**Дерево решений и MDL**

**Введение**

В работе рассматриваются деревья принятия решений и принцип минимальной длины описания. Целью деревьев является решение серии вопросов (деревьев), которые заканчиваются решениями (листьями). Рассматриваемый как регулирующий механизм, MDL показывает, что лучшая гипотеза для данного набора данных - это то, что приводит к наилучшему сжатию данных.

**Дерево решений**

Здесь у нас приведен метод, в котором мы создаем к'-нарное дерево. Каждый узел представляет собой

вопрос об особенностях, Θn. Каждая ветвь представляет собой ответ на свой исходный узел, один из вариантов "К". Каждый лист представляет собой решение. Дерево решений легко реализуетсяв различных приложениях, и оно полезно как для классификации, так и для регрессии.

В данном докладе представлена нотация двоичного дерева решений, но распространение её на к'-нары - это прямолинейно.

Пусть {xi} будет i-й образец, i = 1, 2, ..., m, каждый из которых имеет n признаков, и Θ - это вопрос, задаваемый в каждом узле.

**Пример 1**

Предположим Х - папайя. Цель состоит в том, чтобы решить вкусно это или нет, с помощью дерева решений. Каждый узел представляет собой вопрос о папайи. Как видно из примера (рис.1), верхний узел запрашивает цвет папайи, который, если цвет не правильный, приводит к решению, что папайя не является вкусной. Если цвет правильный, переходим к следующему узлу и задаем следующий вопрос. Если ответы на вопрос и о цвете, и о мягкости были "правильными" (т.е. согласно нашим определениям вкусной папайи), мы решаем, что папайя вкусная.

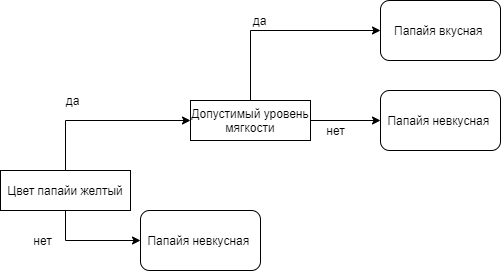
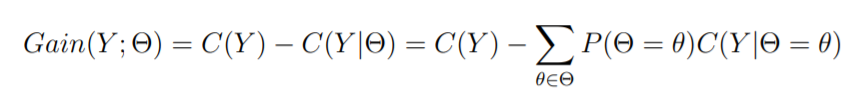


Рис 1. Классическое дерево решений для классификации папайи с использованием двух характеристик: цвета и мягкости.

Функция затрат

Пусть {xi,j} будет j'ой особенностью i'того образца, а {yi}i=1...m - классификациями (метками). Определите коэффициент усиления, где C(x) - функция затрат, в виде



Примеры функций затрат:

Энтропия

Джини коэффициент

Ошибка

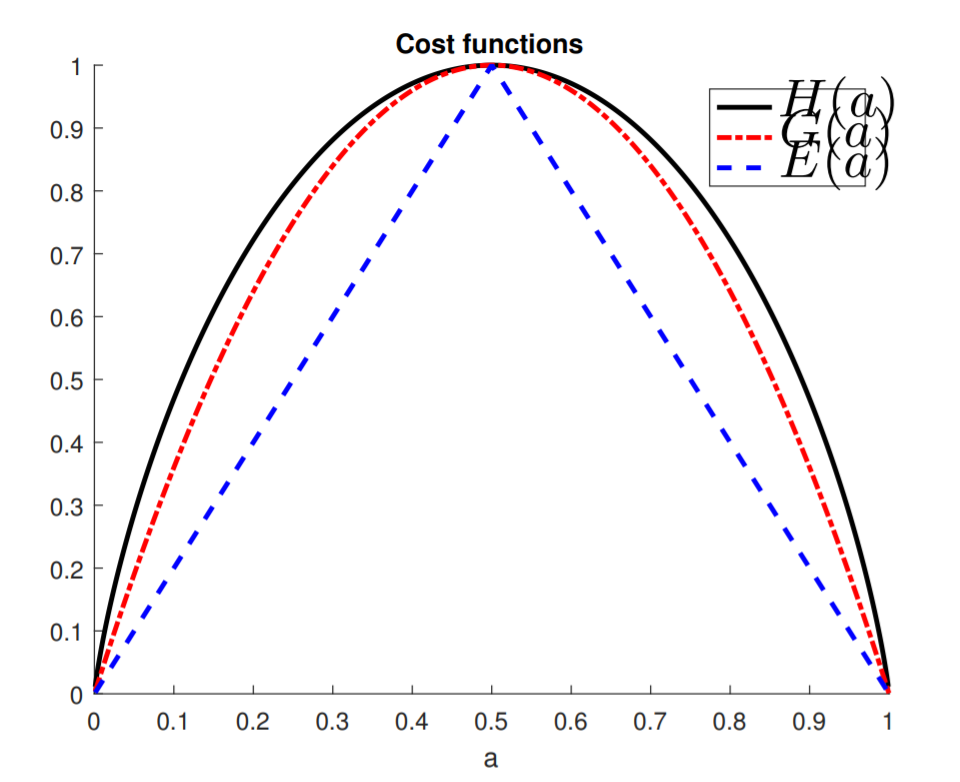


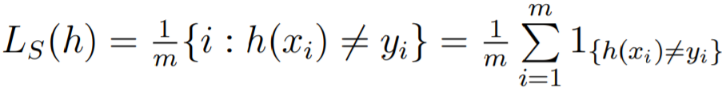
Рис. 2. Функция общей стоимости

Получение энтропии, как стоимости называется информационный выигрыш, и она широко используется в приложениях.

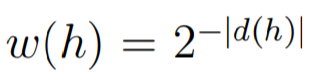


**MDL**

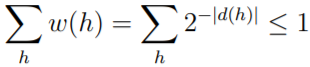
Принцип минимальной длины описания дает описание способа минимизации моделей. Он аналогичен объединению длины (или, в нашем случае, глубины дерева) и стоимости в новую улучшенную функцию стоимости. Цель MDL можно описать как "поиск регулярности в данных", где "регулярность" означает сжимаемость. MDL сочетает в себе те знания, которые он получает, рассматривая обучение как сжатие данных: он говорит нам, что для заданного набора гипотез h и набора данных S мы должны попытаться найти гипотезу или комбинацию гипотез в H, которая сжимает S больше всего. Процедуры MDL автоматически и по своей природе защищают от переподгонки и могут быть использованы для оценки как параметров, так и структуры модели. Кроме того, в отличие от других статистических методов, процедуры MDL имеют четко выраженную статистическую интерпретацию

Пусть "P" будет распределением, - набор образцов, а h - гипотеза (функция) Определим d(h) как описание h с префиксом свойства( такого, что в системе нет полного описания, которое было бы префиксом любого другого описание в системе). Определим как теоретический риск, а эмпирический

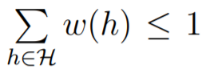
риск как



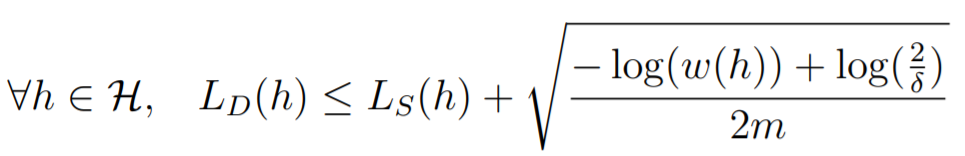
Определим длину описание h как

Также запишем неравенство Крафта

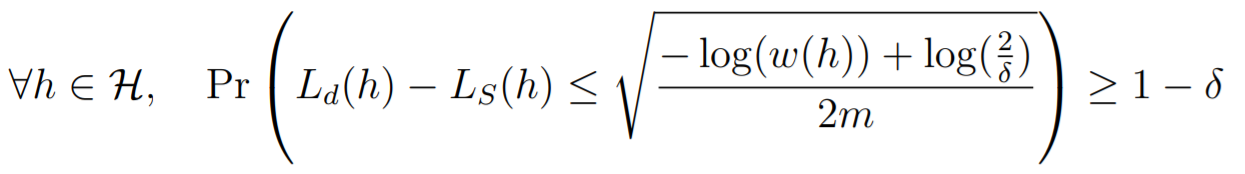
**Теорема 1 (MDL-связь):**

Пусть w будет таким что тогда с вероятностью от

до мы имеем:



Или



Доказательство наличия связи MDL основано в основном на неравенстве Хеффдинга, где при задании некоторых свободных параметров мы приводим неравенство к тому, чтобы оно воспринималось как необходимая нам граница. Отсюда доказательство прямолинейно.

**Обрезка**

Оптимальный конечный размер дерева в алгоритме дерева решений не является тривиальным вопросом. Слишком большое дерево рискует превысить обучающие данные, что сделает его плохо обобщаемым для новых примеров. И наоборот, слишком маленькое дерево может не захватить важную структурную информацию о пространстве выборки. Однако определить, когда алгоритм дерева должен остановиться, сложно, так как нельзя знать, приведет ли добавление одного лишнего узла к резкому снижению точности. Эта проблема, известная как горизонтальный эффект, обычно решается путем выращивания дерева до тех пор, пока каждый узел не будет содержать небольшое количество экземпляров, а затем с помощью обрезки удалять узлы, не предоставляющие дополнительную информацию. При обрезке следует уменьшить размер обучающего дерева, не снижая его предиктивную точность, измеряемую набором для перекрестной проверки. Существует множество методов обрезки деревьев, которые отличаются друг от друга с точки зрения измерения, используемого для оптимизации производительности.

1) Методы:

Вообще говоря, методы обрезки деревьев можно классифицировать либо как обрезку сверху вниз - траверсные узлы и обрезку поддеревьев, начинающуюся от корня, либо как обрезку снизу вверх - начинающуюся от узлов листа. Каждый метод обрезки имеет свои преимущества и недостатки, которые, если они известны и понятны, могут помочь нам в выборе наиболее подходящего метода в каждом конкретном случае. Здесь мы познакомимся с двумя популярными алгоритмами, один из которых работает по принципу "сверху вниз", другой - "снизу вверх".

**1) Обрезка по сложности (сверху вниз) -**

Давайте определим:

 T0 начальное дерево, а Tm только корень

коэффициент ошибок дерева - err(T, S), где T - это текущее дерево, а S – это набор данных.

Функция – prune(T, t) - определяет результирующее дерево после удаления поддеревьев t из T.

Функция - листья(T) - определяет количество листьев на дереве T.

На каждом шаге i дерево формируется путем удаления поддерева t из дерева i - 1 и замены его узлом листа, значение которого выбирается как в алгоритме построения дерева. Сначала мы находим ошибку err(Ti, S) для текущего дерева Ti, а затем ищем для поддерева, которое минимизирует выражение:

**(err(prune(Ti,t),S)−err(Ti,S))/(|leaves(Ti)|−|leaves(prune(Ti,t))|)**

После создания серии деревьев, лучшее дерево выбирается на основе обобщенной точности, измеренной с помощью обучающего набора или путем перекрестной проверки.

2) Уменьшение ошибок при обрезке (снизу вверх) -

Начиная с листьев, каждый узел заменяется на наиболее популярный класс. Если точность прогноза не изменяется, то изменение сохраняется. Очень простая в применении, эта техника обладает дополнительным преимуществом - быстротой.