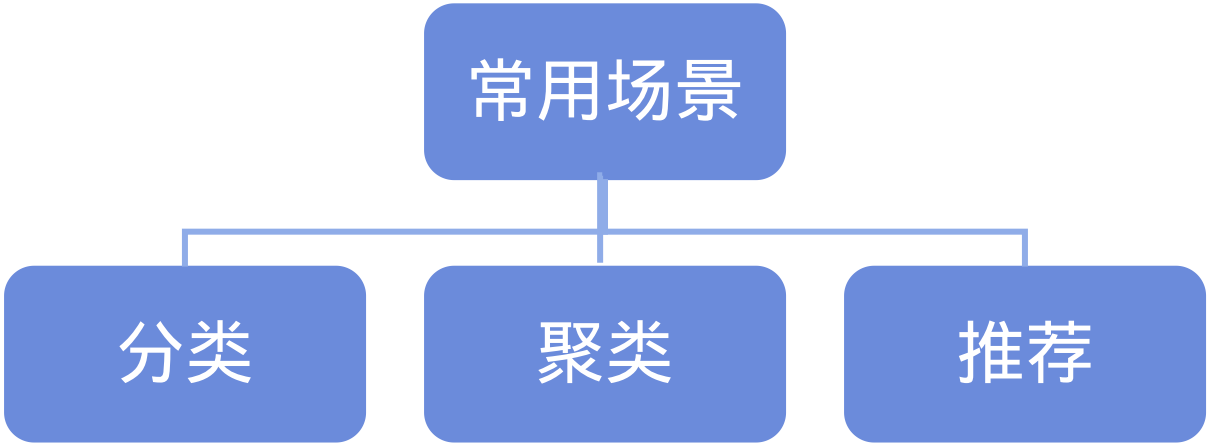
数据挖掘常用方法浅析

徐进

主要内容

背景知识

数据积累与技术进步，挖掘数据中的“宝藏”成为可能



去哪儿? Qunar.Com 聪明你的旅行

数据挖掘概念

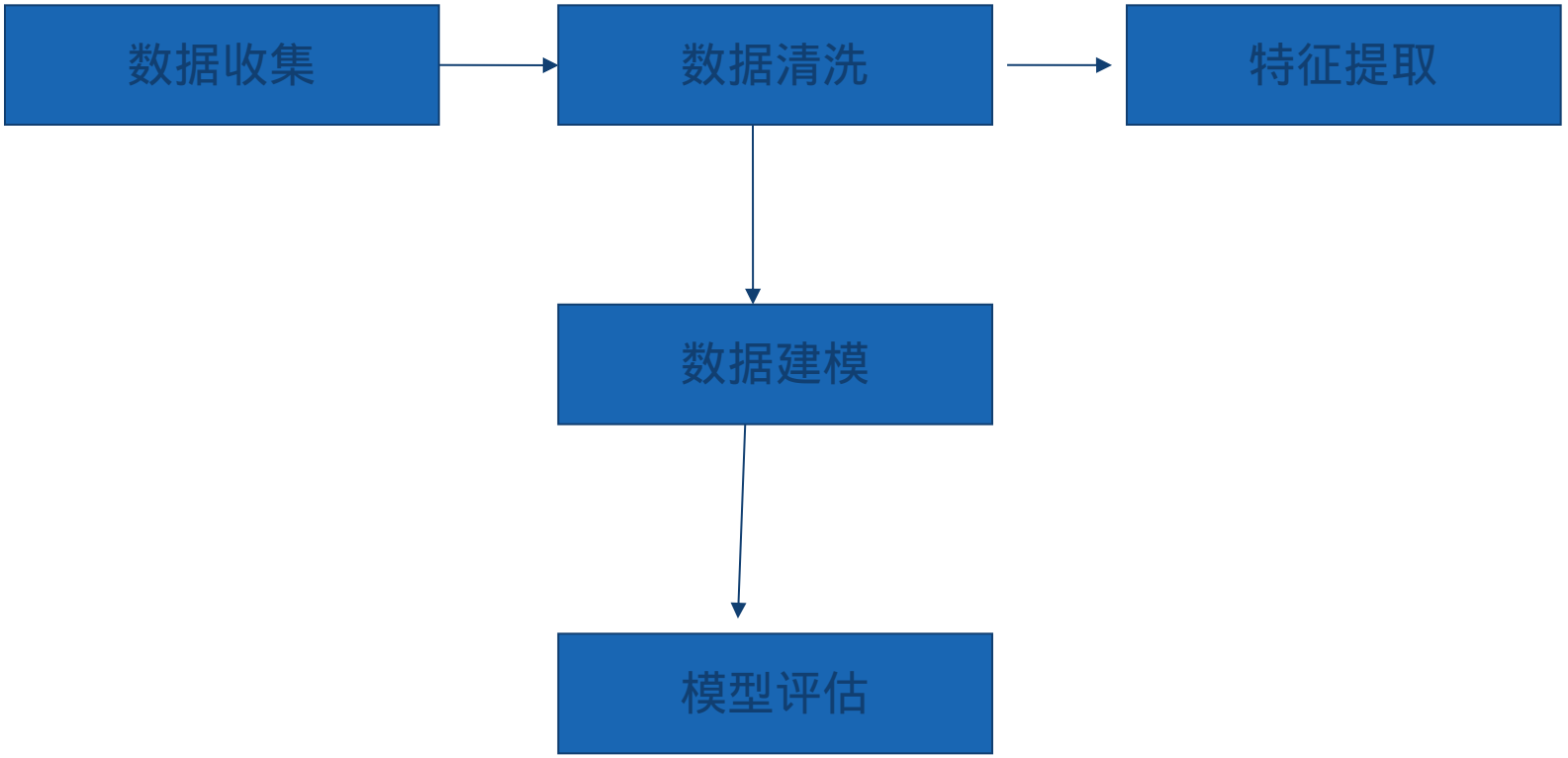
数据挖掘(Data Mining)，是从大量数据中挖掘或抽取出知识

举例：

疾病预测、垃圾邮件识别、商品推荐等等

去哪儿? Qunar.Com 聪明你的旅行

处理流程



去哪儿? Qunar.Com 聪明你的旅行

距离度量

欧几里得距离

$$dist(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

曼哈顿距离

$$dist(X,Y) = \sum_{i=1}^{n}|x_i - y_i|$$

切比雪夫距离

$$dist(X,Y) = \lim_{p\to\infty}(\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} = max\,|x_i - y_i|$$

明可夫斯基距离

$$dist(X,Y) = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p}$$

欧式空间

去哪儿? Qunar.Com 聪明你的旅行

相似度

余弦相似度

$$sim(X,Y) = cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|x\| \cdot \|y\|}$$

杰拉德相似系数

$$Jaccard(X,Y) = \frac{X \cap Y}{X \cup Y}$$

皮尔逊相关系数

$$r = \frac{\sum_{i=1}^{n}(X_i - X)(Y_i - Y)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

皮尔逊相关系数解释：协方差和标准差的商，范围：[-1,1]

去哪儿? Qunar.Com 聪明你的旅行

相似度

博客园相似度计算

昵称	关注数量	共同数量	相似度
蓝枫叶1938	5	4	0.373001923296126
FBI080703	3	3	0.361157559257308
鱼非鱼	3	3	0.361157559257308
Lauce	3	3	0.361157559257308
蓝色蜗牛	3	3	0.361157559257308
shanyujin	3	3	0.361157559257308
Mr.Huang	6	4	0.340502612303499
对世界说你好	6	4	0.340502612303499
strucoder	28	8	0.31524416249564
Mr.Vangogh	4	3	0.312771621085612

去哪儿? Qunar.Com 聪明你的旅行

分类场景

分类指有监督的学习，有明确的类标签

垃圾邮件识别

疾病判断

天气预报

等等

去哪儿? Qunar.Com 聪明你的旅行

分类

常用手段：

- 回归
- 决策树
- 贝叶斯
- 人工神经网络
- 支持向量机

去哪儿? Qunar.Com 聪明你的旅行

回归

1、概念：

基于数学方法在数据集上建立自变量（特征属性）与因变量（分类属性）之间的拟合函数表达式

2、步骤

2.1、寻找拟合函数

2.2、计算参数

2.3、利用拟合函数预测

去哪儿? Qunar.Com 聪明你的旅行

最小二乘法

通过最小化误差的平方和寻找数据的最佳函数匹配

拟合函数：

$$y = a_0 + a_1 x + \ldots + a_k x^k,$$

损失函数：

$$R^2 \equiv \sum_{i=1}^{n} \left[y_i - \left(a_0 + a_1 x_i + \ldots + a_k x_i^k \right) \right]^2.$$

矩阵求根：

$$\begin{bmatrix} n & \sum_{i=1}^n x_i & \cdots & \sum_{i=1}^n x_i^k \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \cdots & \sum_{i=1}^n x_i^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_i^k & \sum_{i=1}^n x_i^{k+1} & \cdots & \sum_{i=1}^n x_i^{2k} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \\ \vdots \\ \sum_{i=1}^n x_i^k y_i \end{bmatrix}.$$

随机梯度下降(SGD)法

沿着倒数的方向，才能最快的逼近极值点

拟合函数：

$$G(x) = \Theta_0 + \Theta_1 x_1 + \cdots + \Theta_n x_n$$

损失函数：

$$\Phi(x) = \frac{1}{2m}\sum_{1}^{m}(G(x) - y)^2$$

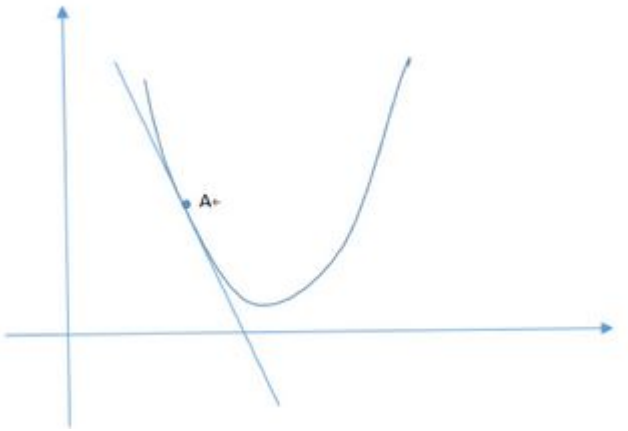
迭代逼近：

$$\theta_0 = \theta_0 - \alpha \frac{\partial \Phi(x)}{\partial \theta_0}$$

$$\theta_1 = \theta_1 - \alpha \frac{\partial \Phi(x)}{\partial \theta_0}$$

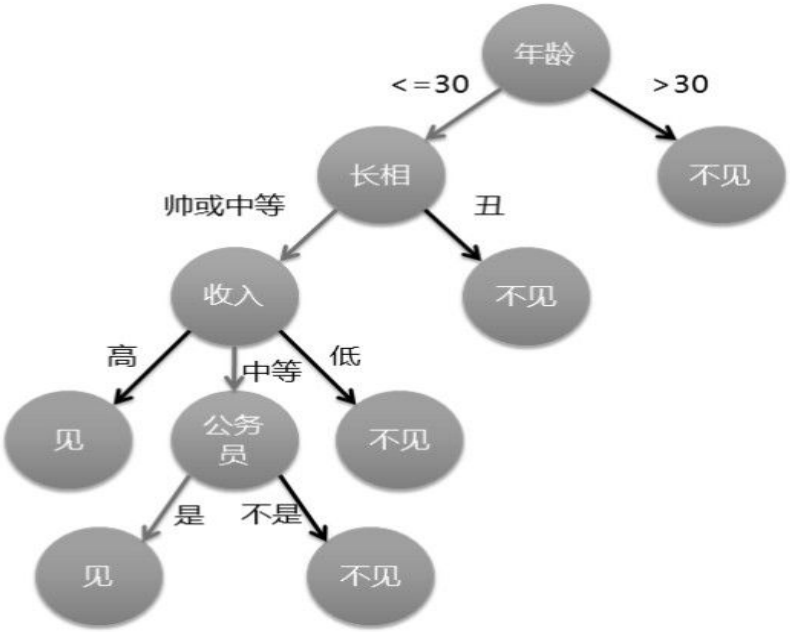
$$\vdots$$

$$\theta_n = \theta_n - \alpha \frac{\partial \Phi(x)}{\partial \theta_n}$$



决策树

决策树树结构:非叶节点表示一个特征属性，其不同取值代表不同分支，叶节点代表分类属性



去哪儿? Qunar.Com 聪明你的旅行

决策树

熵值指的是混乱程度，熵越小越纯

1.数据集D的熵：

$$info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

2.D按特征属性A进行划分后的熵：

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j)$$

3.信息增益：

$$gain(A) = info(D) - info_A(D)$$

ID3算法就是在每次需要分裂时，选择熵增益最大的属性进行分裂。

决策树

D3算法的固有问题偏向于多值属性

C4.5选择具有最大增益率的属性作为分裂属性

分裂信息：$split\_info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} log_2(\frac{|D_j|}{|D|})$

增益率：$gain\_ratio(A) = \frac{gain(A)}{split\_info(A)}$

C4.5选择具有最大增益率的属性作为分裂属性

去哪儿? Qunar.Com 聪明你的旅行

贝叶斯

$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$$

朴素贝叶斯分类:　特征属性条件独立

条件独立:　　P(AB|C)=P(A|C)P(B|C)

贝叶斯定理　　$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$

去哪儿? Qunar.Com 聪明你的旅行

朴素贝叶斯

待分类的特征属性 $x = \{a_1, a_2, ..., a_m\}$;

类别属性 $C = \{y_1, y_2, ..., y_n\}$

1、在有分类属性的数据集上，计算：

$$P(a_1|y_1), P(a_2|y_1), ..., P(a_m|y_1); P(a_1|y_2), P(a_2|y_2), ..., P(a_m|y_2); ...; P(a_1|y_n), P(a_2|y_n), ..., P(a_m|y_n)。$$

2、在给定特征属性数据上，有贝叶斯定理：

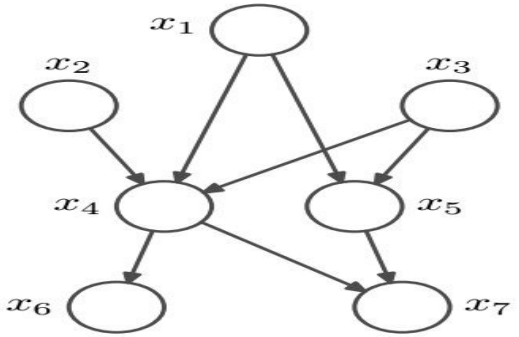
$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

3、已知P(x)是常数，加之特征属性间条件独立，只需求

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i)$$

贝叶斯网络

需要专家预先建立网络的拓扑结构（理论上自适应调整建立贝叶斯网络）



P(x7)概率：

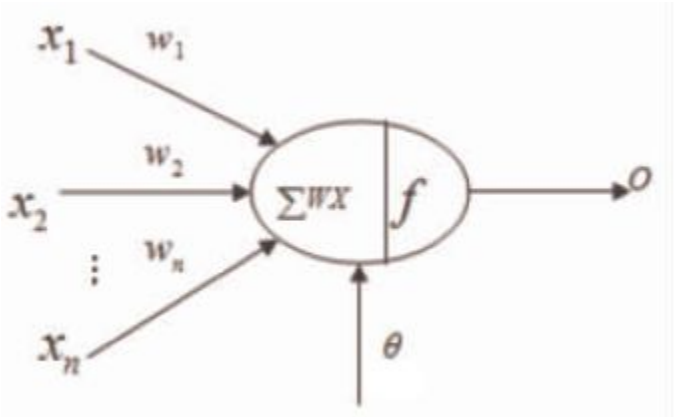
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

一般形式：

$$P(x_1,x_2,...,x_n)=\prod_{i=1}^{n}P(x_i|Parents(x_i))$$

人工神经网络

基于生物学中神经网络的基本原理，借助人脑结构和外界刺激响应机制后，以网络拓扑知识为理论基础，模拟人脑的神经系统对复杂信息的反馈处理机制的一种数学模型



去哪儿? Qunar.Com 聪明你的旅行

人工神经网络

1.建立模型

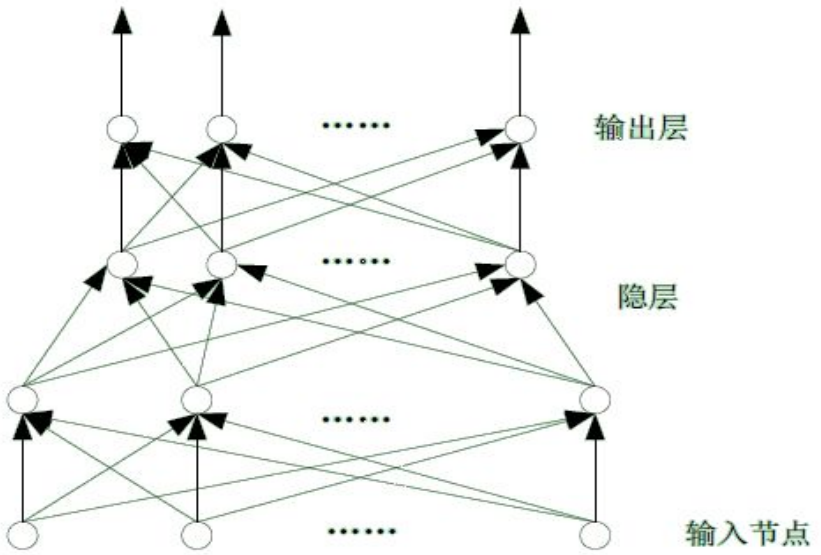
根据数据集，选择网络模型、训练函数

2.调整权值

根据实际输出和期望输出之间的误差进行权值的修正

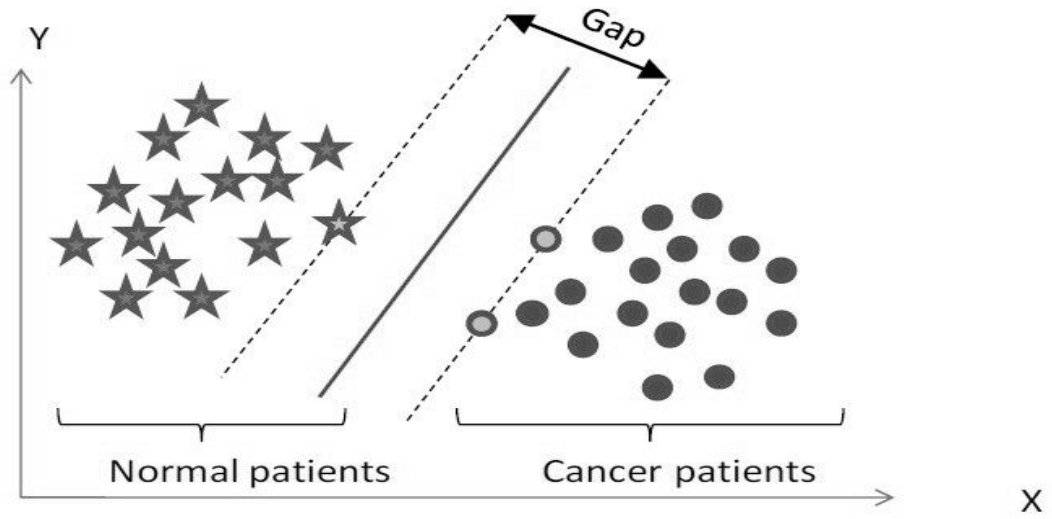
人工神经网络

多神经元、多层次



支持向量机

二类分类模型，特征空间上的间隔最大化



去哪儿? Qunar.Com 聪明你的旅行

支持向量机

1.超平面 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0$

2.距离 $r=\dfrac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b|}{||\boldsymbol{w}||}$

3.分类数据集
$$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\geqslant+1, & y_i=+1\ ; \\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b\leqslant-1, & y_i=-1\ . \end{cases}$$

距离分类超平面最近的点且平行于超平面的训练样本，称为支持向量

$$\gamma=\frac{2}{||\boldsymbol{w}||}$$

支持向量机

5.满足条件

$$\max_{\boldsymbol{w},b} \frac{2}{||\boldsymbol{w}||}$$
$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, m.$$

6.拉格朗日乘子，有 $L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)\right)$ 其中 $\alpha_i \geqslant 0$

7.对w和b求偏导

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$
$$0 = \sum_{i=1}^{m} \alpha_i y_i \ .$$

8.KKT条件

$$\begin{cases} \alpha_i \geqslant 0 \ ; \\ y_i f(\boldsymbol{x}_i) - 1 \geqslant 0 \ ; \\ \alpha_i \left(y_i f(\boldsymbol{x}_i) - 1\right) = 0 \end{cases}$$

去哪儿? Qunar.Com 聪明你的旅行

模型选择

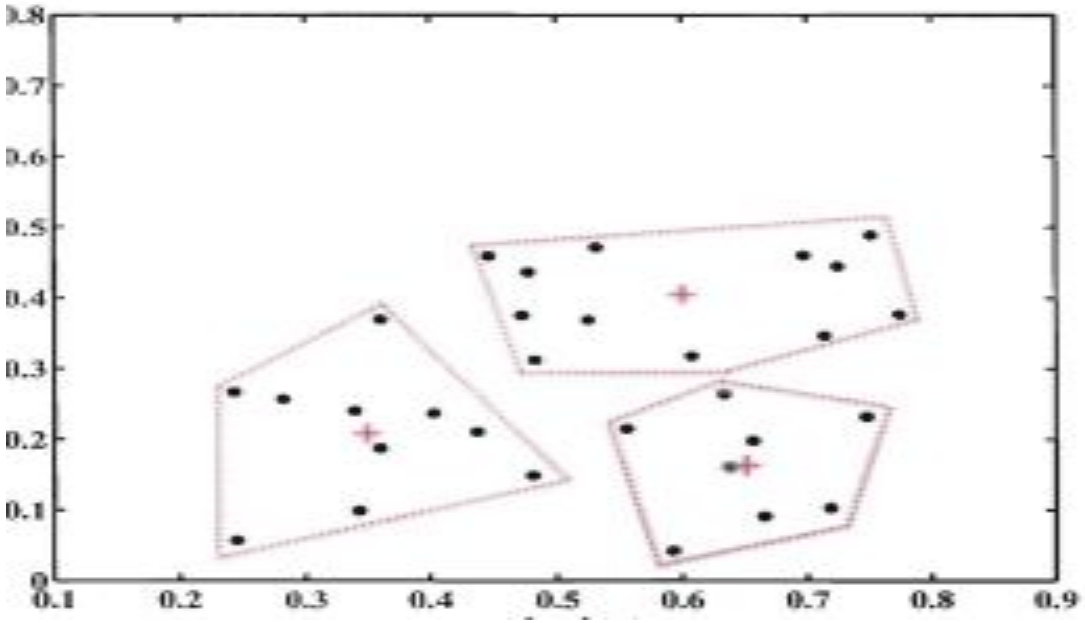
方法	原理实现	计算量	增量训练	可解释性	条件独立	适用场景
回归	简单	小	友好	较强	不要求	受限拟合方程
决策树	简单	大	不友好	较强	不要求	离散数据、属性少
朴素贝叶斯	简单	小	友好	较强	要求	无法感知属性相互作用
K近邻	简单	大	不友好	不强	不要求	受限相似度
支持向量机	复杂	小	不友好	不强	不要求	小样本、高维
人工神经网络	复杂	大	友好	不强	不要求	通用复杂场景

去哪儿? Qunar.Com 聪明你的旅行

聚类

聚类指无监督的学习，没有明确的类标签

“物以类聚，人以群分”



文本聚合（垃圾邮件、新闻聚合）

去哪儿? Qunar.Com 聪明你的旅行

聚类

- K近邻
  - k-means
  - k-modles
- EM
- 层次聚类
  - 归约型
  - 分裂型

去哪儿? Qunar.Com 聪明你的旅行

K近邻

1、随机地选择k个对象，每个对象初始地代表了一个簇的平均值或中心

2、剩余的每个对象，根据其与各簇中心的距离，将它赋给最近的簇

3、重新计算每个簇的平均值，这个过程不断重复，直到准则函数收敛

k-means：以平均值为中心

k-modles：以最靠近平均值的点为中心

EM

1.初始化分布参数

2.聚类

2.1.E步骤：用分布参数计算每个实例的聚类概率

2.2.M步骤：重新估计分布参数(如不同聚类簇的方差、期望等)，使数据的似然性尽可能大

迭代计算2.1，2.2，直至收敛

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

EM

1.似然函数

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

2 .取对数

$$\log L(\mu,\ \sigma^2)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

3.求导，极大值

$$\begin{cases}\dfrac{\partial\log L(\mu,\ \sigma^2)}{\partial\mu}=\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0,\\ \dfrac{\partial\log L(\mu,\ \sigma^2)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0,\end{cases}$$

4.极大似然估计

$$\mu^*=\overline{X}\ ,\quad \sigma^{*2}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\ .$$

层次聚类

归约型与分裂型

1.归约型

1.1、 将每个对象看作一类，计算两两之间的最小距离；

1.2、将距离最小的两个类合并成一个新类；

1.3、 重新计算新类与所有类之间的距离；

1.4、 重复1.2、1.3，直到所有类最后合并成N类。

2.分裂型略

去哪儿? Qunar.Com 聪明你的旅行

推荐

协同过滤

- 基于user
- 基于item

关联规则

- Apriori
- FP_growth

去哪儿? Qunar.Com 聪明你的旅行

协同过滤场景

协同过滤指利用集体智慧的进行过滤

社交网络推荐

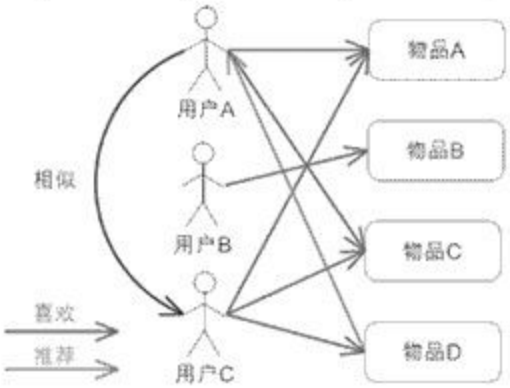
商品推荐

去哪儿? Qunar.Com 聪明你的旅行

基于user协同过滤

找有相似度的用户进行推荐，适合user少、item多的场景

用户/物品	物品A	物品B	物品C	物品D
用户A	√		√	推荐
用户B		√		
用户C	√		√	√

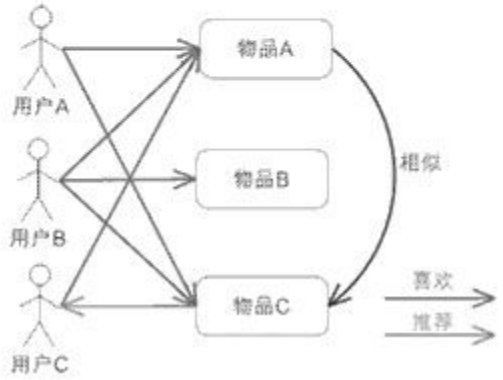


去哪儿? Qunar.Com 聪明你的旅行

基于item协同过滤

找有相似度的item进行推荐，适合item少、user多的场景

用户/物品	物品A	物品B	物品C
用户A	√		√
用户B	√	√	√
用户C	√		推荐



去哪儿? Qunar.Com 聪明你的旅行

关联规则场景

寻找数据项之前频繁出现的组合

“啤酒尿布问题”

TID	Items
T1	{牛奶,面包}
T2	{面包,尿布,啤酒,鸡蛋}
T3	{牛奶,尿布,啤酒,可乐}
T4	{面包,牛奶,尿布,啤酒}
T5	{面包,牛奶,尿布,可乐}

1.support({啤酒}–>{尿布})：
啤酒和尿布同时出现的次数/数据记录数 = 3/5=60%

2.confidence({尿布}–>{啤酒})：
啤酒和尿布同时出现的次数/尿布出现的次数 = 3/4 = 75%

去哪儿? Qunar.Com 聪明你的旅行

关联规则

1、寻找数据中事物之间可能存在的关联或者联系

2、两个主要阶段：

2.1、根据给定的最小支持度supmin，从事务集中找出全部的频繁项集；

2.2、根据最小可信度supcon，由已知频繁项集中生成感兴趣的关联规则。

去哪儿? Qunar.Com 聪明你的旅行

Apriori

1、扫描事务集T求出每个1项集的支持度，即得到频繁1项集的集合；

2、循环计算频繁k项集。

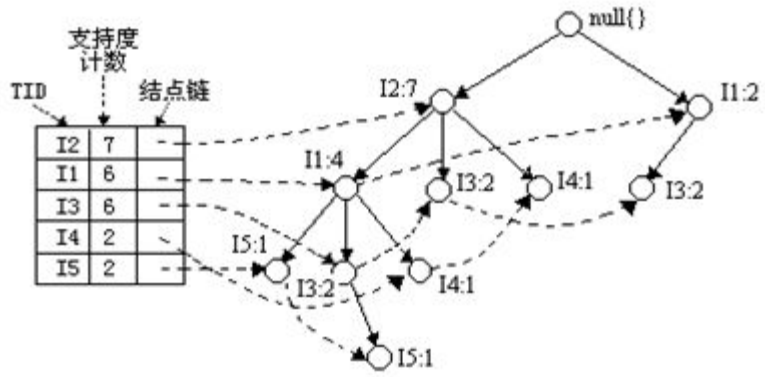
2.1、连接：由两个有且只有一个项不同的k-1频繁项集连接得出k频繁项集的候选集；

2.2、剪枝：上述得出的是k频繁项集的候选集，需要对候选集k中的k-1项子集进行判断。若k-1子集不是频繁项集，则直接剔除掉；

3、扫描计算所得的频繁项集，依据给出的置信度等筛选条件确定感兴趣的关联规则。

FP_growth

在不生成候选项的情况下,寻找频繁项集



一颗growth树，每个分支代表一个数据项，内容该分支节点内容集合

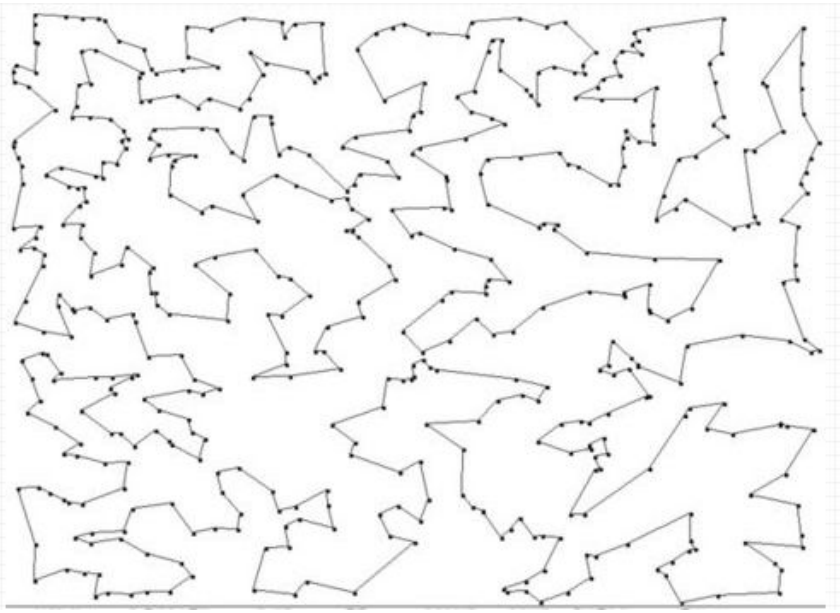
自底向上遍历树的分支，可以得到频繁项集

去哪儿? Qunar.Com 聪明你的旅行

遗传算法场景

模拟自然选择的启发式算法

1、旅行商问题



2、复杂方程求最值

去哪儿? Qunar.Com 聪明你的旅行

遗传算法

1 .种群

生物的进化以群体的形式进行

2.个体

组成种群的单个生物

3.基因

一个遗传因子

4.染色体

包含一组的基因

5 .交叉

新个体会遗传父母双方部分基因

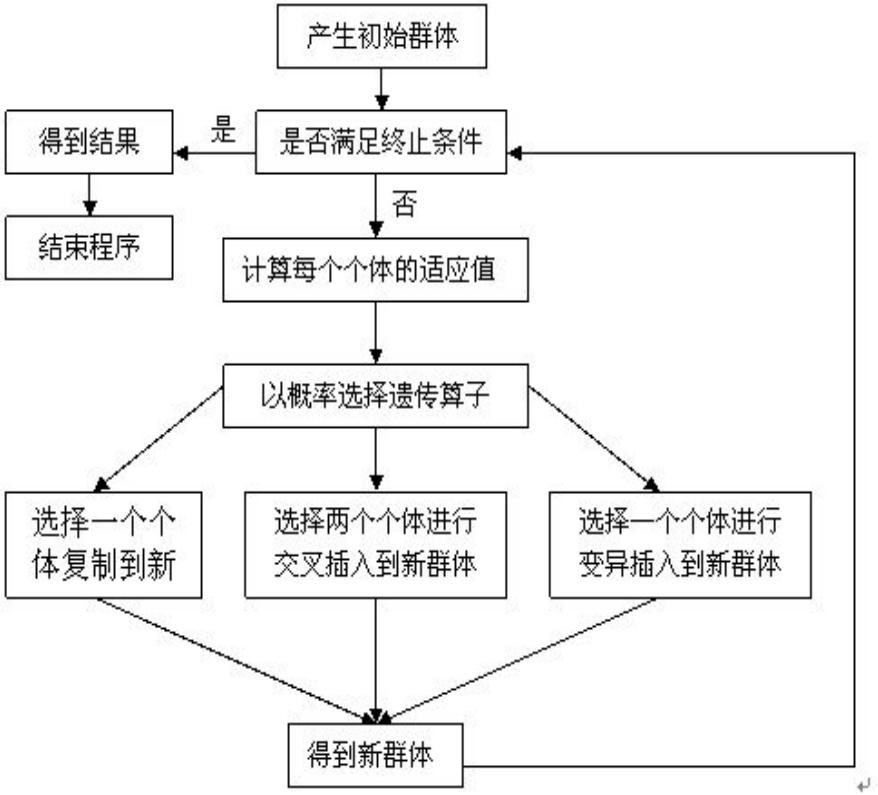
6.变异

基因小概率发生突变

去哪儿? Qunar.Com 聪明你的旅行

遗传算法

在繁殖过程中，会发生染色体交叉、基因突变，适应度低的个体会被逐步淘汰，而适应度高的个体会越来越多；经过N代的自然选择后，种群整体适应度会提高。



去哪儿? Qunar.Com 聪明你的旅行

开源框架

开源框架	机构	语言	适用范围
Mahout	Apache	Java	数据挖掘
Weka	社区	Java	数据挖掘
MLlib	Apache	Scala	数据挖掘
Tensorflow	Google	C++/Python	人工神经网络
Caffe	社区	C++/CUDA	人工神经网络
Paddle	百度	C++/Python	人工神经网络

去哪儿?
Qunar.Com
聪 明 你 的 旅 行

致谢

Thanks