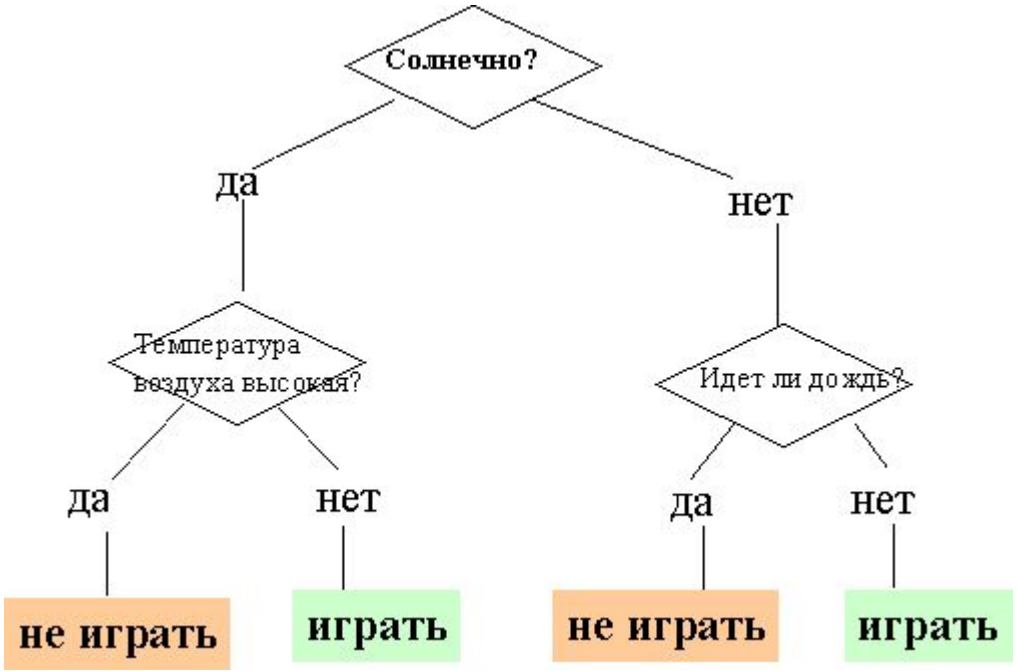
Деревья принятия решений

Определение



Дерево решений – представленный в виде связного ациклического графа план, при помощи которого оценивается значение целевого атрибута объекта по набору значений независимых атрибутов.

Деревья решений

Если зависимая переменная принимает дискретные значения – решает задачу классификации. Если непрерывные – задачу регрессии (численного прогнозирования).

Впервые предложены в конце 50х годов прошлого века.

При проходе от корня к листьям дерева определяется значение зависимой переменной. Внутренний узел представляет разбиение множества возможных значений той или иной независимой переменной. Атрибуты, соответствующие внутренним узлам дерева – **атрибуты расщепления** (прогнозирующие атрибуты).

Каждая ветвь от внутреннего узла отмечается **предикатом расщепления**.

Информация об атрибутах и предикатах расщепления в узле – **критерий расщепления**.

Пример дерева и исходных данных

Решение	Дом	Возраст > 40	Образование	Доход > 5k
Отказать	Нет	Да	Среднее	Нет
Выдать	Нет	Да	Специальное	Да
Выдать	Да	Да	Среднее	Нет
Выдать	Да	Нет	Высшее	Нет
Отказать	Да	Нет	Среднее	Нет
Выдать	Да	Нет	Специальное	Нет
...	...	...	...	...

Преимущества и недостатки

Преимущества деревьев решений:

- Просты в понимании и интерпретации.
- Не требуют подготовки данных.
- Используют модель «белого ящика».
- Позволяют оценить модель при помощи статистических тестов.
- Дают возможность извлекать из базы данных правила на естественном языке.
- Позволяют создавать классификационные модели в тех областях, где аналитику достаточно сложно формализовать знания.
- Алгоритм конструирования дерева решений не требует от пользователя выбора входных атрибутов.
- Быстро обучаются.

Преимущества и недостатки

Недостатки деревьев решений

- Проблема получения оптимального дерева решений бывает NP-полной.
- Могут появиться слишком сложные конструкции, которые при этом недостаточно полно представляют данные.
- Существуют концепты, которые сложно понять из модели, так как модель описывает их сложным путем.
- Для данных, которые включают категориальные переменные с большим набором уровней, больший информационный вес присваивается тем атрибутам, которые имеют большее количество уровней.

Общий алгоритм построения

1. Выбираем целевой атрибут
2. Выбираем критерий расщепления
3. Разделяем обучающую выборку
4. Исключаем атрибут расщепления из выборки
5. Для всех полученных подвыборок переходим на шаг 2

Информационная энтропия

Ансамбль – множество сообщений, каждому из которых соответствует вероятность посылки. Пусть X = {x_1, x_2, …, x_n} – наш ансамбль. Соответственно имеем p(x_1) = p_1 , p(x_2) = p_2, …, p(x_n) = p_n.

Если x_1, x_2, …, x_n независимы и некоторый x_i обязательно отправляется, то $\sum_{i=1}^{n} p_i = 1$.

Мера средней неопределённости ансамбля до посылки сообщения – информационная энтропия ансамбля.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

Информационная энтропия

Информационная энтропия:

- Мера неопределённости выбора сообщения из ансамбля
- Численно равна среднему количеству бит, необходимых для однозначной кодировки всех сообщений ансамбля

Условная энтропия: для ансамблей, в которых известна вероятность появления одного сообщения после другого, или для описания потерь в канале с помехами

$$H(Y \mid x) = -\sum_i p(y_i) \log p(y_i \mid x)$$

Взаимная энтропия

Взаимная энтропия двух ансамблей:

$$H(X \mid Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i \mid y_j) = \sum_j p(y_j) H(X \mid y_j)$$

Некоторые свойства энтропии

Энтропия:

1. Неотрицательна: H(X)≥0
2. Ограничена сверху: $H(X) = -\sum_{i=1}^{n} p_i \log p_i \le \log n$
3. Для независимых A и B справедливо: H(AB) = H(A)+H(B)

Взаимная информация

Взаимная информация (information gain):

$I(Y|X) = H(Y) – H(Y|X)$ – мера неопределённости, снятой посылкой сообщения из ансамбля.

В случае с конструированием деревьев решений целесообразно использовать её в качестве критерия выбора новых атрибутов расщепления.

Пороговая энтропия

При наличии непрерывных атрибутов надо бы поискать пороговые значения, которые надо выставлять в узлах. Для этого тоже можно хорошо приспособить энтропию и information gain. Надо определить, какие значения непрерывных атрибутов дадут наибольший прирост.

Пороговая энтропия:

$$H(Y \mid X:t) = p(X < t)H(Y \mid X < t) + p(X \geq t)H(Y \mid X \geq t)$$

$$I(Y \mid X:t) = H(Y) - H(Y \mid X:t)$$

$$I^*(Y, X) = \max_t I(Y \mid X:t)$$

Алгоритм ID3

На старте имеем таблицу примеров и набор атрибутов с заранее определённым целевым.

1. Если все примеры принадлежат одному классу, возвратим соответствующий лист.
2. Если множество атрибутов пусто, вернуть наиболее часто встречающийся в таблице примеров класс.
3. Найти атрибут с наибольшим приростом информации (а для количественных атрибутов также найти оптимальный порог).
4. Создать узел дерева для найденного атрибута:
    1. Поместить атрибут в узел
    2. Для всех возможных значений атрибута добавить новую ветвь дерева с соответствующим предикатом и рекурсивно вызвать алгоритм для разделённой по атрибуту обучающей выборки
5. Завершить, если
    1. атрибуты закончились
    2. Все элементы таблицы примеров имеют одно значение целевого атрибута