R_3 = \{X | Years \geq 4.5, Hits \geq 117.5\}.$$

Эти данные содержат бинарный результат HD для 303 пациентов с болью в груди.

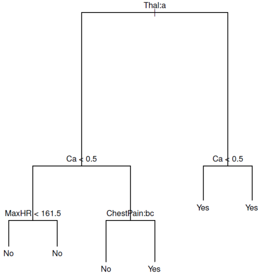

Рис. . классификационная модель

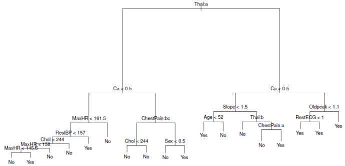

Рис. . переобученная классификационная модель

Пример классификации: сердце

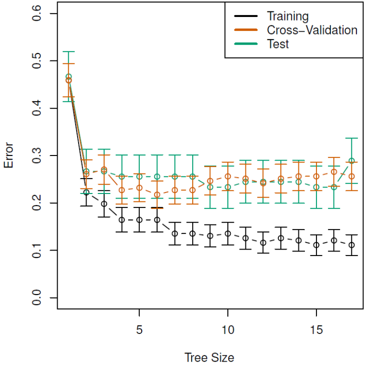

Рис. . среднеквадратическая ошибка в зависимости от размера дерева

Сравнение деревьев с линейными моделями

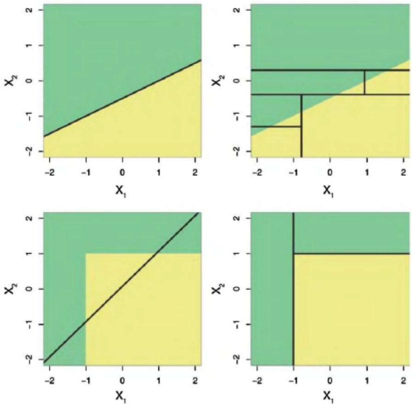

Рис. . сравнение моделей

Преимущества:

1. Классификация + регрессия.
2. Легко визуализировать.
3. Легко интерпретировать.
4. Интуитивность.

Недостатки:

1. Небольшая точность.
2. Переобучение.