5. 得分点：数据抓取、预处理、特征工程、分类、聚类、回归、推荐、神经网络、其他算法、模型评估、数据可视化、MVC.net数据可视化、Python数据库操作、大作业文档、同学们打分、答辩效果

# 一、数据抓取

数据抓取：常用抽取网页数据的方式有三种：**正则表达式、Beautiful Soup、lxml**

|  |  |  |  |
| --- | --- | --- | --- |
| 三种方式的比较 | | | |
| 抓取方法 | 性能 | 使用难度 | 安装难度 |
| 正则表达式 | 快 | 困难 | 简单（内置模块） |
| Beautiful Soup | 慢 | 简单 | 简单（纯python） |
| Lxml | 快 | 简单 | 相对困难 |

# 二、数据预处理

**数据预处理**主要包括**数据清洗、数据集成、数据变换**和**数据规约**四个部分。

## （一）数据清洗

### 1、概念

删除原始数据集中的无关数据、重复数据、平滑噪声数据，处理缺失值、异常值等。

### ****2、数据清洗的步骤****

（1）缺失值处理（通过describe与len直接发现、通过0数据发现）

（2）异常值处理（通过散点图发现）

一般遇到缺失值，处理方式为：删除、插补、不处理

 插补的方式主要有：均值插补、中位数插补、众数插补、固定值插补、最近数据插补、回归插补、插值法等等。

一般遇到异常值，一般处理方式为：视为缺失值、删除、修补（平均数、中位数等等）。

## ****（二）、数据集成****

将多个数据源合并存在一个一致的数据存储中，要考虑实体识别问题和属性冗余问题，从而将数据在最低层上加以转换、提炼和集成。

## ****（三）、数据变换****

### 1、概念

即对数据规范化处理，主要包括以下几种：

（1）**简单函数变换**。（2）**数据规范化**。（3）**数据离散化**。（4）**属性构造。**

### 2、简单函数变换的常见方法

平方、开方、对数、差分运算等。

### ****3、数据规范化的常见方法****

（1）离差标准化（最小-最大标准化）--消除量纲（单位）影响以及变异大小因素的影响。（最小-最大标准化）

                        x1=（x-min）/（max-min）

  （2）标准差标准化--消除单位影响以及变量自身变异影响。（零-均值标准化）

                       x1=（x-平均数）/标准差

  （3小数定标规范化--消除单位影响。

                       x1= x/10\*\*(k)、k=log10(x的绝对值的最大值)

### 4、 ****数据离散化的常见方法****

   （1）等宽离散化。

   （2）等频率离散化。

   （3）一维聚类离散化。

### 5、****属性构造****

即通过现有的一个或多个属性数据经过处理变成新的属性数据。

## ****（四）、数据规约****

### ****1、概念****

产生更小且保持数据完整性的新数据集。意义在于降低无效、错误数据；降低存储成本；少量且具有代表性的数据大幅加快，主要分为以下两类：

### 2、****属性规约****

（1）属性合并或删除无关维，目标是寻找最小子集使子集概率分布尽可能与原来相同。

 （2） 常用方法：

  ①合并属性 将就属性合并为新属性 {A1,A2,A3,B1,B2,C}——{A,B,C}  
    ②逐步向前选择 从空集开始，逐个加入最优属性，直到无最优或满足条件 {}—{A1}—{A1,A4}  
     ③逐步向后删除 从全集开始，每次删除最差属性，直到无最差或满足阈值  
     ④决策树归纳 利用决策树归纳能力进行分类，删除未出现的属性，即效果差的属性  
   ⑤主成分分析 用少量变量解释大部分变量，保留大部分信息，将相关性高的数据转为彼此独立

### 3、****数值规约****

通过选择替代的、较小的数据来较少数据量，包括有参数(回归、对数线性模型)和无参数方法(直方图、聚类、抽样)。

# 三、K-means算法

1、K-means算法中的k表示的是聚类为k个簇，means代表取每一个聚类中数据值的均值作为该簇的中心，或者称为质心，即用每一个的类的质心对该簇进行描述。即，K-Means算法接受参数K；然后将样本数据集划分为K个聚类。获得的聚类需要满足：同一个聚类中的样本数据集相似度较高；而不同聚类中的样本数据集相似度较小。

算法思想为：以空间中K个点为中心进行聚类（即先从样本集中随机选取 k个样本作为簇中心），对最靠近他们的对象归类（所有样本与这 k个“簇中心”的距离，对于每一个样本，将其划分到与其距离最近的“簇中心”所在的簇中）。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

2、算法流程：

第一步，先从没有标签的元素集合A中随机取K个元素，作为K个子集各自的质心。

第二步，分别计算剩下的元素到K个子集质心的距离，根据距离将元素分别划分到最近的子集。

第三步，根据聚类结果，重新计算质心(计算方法为子集中所有元素各个维度的算术平均数)

第四步，将集合A中全部元素按照新的质心然后再重新聚类。

重复第四步，直到聚类结果不再发生变化。

# 四、岭回归

关于Ridge？

原理:岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价，获得回归系数更为符合实际、更可靠的回归方法，对病态数据的耐受性远远强于最小二乘法。

缺点：通常岭回归方程的R平方值会稍低于普通回归分析，但回归系数的显著性往往明显高于普通回归，在存在共线性问题和病态数据偏多的研究中有较大的实用价值。

适用情况：主要适用于过拟合严重或各变量之间存在多重共线性的时候

1.岭回归可以解决特征数量比样本量多的问题

2.岭回归作为一种缩减算法可以判断哪些特征重要或者不重要，有点类似于降维的效果

3.缩减算法可以看作是对一个模型增加偏差的同时减少方差