上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

分类、聚类、回归、关联规则

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

目录

分类问题

- 通过比较事物的特征，把具有某些共同点或相似特征的事物归属于一个给定的集合，用于进行预测和判断，这就是分类方法分类分析属于监督式学习。
- 在通常的应用中，所提取的特征可以是一组分类（如血型的A、B、O型），可以是数值（如对长度或高度的测量值），也可以是顺序排位（如一群人按身高的排名）。
- 分类分析属于监督式学习。

分类问题的实例

- 已知某种传染病的病症，在一个人群集合中根据每个人的表现区分是否感染
- 已有一定的电子邮件样本，通过学习，判断一封新的邮件是否是垃圾邮件

分类问题

- 分类方法：
- 贝叶斯法、非参数法、决策树法、规则归纳法、神经网络法、SVM

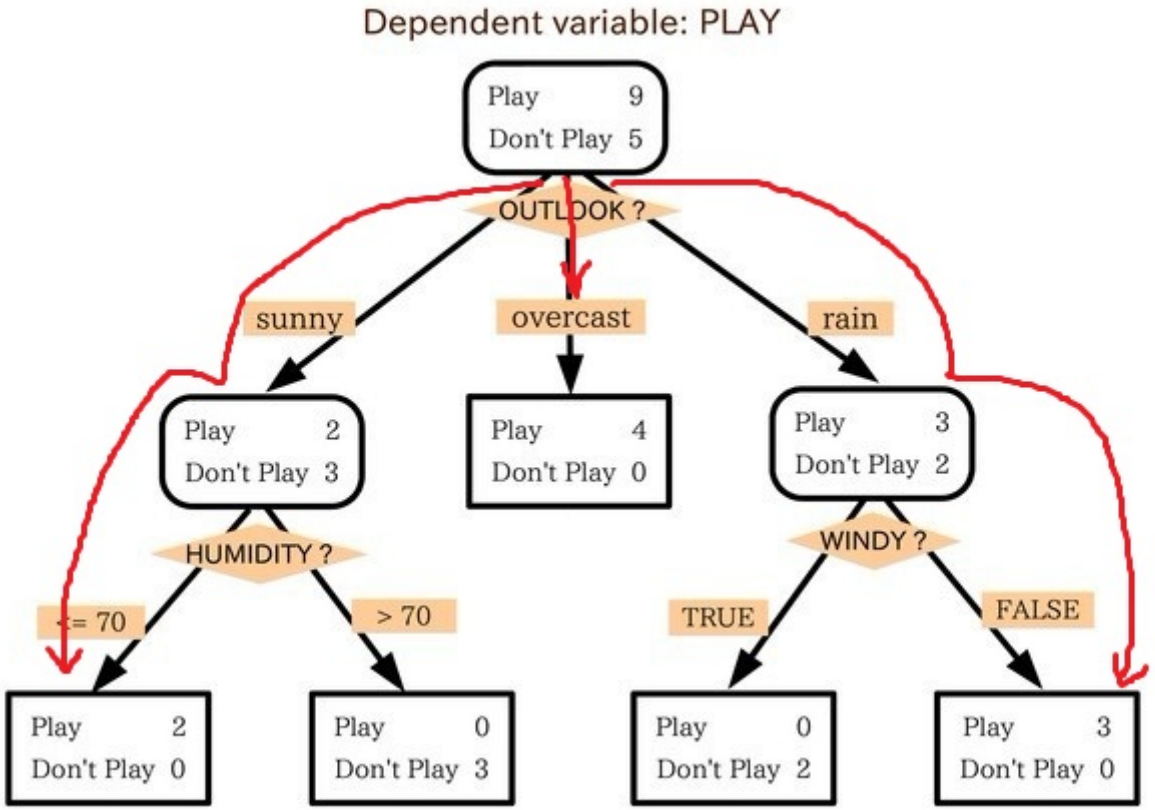
贝叶斯法

- 原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。
- 应用贝叶斯网络分类器进行分类主要分成两阶段：
- 第一阶段是贝叶斯网络分类器的学习，即从样本数据中构造分类器，包括结构学习和CPT 学习；
- 第二阶段是贝叶斯网络分类器的推理，即计算类结点的条件概率，对分类数据进行分类。

朴素贝叶斯分类

- 1、设x={a_1,a_2,..,a_m}为一个待分类项，而每个a为x的一个特征属性。
- 2、有类别集合C={y_1,y_2,..,y_m}。
- 3、计算P(y_1|x), P(y_2|x),.., P(y_m|x) 。
- 4、如果P(y_k|x)=max{P(y_1|x), P(y_2|x),.., P(y_m|x) }，则x∈ y_k。

决策树法


Dependent variable: PLAY
Play 9
Don't Play 5
OUTLOOK ?
sunny
overcast
rain
Play 2
Don't Play 3
Play 4
Don't Play 0
Play 3
Don't Play 2
HUMIDITY ?
WINDY ?
<= 70
> 70
TRUE
FALSE
Play 2
Don't Play 0
Play 0
Don't Play 3
Play 0
Don't Play 2
Play 3
Don't Play 0

决策树法

- a.对当前例子集合，计算属性的信息增益；
- b.选择信息增益最大的属性A_i(关于信息增益后面会有详细叙述)；
- c.把在A_i处取值相同的例子归于同于子集，A_i取几个值就得几个子集；
- d.对依次对每种取值情况下的子集,递归调用建树算法，即返回a；
- e.若子集只含有单个属性，则分支为叶子节点，判断其属性值并标上相应的符号，然后返回调用处。

神经网络法

- 人工神经网络首先要以一定的学习准则进行学习，然后才能工作。现以人工神经网络对于写“A”、“B”两个字母的识别为例进行说明，规定当“A”输入网络时，应该输出“1”，而当输入为“B”时，输出为“0”。
- 如果网络作出错误的的判决，则通过网络的学习，增加判决正确的概率，应使得网络减少下次犯同样错误的可能性。
- 人工神经网络就是模拟人思维的第二种方式。这是一个非线性动力学系统，其特色在于信息的分布式存储和并行协同处理。

神经网络法

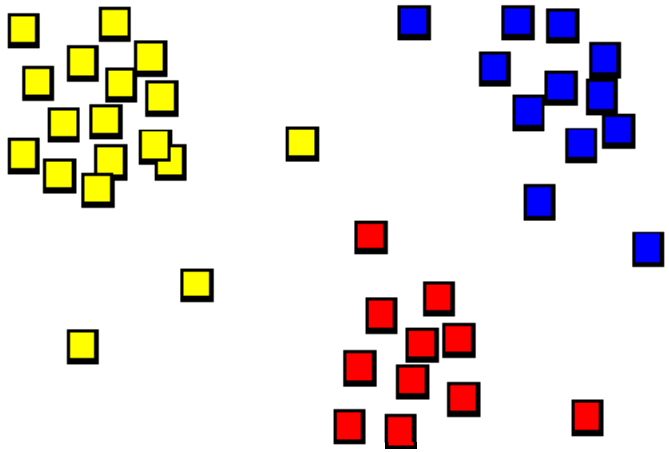
- 1、初始化网络权值和神经元的阈值（最简单的办法就是随机初始化）
- 2、前向传播：按照公式一层一层的计算隐层神经元和输出层神经元的输入和输出。
- 3、后向传播：根据公式修正权值和阈值
- 直到满足终止条件。

聚类问题

- 聚类是将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程，目标是在相似的基础上收集数据来分类。
- 在分类的过程中，不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。
- 聚类分析属于非监督式学习。

聚类问题实例

- 采集一片鸢尾花田中每一株的花萼长宽、花瓣长宽数据，把性状相似的归为同一个集合
- 下图是将平面上的点聚类的一个例子



聚类分析方法

- 主要思想是通过判断各个样本之间的距离，将其中距离较近的归为同一类
- 常用的方法有系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类

- 主要步骤：
- 1. 数据预处理，
- 2. 为衡量数据点间的相似度定义一个距离函数，
- 3. 聚类或分组，
- 4. 评估输出。

系统聚类法

- 首先根据一批数据或指标找出能度量这些数据或指标之间相似程度的统计量；
- 然后以统计量作为划分类型的依据，最初每个样本各自看成一类，把一些相似程度大的样本首先聚合为一类，计算出新类到其他各类的距离，再将距离最近的两类合并，这样直到合并成一类；
- 把每一步分类过程用树状图的形式记录下来，逐步画成一张完整的分类系统图，又称谱系图。其相似程度由距离或者相似系数定义。

动态聚类法

- 首先选择若干个样本点作为聚类中心，再按照事先确定的聚类准则进行聚类，在聚类过程中，根据聚类准则对聚类中心反复修改，直到分类合理为止。
- 在聚类过程中，样品的属类可以发生改变。
- 当样品数较大时，计算量较小，占用内存空间较小。
- 初始分类对最终结果可能有影响，可能会产生局部最优解。

动态聚类法步骤

- （1）随机选取两个点$x_1^{(1)}$和$x_2^{(1)}$作为聚核；
- （2）对于任何点x_k，分别计算到两个聚核之间的距离；
- （3）若$d(x_k, x_1^{(1)}) < d(x_k, x_2^{(1)})$，则将$x_k$划为第一类，否则划为第二类；
- （4）分别计算两个类的重心，得到$x_1^{(2)}$和$x_2^{(2)}$，以其为新的聚核，对空间中的点进行重新分类；
- （5）重复以上步骤。

模糊聚类法

- 一般用在事物聚类之间的界线难以精确划分时，广泛应用在气象预报、地质、农业、林业等方面。
- 一个样本可能同时属于不同的类，通过隶属度来区别。例如把温度分为冷热两类，比如5度，可能属于冷这类的隶属度值为0.7,而属于热这个类的值为0.3。
- 根据研究对象本身的属性来构造模糊矩阵，并在此基础上根据一定的隶属度来确定聚类关系，即用模糊数学的方法把样本之间的模糊关系定量的确定，从而客观且准确地进行聚类。

模糊聚类法步骤

- (1)对数据进行标准化处理；
- (2)建立模糊相似矩阵$R=(s_{ij})^{n\times n}$，其中s_{ij}为相似系数，其定义可以有多种形式：夹角余弦，相关系数或距离。R满足自反性、对称性；
- (3)根据模糊相似矩阵创建模糊等价矩阵R^*， R^*满足自反性、对称性、传递性。可以采用传递闭包法；
- (4)选取截取水平$\lambda(0<\lambda<1)$，直接由模糊相似矩阵R作其λ截矩阵R_λ，再根据R_λ对数据进行划分。

K均值聚类

- 给定一个数据点集合和需要的聚类数目k，k由用户指定，k均值算法根据某个距离函数反复把数据分入k个聚类中。
- 算法的目的是使各个样本与所在类均值的误差平方和达到最小。

K均值聚类步骤

- (1)初始化：输入基因表达矩阵作为对象集X，输入指定聚类类数K，并在X中随机选取K个对象作为初始聚类中心；设定迭代中止条件，比如最大循环次数或者聚类中心收敛误差容限。
- (2)进行迭代：根据相似度准则将数据对象分配到最接近的聚类中心，从而形成一类。初始化隶属度矩阵。
- (3)更新聚类中心：然后以每一类的平均向量作为新的聚类中心，重新分配数据对象。
- (4)反复执行第二步和第三步直至满足中止条件。

回归问题

- 对于给出的两种或多种变量，研究它们之间的相互依赖的定量关系，这样的分析方法称为回归分析，最常用的是线性回归。
- 按照自变量的多少，可分为一元回归分析和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。
- 回归分析属于监督式学习。

回归问题实例

- 已知某个地区近几年的人口变化情况与各种经济数据，求房价走势与这些数据的内在联系并判断未来趋势。
- 已知一段时间内的天气情况和某种农作物的产量，求两者之间的相关关系。

回归问题的求解方法

- （1）确定因变量与自变量间的定量关系表达式，这种表达式称为回归方程；
- （2）对求得的回归方程的可信度进行检验；
- （3）判断自变量对因变量有无影响；
- （4）利用所求得的回归方程进行预测和控制。

线性回归

- 利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模。
- 如果只包括一个自变量和一个因变量，且二者关系可用一条直线近似表示，则称为一元线性回归。
- 回归分析中包括两个两个以上的自变量，且因变量和自变量间是线性关系，则称为多元线性回归。
- 线性回归把焦点放在给定X值的y的条件概率分布，而不是X和y的联合概率分布。

线性回归方法

- 收集的数据中每一个分量可以看做一个特征数据，每个特征至少对应一个未知的参数，可以据此列出一个线性模型函数，向量表示形式为$h_\Theta=\Theta^T X$。
- 可以定义损失函数为所有样本点到拟合得到直线上点的距离的平方和，求其最小值

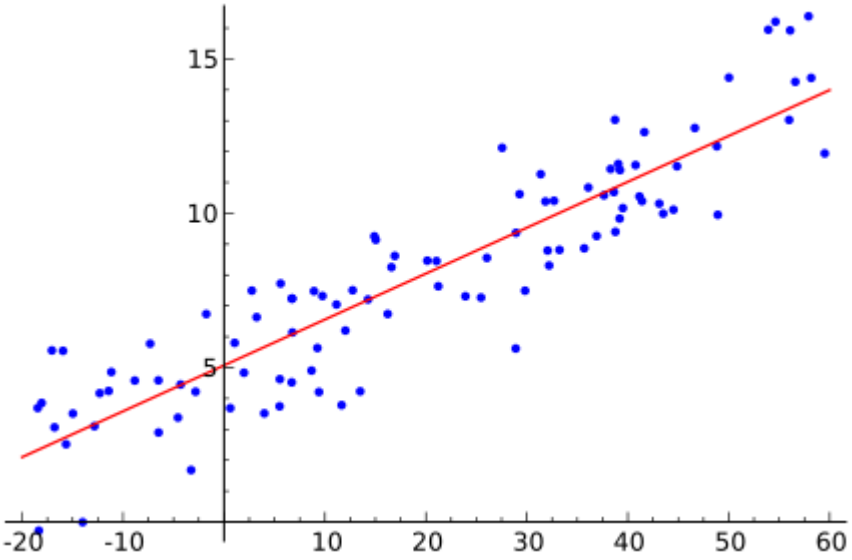
$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$$

$$\min_\theta J_\theta$$

- 利用最小二乘法或梯度下降法求解满足这个方程的误差函数最小的解，得到线性回归方程。

线性回归实例

- 给定平面上的一些点，求一条直线使得所有点到该直线的距离的平方和最小
- 使用最小二乘法求该最小状态



逻辑回归

- 逻辑回归于线性回归的不同之处在于其因变量取值为0或者1，而线性回归的因变量是连续的。
- 一般用逻辑回归来处理分类问题。
- 逻辑回归的因变量满足二项分布，采用sigmoid函数进行类别的判断

$$g(z)=\frac{1}{1+e^{-z}}$$

逻辑回归方法

- (1)寻找预测函数h，函数形式与决策边界有关，线性边界分类的预测函数为

$$\theta_0 + \theta_1 x_1 + ,\ldots, + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i = \theta^T x$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

- (2)基于最大似然函数求导构造损失函数j

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right]$$

- (3) 使得j函数最小并求得回归参数（θ），可以采用梯度下降法。

三者的联系和区别

- 聚类与分类的不同在于，分类中的类的数量与内容是给出的，而聚类所要求划分的类的数量和内容是未知的
- 分类分析和回归分析属于监督式学习，而聚类分析属于非监督式学习
- 三者都可以用于预测
- 分类和聚类的输出是离散的类别值，而回归的输出是一个函数表达式

关联规则

- 关联规则是隐藏在数据项之间的关联或相互关系，即可以根据一个数据项的出现推导出其他数据项的出现。常用在商业分析领域。
- 关联规则模式属于描述型模式，发现关联规则的算法属于非监督式学习。

关联规则实例：尿布与啤酒

- 沃尔玛拥有世界上最大的数据仓库系统，为了能够准确了解顾客在其门店的购买习惯，沃尔玛对其顾客的购物行为进行购物篮分析，想知道顾客经常一起购买的商品有哪些。沃尔玛数据仓库里集中了其各门店的详细原始交易数据。在这些原始交易数据的基础上，沃尔玛利用数据挖掘方法对这些数据进行分析和挖掘。一个意外的发现是：跟尿布一起购买最多的商品竟是啤酒！
- 经过大量实际调查和分析，揭示了一个隐藏在"尿布与啤酒"背后的美国人的一种行为模式：在美国，一些年轻的父亲下班后经常要到超市去买婴儿尿布，而他们中有30%～40%的人同时也为自己买一些啤酒。
- 进一步进行分析，发现产生这一现象的原因是：美国的太太们常叮嘱她们的丈夫下班后为小孩买尿布，而丈夫们在买尿布后又随手带回了他们喜欢的啤酒。

关联规则分类

- 变量可以是布尔型或者数值型，表示类型的判定与数量的差别。
- 抽象层次可以是单层或者多层，进行不同层次之间的映射。例如：IBM台式机=>Sony打印机，是一个细节数据上的单层关联规则；台式机=>Sony打印机，是一个较高层次和细节层次之间的多层关联规则。
- 涉及的数据维数可以是一维或者多维的，例如从平面上的向量到多维空间中的向量的映射。

关联规则分析方法

- (1)先确定分析样本，选定评判的支持度与可信度要求
- (2)从资料集合中找出所有的高频项目组
- (3)利用算法由这些高频项目组中产生关联规则
- (4)分析和评估

- 相关算法：
- Apriori算法
- 基于划分的算法
- FP-树频集算法

关联规则中的基本概念

- 关联规则具有如下两个重要的属性：
- 支持度: P(A∪B)，即A和B这两个项集在事务集D中同时出现的概率。
- 置信度: P(B | A)，即在出现项集A的事务集D中，项集B也同时出现的概率。
- 同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。
- 挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则，也就是产生强规则的问题。

Apriori算法

- 算法的核心是基于两阶段频集思想的递推算法。
- 该关联规则在分类上属于单维、单层、布尔关联规则。
- 在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。
- 算法的缺点：可能产生大量的候选集,可能需要重复扫描数据库。

Apriori算法过程

- 首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。
- 然后由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。
- 然后使用第1步找到的频集产生期望的规则，产生只包含集合的项的所有规则，其中每一条规则的右部只有一项。
- 一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。
- 使用递推的方法生成所有频集。

基于划分的算法

- 先把数据库从逻辑上分成几个互不相交的块，每次单独考虑一个分块并对它生成所有的频集。
- 然后把产生的频集合并，用来生成所有可能的频集，最后计算这些项集的支持度。
- 该算法可以高度并行，可以把每一分块分别分配给某个生成频集。产生频集的每一个循环结束后，处理器之间进行通信来产生全局的候选k-项集。
- 通常这里的通信过程是算法执行时间的主要瓶颈；而另一方面，每个独立的处理器生成频集的时间也是一个瓶颈。

FP-树频集算法

- 算法采用分而治之的策略，在经过第一遍扫描之后，把数据库中的频集压缩进一棵频繁模式树（FP-tree），同时依然保留其中的关联信息。
- 随后再将FP-tree分化成一些条件库，每个库和一个长度为1的频集相关，然后再对这些条件库分别进行挖掘。
- 当原始数据量很大的时候，也可以结合划分的方法,使得一个FP-tree可以放入主存中。
- 此算法对不同长度的规则都有很好的适应性，同时在效率上较之Apriori算法有巨大的提高。

多层关联规则挖掘

- 在对于支持度的设置方面与单层关联规则挖掘有所区别。
- 可以采用两种支持度策略：
- 1) 统一的最小支持度。对于不同的层次，都使用同一个最小支持度。
- 2) 递减的最小支持度。每个层次都有不同的最小支持度，较低层次的最小支持度相对较小。

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

谢　谢！