Решающие деревья

Московский физико-технический институт, МФТИ

Москва

Вводные замечания

Линейный модели имеют множество преимуществ, такие как простота регализации, небольшое количество параметров, быстрота реализации.
Также возможно их расширение на нелинейные случаи, но часто это возможно за счет эвиристического подхода.
Решающие деревья воспроизводят логические схемы, позволяющие получить окончательное решение о классификации обьекта с помощью ответа на иерархически организованную систему вопросов.
Решающие деревья позовляют восстновить нелинейные зависимости любой сложности.

Определение

Граф - множество вершин и ребер менжду ними

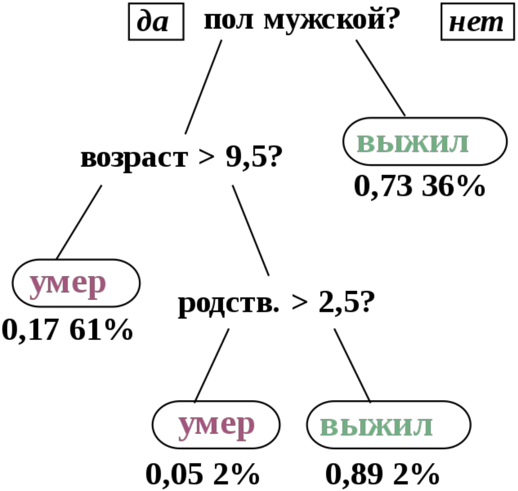
Определение

Дерево - связный граф без циклов

Корневая вершина инцидентна только выходным ребрам, **внутренние вершины**, инцидентные одному входящему ребру и нескольким выходящим, **листья** - концевые вершины, инцидентные только одному входящему ребру.

Основная идея- классификация

Пример решающего дерева



Определение решающего дерева

- каждой внутренней вершине v приписана функция $\beta_v : \mathbb{X} \to \{0, 1\}$
- каждой листовой вершине v приписан прогноз $c_v \in Y$ (в случае с классификацие листу также может быть приписан вектор вероятностей)

Чаще всего использвется предикат, который сравнивает значение одного из признаков с порогом

$$\beta_v(x; j, t) = [x_j < t],$$

также возоможно применение многомерных предикатов.

Построение деревьев

Для любой выборки можно построиоть алгоритм реш. дерева без ошибок, но в этом случае получим алгоритм, у которого не будет *обобщающей способности*.

- пусть задан функционал качества $Q(X, j, t)$.
- найдем лучшие разбиения $R_1(j, t) = \{x | x_j < t\}$ $R_2(j, t) = \{x | x_j \geq t\}$.
- определим оптимальные значения j, t, поставим в соотетствие вершине предикат $[x_j < t]$, так что выборка разбивается на два класса.
- в каждой вершине проверяется, выолнено ли условие останова.

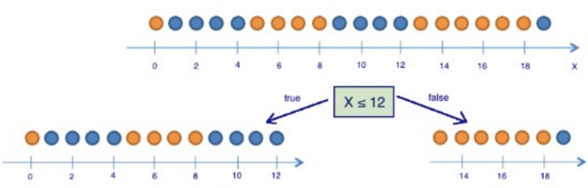
В результате каждый лист содержит ответ, либо значение класса, либо вероятность.

Построение деревьев

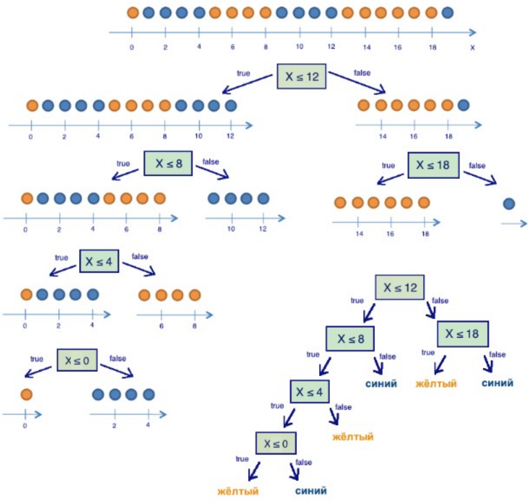
Метод построения решающих деревьев определяется

1. видом предикатов в вершинах;
2. функционалом качества;
3. критерием останова;
4. методом обработки пропущенных значений
5. методом стрижки (pruning)

Пример разбиения



Пример разбиения



Критерий информативности

Рассмотрим функционал следующего вида:

$$Q(R_m, j, s) = H(R_m) - \frac{|R_l|}{|R_m|} H(R_l) - \frac{|R_r|}{|R_m|} H(R_l),$$

где $H(R)$ - критерий информативности, который оценивает качество распределений целевой переменной среди обьектов множества R.
Чем меньще разнообразие целевой переменной, тем меньше должно быть разнообразие критерия информативности.
Значение критерий информативности нужно минимизировать, в то же время максимизириовать фукнционал качества $Q(R_m, j, s)$.

Критерей информативности

Каждй лист дерева должен соответствовать некоторой константе или классу. В связи с этим можно предложить следующую функцию критерия

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c),$$

где $L(y_i, c)$ - некоторая функция потерь

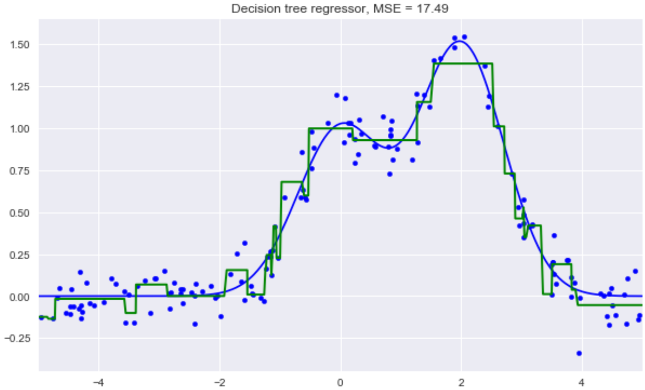
Деревья регрессий

Для обычного квадрата разности

$$H(R) = \min_{c \in Y} \frac{1}{|R| \sum\limits_{(x_i, y_i) \in R}} (y_i - c)^2.$$

Как известно, оптимальное значение константы можно получить, как среднее $\bar{y} = \sum\limits_{(x_i, y_i) \in R} y_j$.

Регрессия иллюстрация



Деревья классификации

Обозначим через p_k долю обьектов класса k $(k = 1..K)$, попавших в вершину

$$p_k = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i = k],$$

Через k_* - класс, чьих представителей больше всего в вершине $k_* = \arg\max_k p_k$.

Ошибка классификации

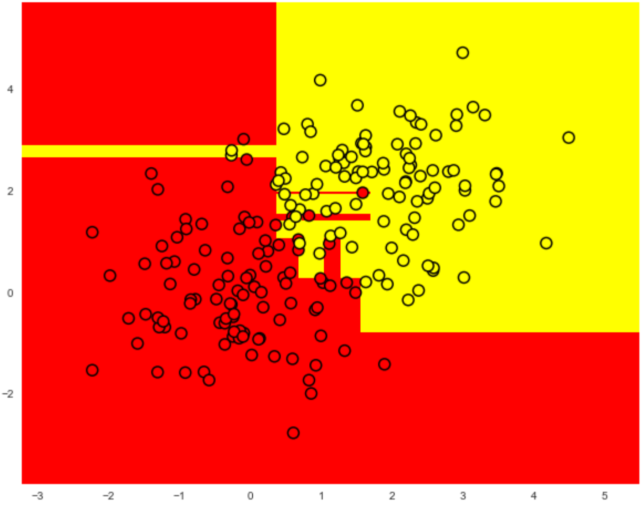
Введем индикатор ошбики для функции потерь:

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq c],$$

оптимальным предсказанием будет наиболее популярный класс k_*

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i \neq k_*] = 1 - p_{k_*}.$$

Классификация иллюстрация



Критерий Джини

Пусть в вершине выдается распределение на всех классах $c = (c_1, .., c_k)$, $\sum_{k=1}^{K} c_k = 1$, посчитаем качество разбиения

$$H(R) = \min_{\sum_k c_k = 1} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} (c_k - [y_i = k])^2,$$

при том, что оптимальный вектор вероятностей состоит из долей классов $c_* = (p_1, .., p_k)$, при этом получаем критерий Джини

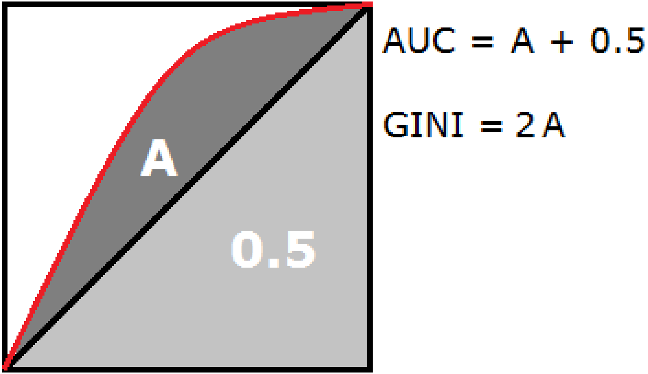
$$H(R) = \sum_{k=1}^{K} p_k (1 - p_k).$$

Критерий Джини

То же самое

$$H(R) = 1 - \sum_{k=1}^{K} p_k^2,$$

максимизацию этого критерия можно интерпретировать как максимизацию числа пар объектов одного класса, оказавшихся в одном поддереве



Энтропийный критерий

Приведем функцию аналго логарифмом правдоподобия:

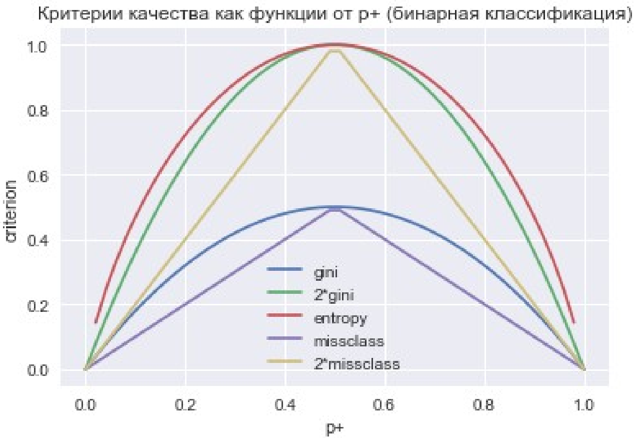
$$H(R) = \min_{\sum_k c_k = 1} \left(-\frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} [y_i = k] \log c_k \right).$$

пользуясь ограничением $\sum_k c_k = 1$, будем искать минимум лагранжиана

$$H(R) = -\sum_{k=1}^{K} p_k \log p_k,$$

где достигается минимум/максимум?

Классификация иллюстрация



Критерий останова

1. ограничение максимальной глубины дерева;
2. ограничения минимального числа обьектов в листе;
3. ограничение максимального количества листьев в дереве;
4. случай, когда все обьекты в листе относятся к одному классу;
5. требование, что функционал качества при дроблении улучшается как минимум на s процентов.

Чистка деревьев (puring)

В данном случае длина дерева не ограничивается, строится переобученное дерево, так что в одном листе содержится по одному обьекту. Далее структура оптмизируется для улучшения *обобщающей способности*.

Альтернативы стрижки:

- случайные леса,
- бустинг.

Чистка деревьев

Воспользуемся регуляризирующим функционалом

$$R_\alpha(T) = R(T) + \alpha|T|,$$

можно показать, что существует последовательность деревьева с однаковыми корнями:

$$T_K \subset T_{K-1} \subset \ldots \subset T_0.$$

На последнем этапе выбирается оптимальное дерево по **отложенной выборке** или с помоьщью **кросс-валидации**.

Обработка пропущенных значений

Пусть для некоторых обьектов V_j не известно значения признаков. Исключим обьекты из выборки с поправкой на потерю информации:

$$Q(R, j, s) \approx \frac{|R \setminus V_j|}{|R|} Q(R \setminus V_j, j, s),$$

Если обьект попал в вершину, предикат которого не может быть вычислен из-за пропуска, то прогнозы для него вычисляются в обоих поддеревьях, и затем усредняются с весами, пропорциональных числу обучающих обьектов в этих поддеревьях.

Учет категориальных признаков

Пусть признак x_j имеет ножество значений $Q = \{u_1, .., u_q\}$, $|Q| = q$, $Q = Q_1 \sqcup Q_2$ $\beta(x) = [x_j \in Q_1]$

Метод построения деревьев

- **ID3** использует энтропийный критерий. Строит дерево до тех пор, пока в каждом листе не окажутся обьекты одного класса, либо пока разбиение вершины дает уменьшение энтропийного критерия.
- **C4.5** использует нормированный энтропийный критерий. Критерий останова - ограничений на число обьектов в листе.
- **CART** критерий Джини. Post-pruning. Для обработки пропусков - метод суррогатных предикатво

Преимущества и недостатки деревьев

Плюсы:

1. Правила классификации как правило поддаются интерпретации;
2. решения для многомерных выборок можно легко визуализировать;
3. Небольшое чисо параметров, небольшое время обработки данных.

Минусы:

1. Чувствительность к нембольшим изменениям входных данных;
2. разделюяющая граница - прямые параллельные осям;
3. поиск оптимального дерева не выгодная с вычислительной точки зрения.

Дальнейшее применение

Наиболее эффективны суперпозиции деревьев, которые называются **ансамблями**. При этом отдельные деревья получаются переобученными.
Stacking, blending - комбинирование различных методов, очень эффектвные способы повышения точности алгоритмов.