Лекция 6

Решающие деревья

Лектор – *Сенько Олег Валентинович*

Методы машинного обучения

Содержание лекции

Решающие деревья

Решающие деревья воспроизводят логические схемы, позволяющие получить окончательное решение о классификации объекта с помощью ответов на иерархически организованную систему вопросов. Причём вопрос, задаваемый на последующем иерархическом уровне, зависит от ответа, полученного на предыдущем уровне. Подобные логические модели издавна используются в ботанике, зоологии, минералогии, медицине и других областях. Пример, решающего дерева, позволяющая грубо оценить стоимость квадратного метра жилья в предполагаемом городе приведена на рисунке 1. Схеме принятия решений, изображённой на рисунке 1, соответствует связный ориентированный ациклический граф – ориентированное дерево. Дерево включает в себя корневую вершину, инцидентную только выходящим рёбрами, внутренние вершины, инцидентную одному входящему ребру и нескольким выходящим, и листья – концевые вершины, инцидентные только одному входящему ребру.

.

Решающие деревья

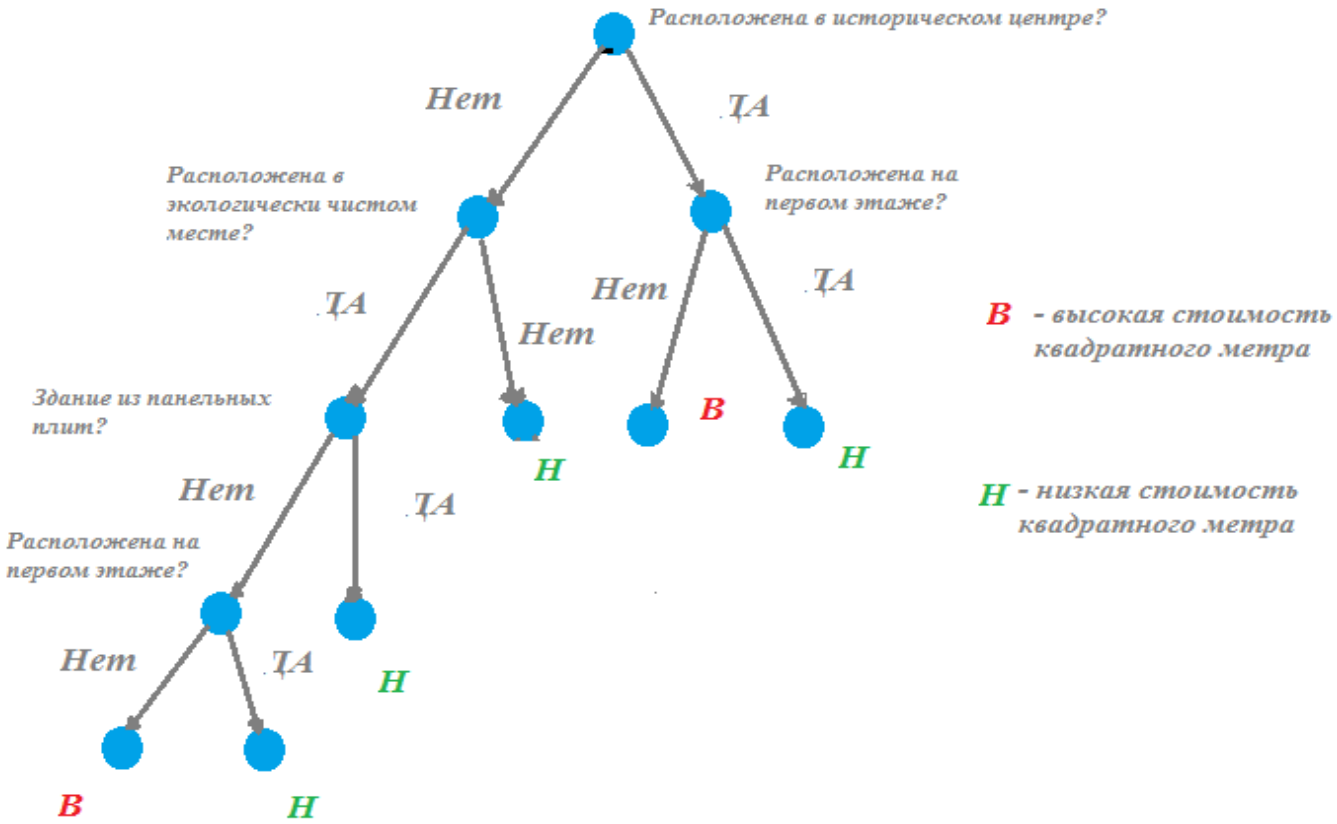

Рис.1

Решающие деревья

Каждой из вершин дерева за исключением листьев соответствует некоторый вопрос, подразумевающий несколько вариантов ответов, соответствующих выходящим рёбрам. В зависимости от выбранного варианта ответа осуществляется переход к вершине следующего уровня. Концевым вершинам поставлены в соответствие метки, указывающие на отнесение распознаваемого объекта к одному из классов. Решающее дерево называется бинарным, если каждая внутренняя или корневая вершина инцидентна только двум выходящим рёбрам. Бинарные деревья удобно использовать в моделях машинного обучения.
Далее графически представлен пример решающего дерева, предназначенного для диагностирования туберкулёза по рентгенограмме и другим показателям. Решающее дерево было построено по выборке из 275 пациентов, для 27% из которых был установлен диагноз туберкулёз.

Пример решающего дерева для диагностики туберкулёза

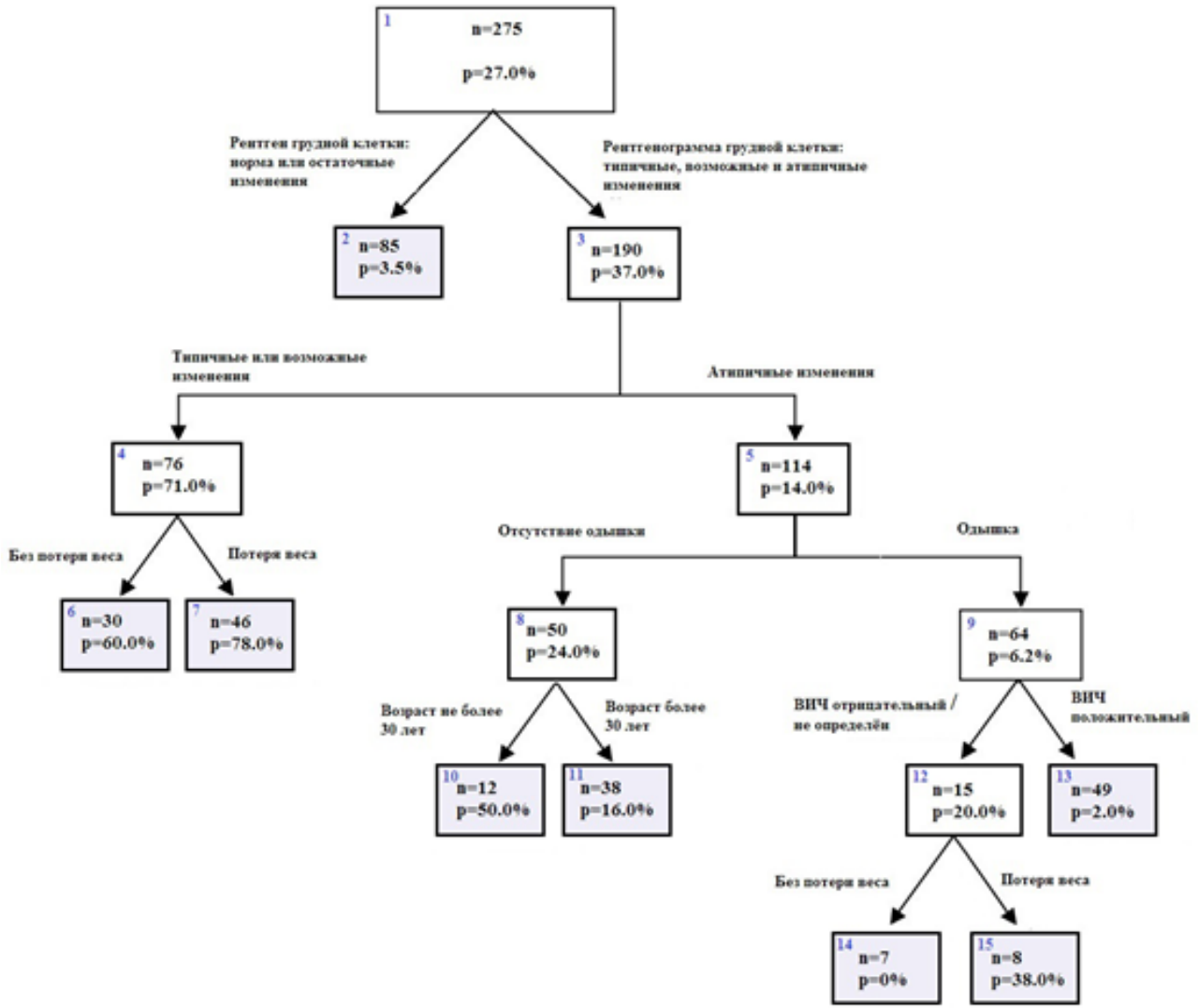

Рис.1

Решающие деревья

Предположим, что бинарное дерево T используется для распознавания объектов, описываемых набором признаков $X_1, \ldots, X_n$.
Каждой вершине ν дерева T ставится в соответствие предикат, касающийся значения одного из признаков. Непрерывному признаку X_j соответствует предикат вида "$X_j \geq \delta_j^\nu$", где δ_j^ν - некоторый пороговый параметр.

Решающие деревья. Обучение

Категориальному признаку $X_{j'}$, принимающему значения из множества $M_{j'} = \{a_1^{j'}, \ldots, a_{r(j')}^{j'}\}$ ставится в соответствие предикат вида "$X_{j'} \in M_{j'}^{\nu 1}$" , где $M_{j'}^{\nu 1}$ является элементом дихотомического разбиения $\{M_{j'}^{\nu 1}, M_{j'}^{\nu 2}\}$ множества $M_{j'}$. Выбор одного из двух, выходящих из вершины ν рёбер производится в зависимости от значения предиката.

Процесс распознавания заканчивается при достижении концевой вершины (листа). Объект относится классу согласно метке, поставленной в соответсттвие данному листу.

Обучение решающих деревьев Рассмотрим задачу распознавания с классами $K_1, \ldots, K_L$. Обучение производится по обучающей выборке $\widetilde{S}_t$ и включает в себя поиск оптимальных пороговых параметров или оптимальных дихотомических разбиений для признаков $X_1, \ldots, X_n$. При этом поиск производится исходя из требования снижения среднего индекса неоднородности в выборках, порождаемых искомым дихотомическим разбиением обучающей выборки $\widetilde{S}_t$.

Индекс неоднородности вычисляется для произвольной выборки $\widetilde{S}$, содержащей объекты из классов $K_1, \ldots, K_L$. При этом используется несколько видов индексов, включая:

- энтропийный индекс неоднородности,
- индекс Джини,
- индекс ошибочной классификации.

Энтропийный индекс неоднородности вычисляется по формуле

$$\gamma_e(\widetilde{S}) = -\sum_{i=1}^{L} P_i \ln P_i, \tag{1}$$

где P_i - доля объектов класса K_i в выборке $\widetilde{S}$. При этом принимается, что $0 \ln(0) = 0$. Наибольшее значение $\gamma_e(\widetilde{S})$ принимает при равенстве долей классов. Наименьшее значение $\gamma_e(\widetilde{S})$ достигается при принадлежности всех объектов одному классу.

Решающие деревья. Обучение

Индекс Джини вычисляется по формуле

$$\gamma_g(\widetilde{S}) = 1 - \sum_{i=1}^{L} P_i^2. \tag{2}$$

Индекс ошибочной классификации вычисляется по формуле

$$\gamma_m(\widetilde{S}) = 1 - \max_{1,\ldots,L}(P_i). \tag{3}$$

Нетрудно понять, что индексы (2) и (3) также достигают минимального значения при принадлежности всех объектов обучающей выборке одному классу.
Предположим, что в методе обучения используется индекс неоднородности γ_*. Для оценки эффективности разбиения обучающей выборки $\widetilde{S}_t$ на непересекающиеся подвыборки $\widetilde{S}_t^l$ и $\widetilde{S}_t^r$ используется уменьшение среднего индекса неоднородности в $\widetilde{S}_t^l$ и $\widetilde{S}_t^r$ по отношению к $\widetilde{S}_t$

Решающие деревья. Обучение

Данное уменьшение вычисляется по формуле

$$\Delta(\gamma_*, \widetilde{S}_t) = \gamma_*(\widetilde{S}_t) - P_l\gamma_*(\widetilde{S}_t^l) - P_r\gamma_*(\widetilde{S}_t^r),$$

где P_l и P_r являются долями $\widetilde{S}_t^l$ и $\widetilde{S}_t^r$ в выборке $\widetilde{S}_t$. На первом этапе обучения бинарного решающего дерева ищется оптимальный предикат соответствующий корневой вершине. С этой целью оптимальные разбиения строятся для каждого из признаков из набора $X_1, \ldots, X_n$. Выбирается признак $X_{i_{max}}$ с максимальным значением индекса $\Delta(\gamma_*, \widetilde{S}_t)$. Подвыбороки $\widetilde{S}_t^l$ и $\widetilde{S}_t^r$, задаваемые оптимальным предикатом для $X_{i_{max}}$ оцениваются с помощью критерия остановки. В качестве критерия остановки может быть использован простейший критерий достижения полной однородности по одному из классов. В случае, если какая-нибудь из выборок $\widetilde{S}_t^*$ удовлетворяет критерию остановки, то соответствующая вершина дерева объявляется концевой и для неё вычисляется метка класса. В случае, если выборка $\widetilde{S}_t^*$ не удовлетворяет критерию остановки, то формируется новая внутренняя вершина, для которой процесс построения дерева продолжается.

Решающие деревья. Обучение

Однако вместо обучающей выборки $\widetilde{S}_t$ используется соответствующая вновь образованной внутренней вершине ν выборка $\widetilde{S}_\nu$, которая равна $\widetilde{S}_t^*$. Для данной выборки производятся те же самые построения, которые на начальном этапе проводились для обучающей выборки $\widetilde{S}_t$. Обучение может проводиться до тех пор, пока все вновь построенные вершины не окажутся однородными по классам. Такое дерево может быть построено всегда, когда обучающая выборка не содержит объектов с одним и тем же значениям каждого из признаков, принадлежащих разным классам. Однако абсолютная точность на обучающей выборке не всегда приводить к высокой обобщающей способности в результате эффекта переобучения.
Одним из способов достижения более высокой обобщающей способности является использования критериев остановки, позволяющих останови процесс построения дерева до того, как будет достигнута полная однородность концевых вершин.

Решающие деревья. Обучение.

Рассмотри несколько таких критериев.
1. Критерий остановки по минимальному допустимому числу объектов в выборках, соответствующих концевым вершинам.
2. Критерий остановки по минимально допустимой величине индекса $\Delta(\gamma_*, \widetilde{S})$. Предположим, что некоторой вершине ν соответствует выборка $\widetilde{S}_\nu$, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\widetilde{S}_\nu^l, \widetilde{S}_\nu^r\}$. Вершина ν считается внутренней, если индекс $\Delta(\gamma_*, \widetilde{S})$ превысил пороговое значение τ и считается концевой в противном случае.
3.Критерий остановки по точности на контрольной выборке. Исходная выборка данных случайным образом разбивается на обучающую выборку $\widetilde{S}_t$ и контрольную выборку $\widetilde{S}_c$. Выборка $\widetilde{S}_t$ используется для построения бинарного решающего дерева. Предположим, что некоторой вершине ν соответствует выборка $\widetilde{S}_\nu$, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\widetilde{S}_\nu^l, \widetilde{S}_\nu^r\}$.

Решающие деревья. Обучение

На контрольной выборке $\widetilde{S}_c$ производится сравнение эффективность распознающей способности деревьев T_ν и T_ν^{++}.
Деревья T_ν и T_ν^{++} включает все вершины и рёбра, построенные до построения вершины ν . В дереве T_ν вершина ν считается концевой. В дереве T_ν^{++} вершина ν считается внутренней, а концевыми считаются вершины, соответствующие подвыборкам $\widetilde{S}_\nu^l$ и $\widetilde{S}_\nu^r$. Распознающая способность деревьев T_ν и T_ν^{++} сравнивается на контрольной выборке $\widetilde{S}_c$. В том, случае если распознающая способность T_ν^{++} превосходит распознающую способность T_ν все дальнейшие построения исходят из того, что вершина ν является концевой. В противном случае производится исследование $\widetilde{S}_\nu^l$ и $\widetilde{S}_\nu^r$.
4. Статистический критерий. Заранее фиксируется пороговый уровень значимости (P<0.05,p<0.01 или p<0.001). Предположим, что нам требуется оценить, является ли концевой вершина , для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\widetilde{S}_\nu^l, \widetilde{S}_\nu^r\}$.

Решающие деревья. Обучение

Исследуется статистическая достоверность различий между содержанием объектов распознаваемых классов в подвыборках $\widetilde{S}_\nu^l$ и $\widetilde{S}_\nu^r$. Для этих целей может быть использованы известные статистический критерий: Хи-квадрат и другие критерии. По выборкам $\widetilde{S}_\nu^l$ и $\widetilde{S}_\nu^r$ рассчитывается статистика критерия и устанавливается соответствующее p-значение. В том случае, если полученное p-значение оказывается меньше заранее фиксированного уровня значимости вершина ν считается внутренней. В противном случае вершина ν считается концевой.

Использование критериев ранней остановки не всегда позволяет адекватно оценить необходимую глубину дерева. Слишком ранняя остановка ветвления может привести к потере информативных предикатов, которые могут быть на самом деле найдены только при достаточно большой глубине ветвления.

Решающие деревья. Подрезка

В связи с этим нередко целесообразным оказывается построение сначала полного дерева, которое затем уменьшается до оптимального с точки зрения достижения максимальной обучающей способности размера путём объединения некоторых концевых вершин. Такой процесс в литературе принято называть «pruning» («подрезка»).
При подрезке дерева может быть использован критерий целесообразности объединения двух вершин, основанный на сравнение на контрольной выборке точности распознавания до и после проведения «подрезки».
Ещё один способ оптимизации обобщающей способности деревьев основан на учёте при «подрезке» дерева до некоторой внутренней вершины ν одновременно увеличения точности разделения классов на обучающей выборке и увеличения сложности, которые возникают благодаря ветвлению из ν.

При этом прирост сложности, связанный с ветвлением из вершины ν, может быть оценён через число листьев в поддереве T_{ν}^{sub} полного решающего дерева с корневой вершиной ν. Следует отметить, что рост сложности является штрафующим фактором, компенсирующим прирост точности разделения на обучающей выборке с помощью включения поддерева T_{ν}^{sub} в решающее дерево. Разработан целый ряд эвристических критериев, которые позволяют оценить целесообразность включения T_{ν}^{sub}. Данные критерии учитывают одновременно сложность и разделяющую способность.