Обзор методов классификации

Курс «Интеллектуальные информационные системы»

Кафедра управления и информатики НИУ «МЭИ»

Осень 2018 г.

Наивный байесовский метод (НБ)

теорема Байеса: $$P(Q_k \mid \vec{X}) = \frac{P(\vec{X} \mid Q_k)P(Q_k)}{P(\vec{X})}$$

позволяет определить вероятность какого-либо события при условии, что произошло другое статистически взаимозависимое с ним событие.

- $P(\vec{X})$ - одинакова для различных классов и может быть исключена из дальнейшего рассмотрения
- Допущение: признаки (свойства, термины,…), которыми описывается объект, независимы между собой.
- Данное допущение значительно упрощает задачу, но крайне редко выполняется на практике.

$$P(\vec{X} \mid Q_k) = \prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

$$P(Q_k \mid \vec{X}) = P(Q_k) \prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

Наивный байесовский метод (2)

$$P(Q_k \mid \vec{X}) = P(Q_k)\prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

- $\widehat{P}(Q_k) = \frac{N_k}{N}$ - оценка для P(Qk) – вероятность встретить объект класса Qk в выборке

- $\widehat{P}(x^{(i)} \mid Q_k) = \frac{N_{ik}}{N_k}$ - вероятность встретить признак x(i) в классе Qk

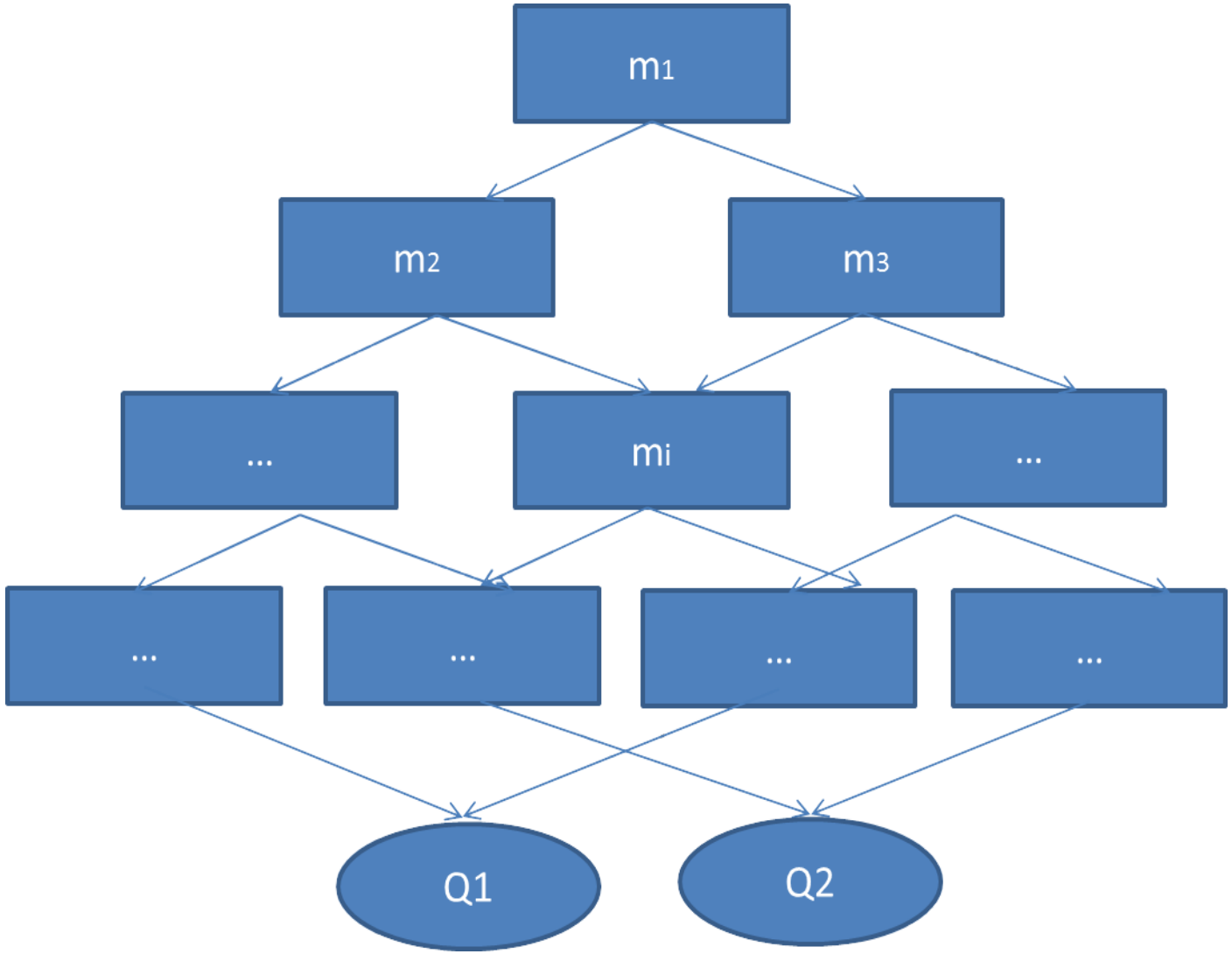
- Часто (особенно для задач Text mining) используется уточненная формула: $\widehat{P}(x^{(i)} \mid Q_k) = \frac{1 + N_{ik}}{M + N_k}$
  где M – общее количество признаков (терминов) во всех объектах (документах) выборки

$$P(Q_k \mid \vec{X}) = \arg\max_{k \in K} \frac{N_k}{N} \prod_{i=1}^{M} \frac{1 + N_{ik}}{M + N_k}$$

Метод деревьев решений

Средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей.

В методе деревьев решений проводится последовательное разделение множества объектов на основе значений выбранного признака, в результате чего строится дерево, содержащее нетерминальные узлы (узлы проверок), в которых происходит разбиение по выбранному атрибуту, и терминальные узлы (узлы ответа), в которых должны находиться элементы одного класса.



Метод деревьев решений. Критерий прироста информации

Для выбора наиболее информативного признака, по которому проводится разбиение, в методе деревьев решений чаще всего используется *теоретико-информационный (энтропийный) подход*.

Хотим найти такой признак $x^{(S)}$, при разбиении по которому один из классов имел наибольшую вероятность появления. Это возможно, если величина прироста информации *Gain* будет достигать своего максимума.

$$Gain(x^{(s)},T) = I(T) - I(x^{(s)},T)$$

$$I(T) = \sum_{k=1}^{K} P_k \log_2 \frac{1}{P_k} = -\sum_{k=1}^{K} \frac{N_k}{N} \log_2 \frac{N_k}{N}$$

- среднее количество информации (энтропия), необходимое для определения класса примера из обучающей выборки *T*

$$I(x^{(s)},T) = \sum_{i=1}^{S} \frac{N_s}{N} I(T_s) = \sum_{s=1}^{S} \frac{N_s}{N} \left(-\sum_{k=1}^{K} \frac{N_{ks}}{N_s} \log_2 \frac{N_{ks}}{N_s} \right)$$

- среднее количество информации, необходимое для идентификации класса примера в каждом подмножестве после разбиения по признаку $x^{(S)}$

Метод деревьев решений. Меры неоднородности

Еще один подход к выявлению признака, по которому стоит проводить разбиение – использовать меры неоднородности ф. Здесь вектор **p** состоит из **m** вероятностей меток встречающихся в некотором подмножестве обучающего множества

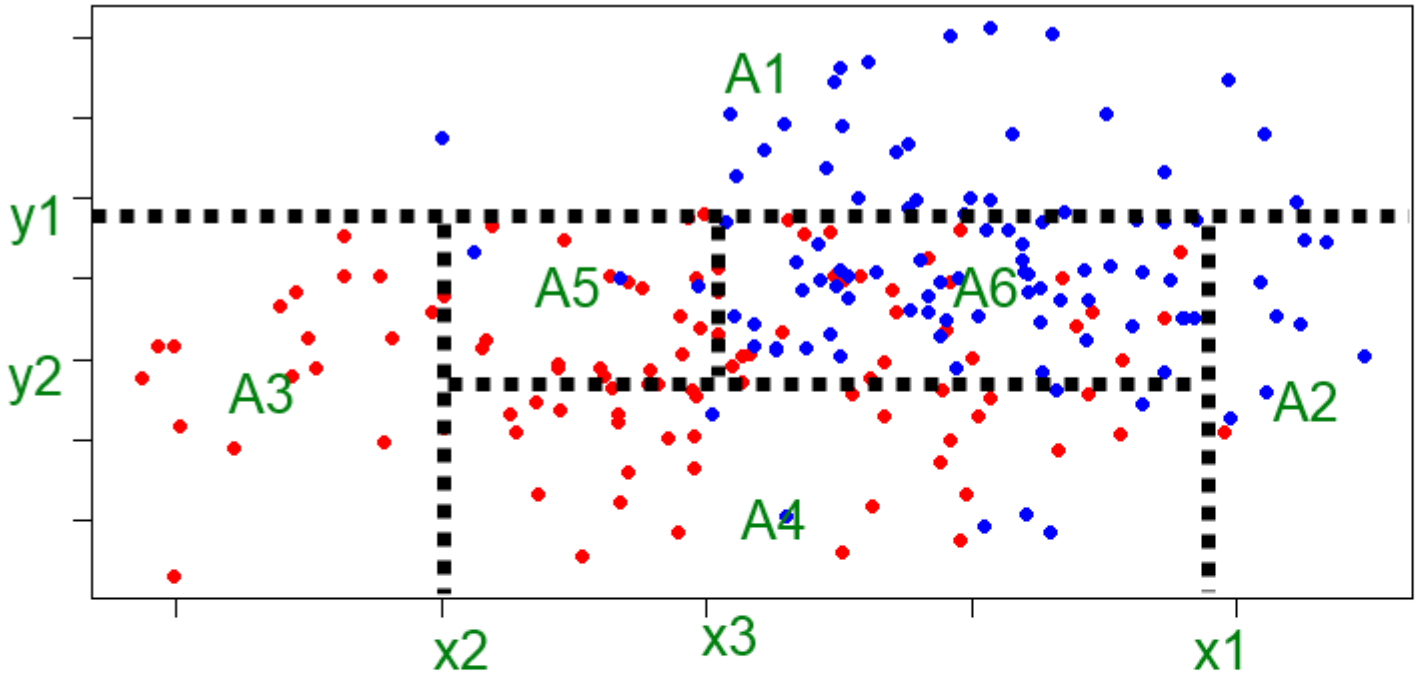
$$\phi(\vec{p}) = 1 - \max(\vec{p})$$ Наиболее часто встречаемый класс

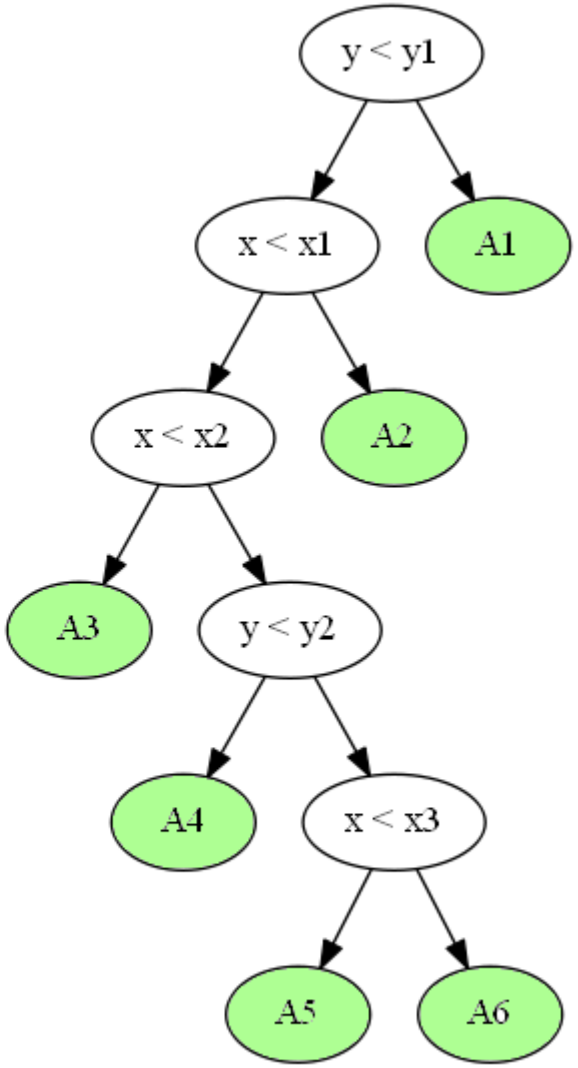
$$\phi(\vec{p}) = \sum_{i=1}^{m} p_i(1 - p_i)$$ Индекс (коэффициент) Джини (Gini index)

$$\phi(\vec{p}) = -\sum_{i=1}^{m} p_i \log(p_i)$$ Перекрестная энтропия

На каждой итерации для входного подмножества обучающего множества строится такое разбиение пространства гиперплоскостью (ортогональной одной их осей координат), которое минимизировало бы среднюю меру неоднородности двух полученных подмножеств. Данная процедура выполняется рекурсивно для каждого полученного подмножества до тех пор, пока не будут достигнуты критерии остановки.

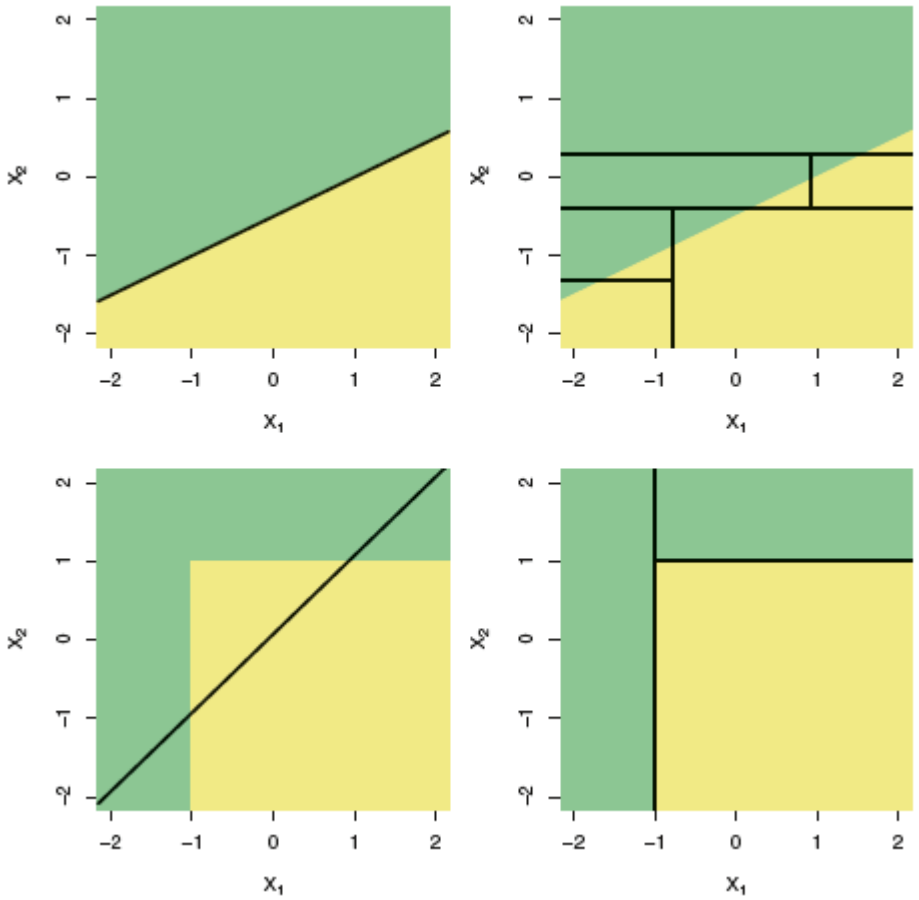
Метод деревьев решений. Пример разбиения





Метод деревьев решений. Область использования

Сравнение линейных алгоритмов и алгоритмов, основанных на деревьях решений:



Недостатки:

- Алгоритмы «локальны», не могут обеспечить оптимальность всего дерева в целом
- Свойственна проблема «переобученности»

Достоинства

- Прост в понимании и интерпретации
- Не требует предварительной обработки данных
- Метод хорошо работает даже в том случае, если были нарушены первоначальные предположения, включенные в модель.

Случайный лес (Random Forest)

Пусть обучающая выборка состоит из N примеров, размерность пространства признаков равна M, и задан параметр m

Все деревья комитета строятся независимо друг от друга по следующей процедуре:

1. Сгенерируем случайную подвыборку **с повторением** размером *N* из обучающей выборки. (Таким образом, некоторые примеры попадут в неё несколько раз, а в среднем $N(1 - 1/N)^N$, т.е. примерно *N/e* примеров не войдут в неё вообще)
2. Построим решающее дерево, классифицирующее примеры данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех M признаков, а лишь из m случайно выбранных.
3. Проводится построение дерева

Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.