分类算法

2020年3月6日 10:00

ID3

思想：信息熵，贪心

terative Dichotomiser 3

即 迭代二叉树 3 代

ID3算法最早是由罗斯昆（J. Ross Quinlan）于1975年在悉尼大学提出的一种分类预测算法，算法的核心是“信息熵”。ID3算法通过计算每个属性的信息增益，认为信息增益高的是好属性，每次划分选取信息增益最高的属性为划分标准，重复这个过程，直至生成一个能完美分类训练样例的决策树。

（1）熵

在信息论中，熵（entropy）是随机变量不确定性的度量，也就是熵越大，则随机变量的不确定性越大。设X是一个取有限个值得离散随机变量，其概率分布为：

$$P(X = x_i) = p_i, \qquad i = 1,2, \dots, n$$

则随机变量X的熵定义为：

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

（2）条件熵

设有随机变量（X, Y），其联合概率分布为：

$$P(X = x_i, Y = y_i) = p_{ij}, \qquad i = 1,2, \dots, n, j = 1,2, \dots, m$$

条件熵H(Y|X)表示在已知随机变量X的条件下，随机变量Y的不确定性。随机变量X给定的条件下随机变量Y的条件熵H(Y|X)，定义为X给定条件下Y的条件概率分布的熵对X的数学期望：

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i)$$

其中$p_i = P(X = x_i)$，$i = 1,2,\dots,n$

当熵和条件熵中的概率由数据估计得到时（如极大似然估计），所对应的熵与条件熵分别称为经验熵和经验条件熵。

（3）信息增益

定义：信息增益表示由于得知特征A的信息后儿时的数据集D的分类不确定性减少的程度，定义为：

$$Gain(D,A) = H(D) - H(D|A)$$

即集合D的经验熵H(D)与特征A给定条件下D的经验条件熵H(H|A)之差。

理解：选择划分后信息增益大的作为划分特征，说明使用该特征后划分得到的子集纯度越高，即不确定性越小。因此我们总是选择当前使得信息增益最大的特征来划分数据集。

缺点：

1.倾向于选择取值比较多的属性，分类过多会影响信息熵，对结果造成影响，可采用聚类、聚集或上钻较少其取值。

2.没有考虑连续特征，如成绩，工资，会把不同数值的属性分别分类

3.D3算法没有考虑“过拟合”overfitting的问题，对于样本内的数据表现很好，对于样本外数据不能很好的拟合

C4.5

针对ID3缺点1，引入信息增益率概念，将选择特征的方法由信息增益改成信息增益比。

CART

改用Gini指数为标准分类

Gini指数

分类问题中，假设有K个类，样本点属于第k类的概率为 p_k ，则概率分布的基尼指数定义为：

$$\text{Gini(p)} = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} {p_k}^2$$

备注： p_k 表示选中的样本属于k类别的概率，则这个样本被分错的概率为 $(1\text{-}p_k)$ 。

对于给定的样本集合D，其基尼指数为：

$$\text{Gini(D)} = 1 - \sum_{k=1}^{K} (\frac{|C_k|}{|D|})^2$$

备注：这里 c_k 是D中属于第k类的样本自己，K是类的个数。

如果样本集合D根据特征A是否取某一可能值a被分割成D1和D2两部分，即：

$$D_1 = \{(x,y) \in D | A(x) = a\}, D_2 = D - D_1$$

则在特征A的条件下，集合D的基尼指数定义为：

$$Gini(D,A) = \frac{|D_1|}{|D|}\text{Gini}(D_1) + \frac{|D_2|}{|D|}\text{Gini}(D_2)$$

基尼指数Gini(D)表示集合D的不确定性，基尼指数Gini(D,A)表示经A=a分割后集合D的不确定性。基尼指数值越大，样本集合的不确定性也就越大，这一点跟熵相似。

当p1=p2=...=pK=1/K时，G(p)=取得最大值，此时随机变量最不确定。证明：拉格朗日乘子法。

连续属性离散化

分段划分

非监督离散化：等宽离散化、等频离散化、聚类

- 等宽离散化将属性划分为宽度一致的若干个区间
- 等频离散化将属性划分为若干个区间，每个区间的数量相等
- 聚类将属性间根据特性划分为不同的簇，以此形式将连续属性离散化

过拟合问题

是啥？训练样本的正确率高，检验样本误差很大，泛化程度差。

防止过拟合：剪枝，树的深度不能太高

剪枝方法：错误率降低剪枝(REP)：简单快速，数据集大效果不错，小反而不好；一般在第三层以后剪枝

分类效果的评价

指标：训练误差、泛化误差、准确率、错误率等

样本的分类

- 样本为正例，被分类为正例，称为真正类(TP)
- 样本为正例，被分类为反例，称为假反类(FN)
- 样本为反例，被分类为正例，称为假正类(FP)
- 样本为反例，被分类为反例，称为真反类(TN)

准确率：（TP+TN）/（TP+TN+FT+FN）识别正确的概率
精确率（precision）： TP/(TP+FP) 识别为正例的概率
召回率（查全率）：TP/(TP+FN) 识别对的正确样本的概率

有时要为了提高某个指标，牺牲其他指标。

F值为精确率和召回率的调和平均

$$F = \frac{(\alpha^2 + 1) \times accuracy \times recall}{\alpha^2(accuracy + recall)}$$

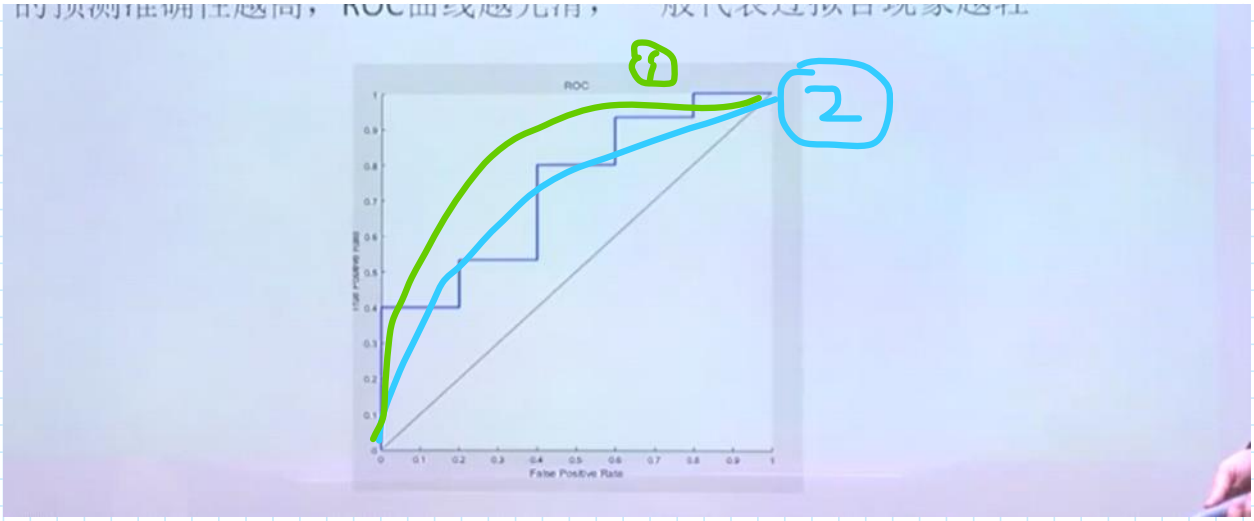
a为调和参数值。
F1 值：当 a = 1 时的取值

$$F_1 = \frac{2 \times accuracy \times recall}{accuracy + recall}$$

受试者工作特征曲线（ROC）

受试者工作特征曲线(ROC)曲线也是一种常用的综合评价指标。假设检验集中共有20个样本，每个样本为正类或反类，根据分类算法模型可以得出每个样本属于正类的概率，将样本按照此概率由高到低排列

ROC曲线下的面积称为AUC(Area under Curve)，AUC值越大，表示分类模型的预测准确性越高，ROC曲线越光滑，一般代表过拟合现象越轻



①优于②

分类效果的评价方法

保留法：将样本集分为训练集和检验集，训练集>样本集

蒙特卡洛交叉验证（重复随机二次采样验证）：多次划分训练集和检验集，多次训练，取平均，本质是多次保留法。

k折交叉验证法：将样本随机划分为k个大小相等的子集，每次选一个为检验集，其余为训练集，重复k次取平均。最常用的是十折交叉验证。

集成学习

分为多棵树，类似多专家决策

装袋法

通过组合多个训练集的分类结果提升分类效果。
每次随机抽取样本中一部分数据集，训练多棵树，最后投票

提升法

引入 权重 概念，多轮训练。
加权平均 效果明显

随机森林

EX装袋法，每次随机抽取t个属性，然后在t个属性中选取最优的作为分支属性，最后投票。

若样本属性有N个，一般取t 为<= log2（N+1）的最大整数。

一般采用CART算法作为决策树。

支持向量机

概念

支持向量机（Support Vector Machine, SVM）是一类按监督学习（supervised learning）方式对数据进行二元分类的广义线性分类器（generalized linear classifier），其决策边界是对学习样本求解的最大边距超平面（maximum-margin hyperplane）[1-3] 。(百度百科)

是分类的一个工具，可以最优化分类。
当在当前维度线性不可分时，可以考虑升维。

核函数

线性核函数
K(x,y) = x·y +c
c是可选常数
主要用于线性可分的情况，一般采用输入空间的内积，维度不变，运算较少

多项式核函数
K(x,y) = [a·x·y+c]^d
a是调参，d是最高项次数，c是可选参数

径向基核函数

$$K(x,y) = \exp\{-[(||x-y||^2)/(2\cdot a^2)]\}$$

即

$$e^{-\frac{\|x-y\|^2}{2a^2}}$$

应用广泛，参数少，类似高斯函数，又称为高斯核函数

Sigmoid核

$$K(x,y) = \tanh(a\cdot x\cdot y+c)$$

a是调参，c是可选参数，一般c取1/n

Tanh 双曲正切

支持向量机应用
适合图像和文本等样本特征较多的应用场合，小样本

贝叶斯

对于任意的x_i，假设存在集合$\varphi(X)$，$\varphi(X) \sqsubseteq \{x_1, \ldots, x_n\}$，使得在$\varphi(X)$确定下，$X$与$\{x_1, \ldots, x_n\} - \varphi(X)$中任意元素条件独立，即有$P(X|x_1, \ldots, x_{i-1}) = p(X|\varphi(x_i))$，于是有：

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\varphi(x))$$

贝叶斯网络大体上就是在有相关关系的点对之间，根据拓扑序连边，权值为相应的条件概率。搞成一个DAG，每次搞得时候可以调整整个网络。

主成分分析和奇异值分解

这个回学校再看，暂时还很懵逼，矩阵都忘干净了。

判别分析

根据已有的分类模型，判断新样本的类别。

主要方式：

按判别的类数：二分类判别分析和多分类判别分析

按所含变量个数，可以分为一元判别分析和多元判别分析

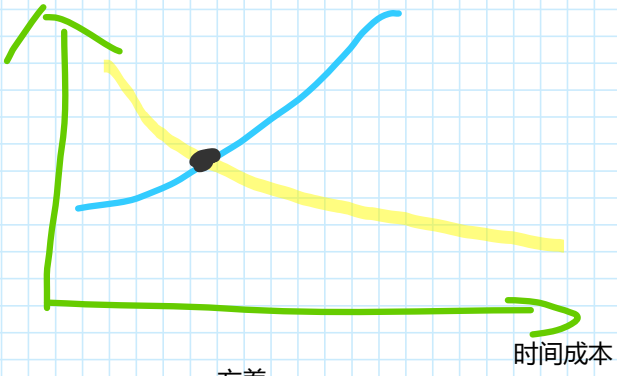
按照判别准则：
距离：求出每个分类的中心坐标，对于新样本，算出它距每个中心坐标的距离，归为最小的一类，距离一般采用马氏距离（协方差距离）或欧氏距离。
Fisher：投影，将高维空间投影到低维空间，在低维空间分类，使类内的离差较小（离差：$\sum xi - x$）。
贝叶斯：在考虑先验概率的前提下，利用贝叶斯公式，按照一定准则构成一个判别函数，计算新样本落入每类的概率。

判别分析-LDA/QDA

- 线性判别分析（LDA）和二次判别分析（QDA）
- 如何选择LDA或QDA？
- 二次判别分析是针对服从高斯分布，且均值不同，方差也不同的样本数据而设计的。对高斯分布的协方差矩阵不做任何假设，直接用每个分类下的协方差矩阵，因为数据方差相同的时候，一次判别就可以，但如果类别间的方差相差较大时，就变成一个关于x的二次函数，则需使用二次决策平面

方差和误差的取舍

高方差低误差，容易学习到样本本身的特征，容易过拟合。
高误差低方差，过于迟钝，容易偏离真实值。



黑点表示一个适合的学习时间。

LDA相对方差更低，QDA相对误差更低。
很明显，在样本量小的时候，选择LDA，样本量大的时候，选择QDA。