# МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор
Сенько Олег Валентинович

Лекция 8

# 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.

# Пример решающего дерева

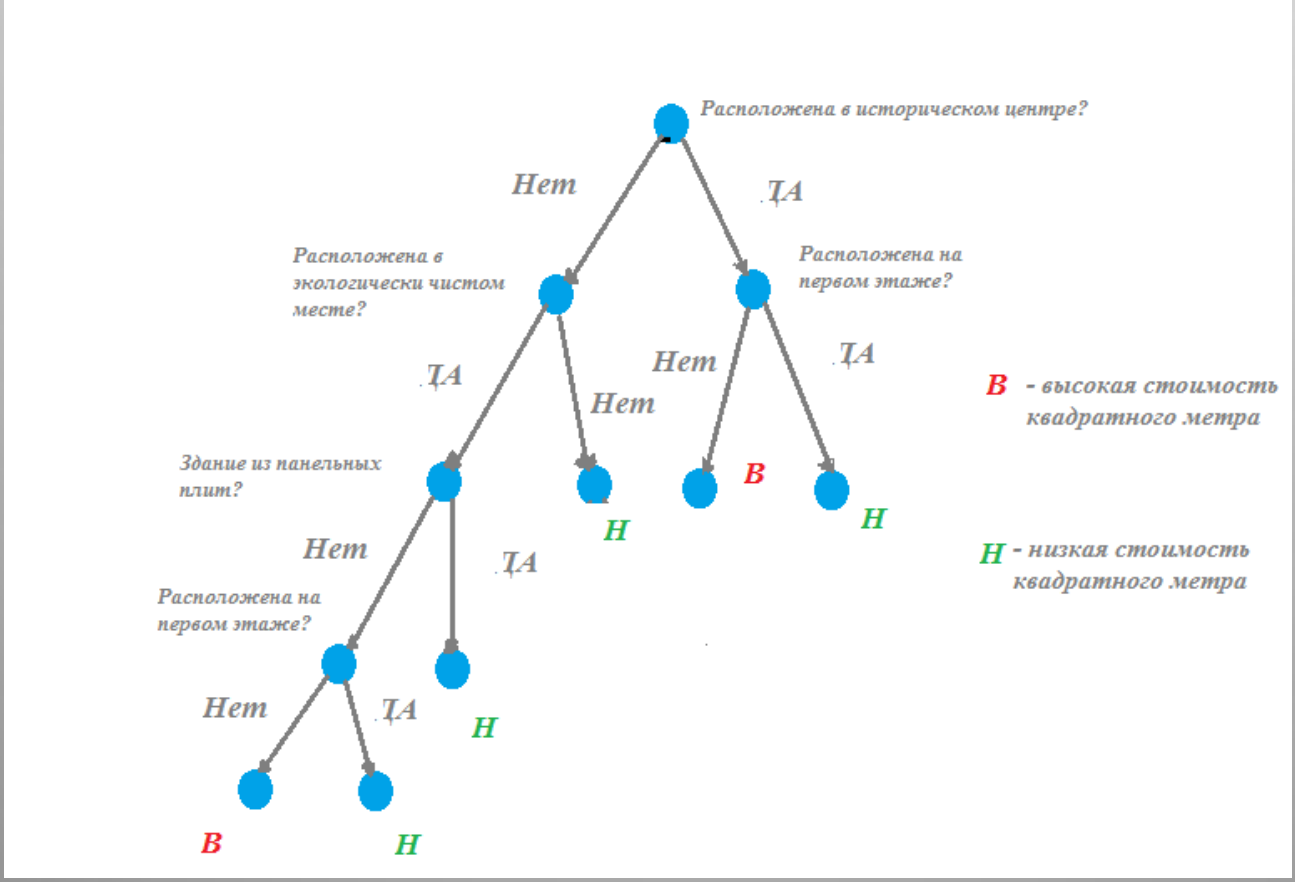


Рис. 1

# Решающие деревья

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.

Каждой из вершин дерева за исключением листьев соответствует некоторый вопрос, подразумевающий несколько вариантов ответов, соответствующих выходящим рёбрам.

# Решающие деревья

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.

# Решающие деревья

Предположим, что бинарное дерево $\mathbf{T}$ используется для распознавания объектов, описываемых набором признаков $X_1, \ldots, X_n$. Каждой вершине дерева $v$ ставится в соответствие предикат, касающийся значения одного из признаков.

Непрерывному признаку $X_{j'}$ соответствует предикат вида "$X_{j'} \geq \delta_{j'}^{v}$", где $\delta_{j'}^{v}$ - некоторый пороговый параметр.

# Решающие деревья

Категориальному признаку $X_{j''}$, принимающему значения из множества $M_{j''} = \{a_{j''}^1, \ldots, a_{j''}^{r_{j''}}\}$ ставится в соответствие предикат вида $"X_{j''} \in M_{j''}^{gv}"$, где $\{M_{j''}^{gv}, M_{j''}^{cv}\}$ - дихотомическое разбиение множества $M_{j''}$ .

Выбор одного из двух, выходящих из вершины $v$ рёбер производится в зависимости от значения предиката.

Обучение производится по обучающей выборке $\tilde{S}_t$ и включает в себя поиск оптимальных разбиений или поиск оптимальных пороговых параметров. .

# Решающие деревья

Рассмотрим задачу распознавания с классами $K_1,\ldots,K_L$ .

На первом этапе обучения бинарного решающего дерева ищется оптимальный предикат соответствующий корневой вершине. При этом поиск производится исходя из требования уменьшения среднего индекса неоднородности в выборках $\tilde{S}_t^l$ и $\tilde{S}_t^r$ , принадлежащих дихотомическому разбиения обучающей выборки , задаваемому предикатом.

# Решающие деревья

Индекс неоднородности вычисляется для произвольной выборки $\tilde{S}$ , содержащей объекты из классов $K_1,\ldots,K_L$.

При этом используется несколько видов индекса:

а) Энтропийный индекс неоднородности вычисляется по формуле $\gamma_e(\tilde{S}) = -\sum_{i=1}^{L} P_i \ln(P_i)$, где $P_i$ - доля объектов класса $K_i$ в выборке $\tilde{S}$ . При этом принимается, что $0\ln(0) = 0$.

Наибольшее значение $\gamma_e(\tilde{S})$ принимает при равенстве долей классов. Наименьшее значение $\gamma_e(\tilde{S}) = 0$ достигается при принадлежности всех объектов одному классу.

# Решающие деревья

б) Индекс Джини вычисляется по формуле

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

В) Индекс ошибочной классификации

$$\gamma_m(\tilde{S}) = 1 - \max_{i=1,\ldots,L}(P_i)$$

Нетрудно понять, что индексы б) и в) также достигают минимального значения при принадлежности всех объектов обучающей выборке одному классу.

# Решающие деревья

Уменьшение среднего индекса неоднородности в выборках $\tilde{S}_t^l$ и $\tilde{S}_t^r$ по отношению к обучающей выборке $\tilde{S}_t$ вычисляется по формуле

$$\Delta(\gamma, \tilde{S}_t) = \gamma(\tilde{S}_t) - P_l\gamma(\tilde{S}_t^l) - P_r\gamma(\tilde{S}_t^r)$$

где $P_l$ и $P_r$ - доли $\tilde{S}_t^l$ и $\tilde{S}_t^r$ в выборке $\tilde{S}_t$ .

# Решающие деревья

Выбирается индекс неоднородности $\gamma$. Для каждого из признаков ищется оптимальный предикат. Выбирается признак $X_{i_{\max}}$ с максимальным значением индекса $\Delta(\gamma)$. Подвыбороки $\tilde{S}_t^l$ и $\tilde{S}_t^r$, 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.

# Решающие деревья

В случае, если какая-нибудь из выборок $\tilde{S}_t^*$ удовлетворяет критерию остановки, то соответствующая вершина дерева объявляется концевой и для неё вычисляется метка класса. В случае, если выборка $\tilde{S}_t^*$ не удовлетворяет критерию остановки, то формируется новая внутренняя вершина, для которой процесс построения дерева продолжается.

Пусть некоторой вновь образованной внутренней вершине соответствует выборка $\tilde{S}_v$. Для данной выборки производятся те же самые построения, которые на начальном этапе проводились для обучающей выборки $\tilde{S}_t$.

# Решающие деревья

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.

# Решающие деревья

Одним из способов достижения более высокой обобщающей способности является использования критериев остановки, позволяющих остановит процесс построения дерева до того, как будет достигнута полная однородность концевых вершин. Рассмотри несколько таких критериев.

1. Критерий остановки по минимальному допустимому числу объектов в выборках, соответствующих концевым вершинам.

# Решающие деревья

2. Критерий остановки по минимально допустимой величине индекса $\Delta(\gamma, \tilde{S})$. Предположим, что некоторой вершине $v$ соответствует выборка $\tilde{S}_v$, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\tilde{S}_v^l, \tilde{S}_v^r\}$. Вершина $v$ считается внутренней, если индекс $\Delta(\gamma, \tilde{S}_t)$ превысил пороговое значение $\tau$ и считается концевой в противном случае.

# Решающие деревья

3. Критерий остановки по точности на контрольной выборке.

Исходная выборка данных случайным образом разбивается на обучающую выборку $\tilde{S}_t$ и контрольную выборку $\tilde{S}_c$. Выборка $\tilde{S}_t$ используется для построения бинарного решающего дерева.

Предположим, что некоторой вершине $v$ соответствует выборка $\tilde{S}_v$, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\tilde{S}_v^l, \tilde{S}_v^r\}$.

На контрольной выборке $\tilde{S}_c$ производится сравнение эффективность распознающей способности деревьев $\mathbf{T}_v$ и $\mathbf{T}_v^{++}$.

# Решающие деревья

Деревья $\mathbf{T}_v$ и $\mathbf{T}_v^{++}$ включает все вершины и рёбра, построенные до построения вершины $v$ . В дереве $\mathbf{T}_v$ вершина $v$ считается концевой. В дереве $\mathbf{T}_v^{++}$ вершина $v$ считается внутренней, а концевыми считаются вершины, соответствующие подвыборкам $\tilde{S}_v^l \, u \, \tilde{S}_v^r$ . Распознающая способность деревьев $\mathbf{T}_v$ и $\mathbf{T}_v^{++}$ сравнивается на контрольной выборке $\tilde{S}_c$ . В том, случае если распознающая способность $\mathbf{T}_v$ превосходит распознающую способность $\mathbf{T}_v^{++}$ все дальнейшие построения исходят из того, что вершина $v$ является концевой. В противном случае производится исследование $\tilde{S}_v^l \, u \, \tilde{S}_v^r$.

# Решающие деревья

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 $v$, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\tilde{S}_v^l, \tilde{S}_v^r\}$.

.Исследуется статистическая достоверность различий между содержанием объектов распознаваемых классов в подвыборках $\tilde{S}_v^l$ и $\tilde{S}_v^r$.

# Решающие деревья

Для этих целей может быть использованы известные статистический критерий: Хи-квадрат и другие критерии. По выборкам $\tilde{S}_v^l$ и $\tilde{S}_v^r$ рассчитывается статистика критерия, по которой с использованием табулированных распределений устанавливается соответствующее p-значение. В том случае, если полученное p-значение оказывается меньше заранее фиксированного уровня значимости вершина $v$ считается внутренней. В противном случае вершина $v$ считается концевой.

# Решающие деревья

Использование критериев ранней остановки не всегда позволяет адекватно оценить. необходимую глубину дерева. Слишком ранняя  остановка ветвления может привести к потере информативных предикатов. В связи с эти  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»).

# Решающие деревья

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 $v$ одновременно увеличения точности разделения классов на обучающей выборке и увеличения сложности, которые возникают благодаря ветвлению из $v$.

# Решающие деревья

При этом прирост сложности может быть оценён через число листьев в поддереве $\mathbf{T}_v$ решающего дерева с корневой вершиной $v$ Следует отметить, что рост сложности является штрафующим фактором, компенсирующим прирост точности разделения на обучающей выборке с помощью включения поддерева $\mathbf{T}_v$ в решающее дерево. Разработан целый ряд эвристических критериев, зависящих от точности разделения с включением и без включением поддерева $\mathbf{T}_v$ и сложности $\mathbf{T}_v$, которые позволяют оценить целесообразность включения $\mathbf{T}_v$.