附　录

附录 A　开题报告

毕业设计（论文）开题报告

学　院	理学院	教学系	统计系	专业班级	统计学专业 1801
学生姓名	刘士坤	学号	20183744	指导教师	王倩
毕业设计（论文）题目	线性判别分析的原理与应用				

一、选题依据

1. 研究背景及意义

在生产生活中经常遇到如何根据观测到的数据资料对所研究的对象进行判别归类的问题。例如：在医学诊断中，一个病人肺部有阴影，医生要判断他患的是肺结核、肺部良性肿瘤还是肺癌？在气象学中，根据已有气象资料（气温、气压、湿度等）来判断明天是阴天还是晴天，是有雨还是无雨。在地质勘探中，需要从岩石标本的多种特征来判断地层的地质年代，是有矿还是无矿，是富矿还是贫矿。判别分析发展至今天已渗透到各个领域。但不管是哪个领域，判别分析问题都可以这样描述：设有 k 个 m 维总体 $G_1, G_2, \cdots, G_k$，其分布特征已知（如已知分布函数分别为 $F_1(x), F_2(x), \cdots, F_k(x)$, 或知道来自各个总体的训练样本）。对给定的一个新样品 X，我们要判断它来自哪个总体[1][2]。

可以说判别问题的实质就是认为样本的各种指标数据来自于不同的分布，然后估计其中的参数，获得总体的统计特征，以此来进行判别分类。基于此，在进行判别归类时由假设前提、判别依据及处理手法的不同可得到不同判别方法，线性判别分析就是其中一种。线性判别分析是 Fisher 在其 1936 年的经典论文[3]中提出的，论文中同时也公布了现如今经常出现在统计学习或机器学习中的一个示例数据集——鸢尾花数据集，Fisher 就是拿这个数据集做了线性判别分析来对鸢尾花进行分类。但其实 Fisher 在其论文中也明确提到了鸢尾花的测量数据是用来说明 Iris versicolor 是 Iris virginica 与 Iris setosa 的中间类型，拿来实际分类的准确度并不高[4]，三种鸢尾花的判别依据是种子，尽管 Fisher 自己就是用这些测量数据做了一个线性判别分析。线性判别分析基本思想的几何解释是：通过最大化 Fisher 准则，找到最佳的投影方向，使得投影后的数据类间距离大的同时类内距离(方差)小，以此达到判别分类的目的。

本毕业论文通过对线性判别分析的学习，了解其发展过程、基本思想、原理、数学计算方法[5][6]以及判别效果的检验，并将线性判别分析应用到搜集的数据中，同时利用多元统计分析的原理、方法，借助于统计软件进一步探索数据，以期获得一些有用的结论。

2. 国内外研究现状

线性判别分析自 1936 年由 Fisher 首次提出并应用于生物分类。1968 年 Altman 将线性判别分析引入基于财务比率和其他金融变量的破产预测中并提出了 Altman Z-score 模型，是第一个用来系统解释公司进入破产或存活的统计学工具。1996 年 Belhumeur 将其引入模式识别和人工智能领域[7]，线性判别分析作为一种降维特征提取方法[8]，广泛应用于语音识别[9]、人脸识别[10]、步态识别[11]、手势识别[12]、行人再识别[13]等领域。

二、研究内容和研究方法

1. 研究内容

首先对判别分析部分进行概述。距离判别通过仿照似然比导出判别函数的其实就是个二次型的差，和欧几里得空间中的距离是一致的。基于此 Fisher 提出了只考虑判别函数是线性的那种情况，找到最优的线性判别函数，当然这也可以从几何投影的角度来解释（协方差矩阵线性变换）。至于 Bayes 判别，则是直接拿 Bayes 的思想放到判别上，计算出使得平均损失（风险函数）最小的那个判别函数（Bayes 解）。

在线性判别分析部分介绍线性判别分析求解的计算方法，如（广义）特征值，（广义）瑞利商及极值计算。之后介绍判别效果的检验及各变量判别能力的检验。在总体分布未知的情况下去做判别，确实没有什么好的方法，以上方法得到的结果不够理想也只能接受。所以在此说明与其他统计学经典方法及机器学习的分类方法，如 Logistic 回归，SVM、Boosting 模型等的比较。

2. 研究方法

对数据进行探索性数据分析，利用统计软件对获得的数据做探索性数据分析，观察数据的分布特征，如可能可做关于分布的检验。对数据做线性判别分析，对取得的数据做线性判别分析，求得判别函数。进行假设检验评价模型分类效果，如 CV、ROC。与其他分类方法的比较，将其他经典统计方法和机器学习方法应用到上述数据中，比较模型之间的分类效果。

三、预计可获得的成果

本毕业论文的动机在于熟悉多元分析和统计学中经典的线性判别分析及其算法。从探索性数据分析到线性判别分析，从观察到数据分析给出证据，从未知到已知。就如同 80 年前 Fisher 及其同时代的人如 Anderson 走过的路一样，找寻躲藏在随机性背后的小精灵。

通过多种模型的对比，避免单一模型的从数据到结论，能看得到更多的东西。认识到在机器学习、数据挖掘盛行的今天，经典统计学方法或许拟合地不够好，预测地不够准，但其检验和解释却是“黑箱子”所缺少的，这应该也是统计学在面对机器学习、数据挖掘冲击的情况下得以生存的本领吧。

四、工作进度计划

2022.02.20——2022.04.30　回顾多元分析及代数知识，阅读相关书籍和论文；

2022.03.15——2022.03.31　撰写开题报告；

2022.04.01——2022.04.30　搜集论文数据，完成论文初稿；

2022.05.01——2022.05.12　写出其他部分如摘要等；

2022.05.13——2022.05.28　根据导师意见对论文进行修改、校对、定稿。

五、与开题有关的主要参考文献

[1] 陈希孺, 倪国熙. 数理统计学教程[M]. 合肥: 中国科学技术大学出版社, 2009: 305-337.

[2] 高惠璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2005: 175-215.

[3] Fisher, R. A.. The use of multiple measurements in taxonomic problems[J]. Annals of eugenics. 1936, 7(2): 179-188.

[4] Thomas Lumley. The ‘iris’ data[EB/OL]. https://notstatschat.tumblr.com/post/155194690691/the-iris-data, 2016-12-31.

[5] 梁露方. Fisher 线性判别分析问题的求解算法研究[D]. 昆明: 云南师范大学, 2020.

[6] 李卫平, 沈海斌. 基于接近函数的线性判别分析算法研究[J]. 电子技术（上海）, 2017, 46(2):3.

[7] James, G.等. 统计学习导论——基于 R 应用[M]. 王星等, 译. 北京: 机械工业出版社, 2015: 89-117.

[8] 崔自峰, 吉小华. 基于线性判别分析的特征选择[J]. 计算机应用, 2009(10):5.

[9] 谢达东, 吴及, 王作英. 线性判别分析在汉语语音识别中的应用[J]. 计算机工程与应用, 2002, 38(023):1-2,8.

[10] 周大可, 杨新, 彭宁嵩. 改进的线性判别分析算法及其在人脸识别中的应用[J]. 上海交通大学学报, 2005, 39(4):4.

[11] 韩鸿哲, 王志良, 刘冀伟, 李正熙, 陈锋军. 基于线性判别分析和支持向量机的步态识别[J]. 模式识别与人工智能, 2005, 18(2):5.

[12] 温俊芹, 王修晖. 基于线性判别分析和自适应 K 近邻法的手势识别[J]. 数据采集与处理, 2017, 32(3):6.

[13]霍中花, 陈莹. 采用增量式线性判别分析的行人再识别[J]. 小型微型计算机系统, 2017, 38(3):6.

指导教师意见

同意本课题进入设计（论文）阶段。

指导教师签字：　　年　月　日

说明：1.本报告必须在第八学期开学两周内经指导教师审阅并形成正式报告。

2.本报告作为指导教师审查学生能否开展课题研究和是否按时完成进度的检查依据，并接受学校的抽查。