

理学院学生实习系列指导手册之六

统计分析与数据挖掘技能训练

—实习报告

**系 别： 统计系**

**专 业： 统计学**

**班 级： 统计1801**

**姓 名： 刘士坤**

**学 号： 20183744**

**指导教师： 藏薇**

|  |  |
| --- | --- |
| **1.课程实习记录表** | |
| **实习时间** | 2020.11.2——2020.11.7 |
| **实习地点** | 6教201室 |
| **实习内容** | 多元统计分析方法 |
| **实习计划** | |
| **练习并掌握以下几种多元分析方法及其背后理论基础。**   * 聚类分析（Cluster analysis）   + 系统聚类（Hierarchical clustering）   + K‐means 聚类（k‐means clustering） * 线性判别分析（Linear discriminant analysis） * 主成分分析（Principal component analysis） * 因子分析（Factor Analysis）   软件：IBM SPSS  数据集：judges 系统聚类.sav  telco 快速聚类.sav  iris 判别分析.sav  factor1 主成分分析.sav  Olympic88 因子分析.sav | |

|  |
| --- |
| **实习记录** |
| **系统聚类**  对 judges 系统聚类.sav 数据集做变量聚类，此数据集有8个变量，选择组间联结法和Pearson系数，变量不进行标准化，得到如下树状聚类图。    **2020年11月2日** |
| **实习记录** |
| **K-means 聚类**  对 telco 快速聚类.sav 数据集做快速聚类，共有3395个样品，首先对变量进行标准化，即Z-Score处理，选择对变量聚成2类的结果。    从方差分析表中可以看出这6个变量对分类贡献均显著。  **2020年11月3日** |

|  |
| --- |
| **实习记录** |
| **线性判别分析**  对 iris 判别分析.sav 数据集进行判别分析，该数据集样品分为三类：刚毛鸢尾花、变色鸢尾花、佛吉尼亚鸢尾花。我们利用此数据集训练出的分类函数对新的样品进行分类。  分类函数为        从验证的分类结果可以看出分类函数对样本做出了很好的分类。  **2020年11月 4日** |
| **实习记录** |
| **主成分分析**  对 factor1 主成分分析.sav 数据集进行主成分分析，该数据集含有8个变量，提取到的三个主成分分别为：  从方差解释表中能够看出这3个主成分解释了原始变量89.551%的方差。  **2020年11月5日** |

|  |
| --- |
| **实习记录** |
| **因子分析**  对 Olympic88 因子分析.sav 数据集进行因子分析，该数据集共10个变量，用主成分法提取出4个公共因子，分别为  因子模型中公共因子的系数矩阵    **2020年11月6日** |
| **实习记录** |
| 总结各多元分析方法的特点，适用数据集类型，不同分析方法之间的差别，如主成分法和因子分析方法的差别。  回顾各多元分析方法的理论基础，特别是线性判别法、主成分法、因子分析法背后的理论基础。  综合运用聚类方法、判别方法、主成分法和因子分析方法对数据集进行分析，给出分析结果。  **2020年11月7日** |

**2. 学生实习成绩评定表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **姓名** | 刘士坤 | **学号** | 20183744 | **班级** | 统计1801 |
| **实**  **习**  **总**  **结**  **与**  **体**  **会** | 通过此次实习，我掌握了对数据进行处理和分析的多元分析方法。具体来说有聚类方法，如系统聚类和 K-means 聚类；线性判别方法，如Fisher线性判别，Bayes判别；主成分分析方法；因子分析方法。  其中聚类方法是无监督的，判别方法是有监督的。  Fisher线性判别（LDA）本质是投影，使得投影后的数据能够尽可能地分离开来，以此作为分类标准。  Bayes判别是将Bayes思想应用到判别上，计算出使得平均损失最小的分类函数，以此作为分类标准。  主成分方法本质是坐标轴旋转，将数据投影到方差最大的方向，就是旋转坐标轴到方差最大的方向，然后选取前几个方差较大的主成分来代替原始变量。  因子分析方法是找出变量背后的公共因子，将变量表示成公共因子的函数，具体方法有主成分法、主轴因子法、极大似然法、最小二乘法等。采用不同的方法求出的因子不同，通过因子旋转也能得到不同的因子。  要真正地掌握以上几种多元分析方法，就一定要真正地理解协方差矩阵，要知道协方差矩阵的几何意义，要知道协方差矩阵是怎么通过单位矩阵线性变换得到的。还要有较强的线性代数知识，对线性映射、线性变换要掌握的很好。矩阵的特征分解、SVD分解等也要掌握的很好。  总而言之，线性代数的基础非常好的话，这些都没有什么问题，都只是其应用罢了。 | | | | |
| **答**  **辩**  **记**  **录**  **与**  **评**  **价** | 报告共使用了4钟多元分析方法，分别是系统聚类、K-means 聚类、主成分分析、线性判别分析。所采用的数据为某中学火箭班、实验班、重点班、普通班80名同学的期中考试成绩，成绩构成为语文、数学、英语、物理、化学、生物、历史、地理、政治。  用系统聚类方法将这9个科目聚成4类，语文为一类，数学、物理、化学、生物、历史、地理为一类，英语为一类，政治为一类。  用 K-means 聚类将80名同学分为4类。  用5个主成分代替9个变量做线性判别分析，这5个主成分共解释了原始变量95.532%的方差，对5个主成分做均值相等的检验为显著，协方差矩阵的 Box’M 检验为显著，但由分类函数的得出的分类结果看判别效果还是比较好的，有82.5%的比例分类正确，留一验证法也有80.0%的正确率。  本次报告所使用的的幻灯片地址：  <https://github.com/loreliu/Multivariate-analysis/releases/download/2020-11-2/multianalysis.pdf> | | | | |
| **答**  **辩**  **记**  **录**  **与**  **评**  **价** |  | | | | |
| **答**  **辩**  **记**  **录**  **与**  **评**  **价** |  | | | | |
| **答**  **辩**  **记**  **录**  **与**  **评**  **价** |  | | | | |

|  |  |
| --- | --- |
| **答**  **辩**  **记**  **录**  **与**  **评**  **价** | 记录人签名： |
| **实**  **习**  **成**  **绩** | **优 良 中 及格 不及格**  指导教师签名：  年 月 日 |
| **备**  **注** |  |

天津商业大学理学院制2020年10月修订