**课程教学大纲-统计方法与机器学习**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程代码** | **DATA0031131806.01** | **课程性质** | | **专业必修** |
| **课程名称：** | **统计方法与机器学习** | | | |
| **英文名称** | **Statistical Methods and Machine Learning** | | | |
| **学时/学分** | **85学时/4学分** | **其中实验/实践学时** | **34学时** | |
| **开课单位** | **数据科学与工程学院** | **适用专业：** | **数据科学与大数据技术** | |
| **先修课程** | **高等数学、线性代数、概率论与数理统计、离散数学、数据科学与工程的数学基础** | | | |
| **大纲撰写人** | **董启文、倪葎** | **大纲审核人** | **陈宇皓** | |
| **课程网址** | **https://www.shuishan.net.cn/education/course/1565728402490331138** | **授课语言** | **中文** | |

注：课程性质选择下列类别之一：学科基础、大类平台、专业必修、专业选修、教师教育

一、课程说明

经典的统计方法依托概率模型，用于刻画数据中的不确定性。它采用特定的统计手段对被测对象进行描述、推断以及预测，以洞察数据背后的深层规律或者模型。统计学方法是归纳思想的具体应用，它从数据中总结规律。然而，随着互联网、云计算和物联网技术的发展，应用中积累的数据量越来越多，硬件平台的存储和计算能力越来越强，获得的数据常常表现出高维、异构、多源、缺失等特征，仅仅依赖少量参数的概率模型无法很好地刻画复杂的数据，从而难以精确洞察数据中隐含的深层规律。

机器学习运用诸如人工神经网络、决策树、随机森林、支持向量机等计算机算法发掘复杂数据背后隐藏的模型，并运用此模型实现对被测对象的预测。从机器学习模型的实际预测效果来看，在诸如文本、图像和音频数据等应用中取得了非常大的成功，它提供了分析和挖掘数据的新方法和新思路。相比较于经典统计学方法，机器学习方法常常较为复杂，模型参数可能会爆炸式增长，但这可能导致模型预测结果难以解释等问题。实际上，机器学习也是运用归纳思想从数据中发掘规律的方法论体系。因此，统计学的理论和方法可以弥补当前一些机器学习方法的不足。

在大部分高校的培养方案中，统计方法和机器学习方法往往是计算机科学、软件工程、统计学等专业所开设的多门课程中内容。这两类方法的侧重略有不同，统计方法方法往往追求的是所构造一个适合的区间较大的概率覆盖真值，而机器学习方法往往追求的是所输出的预测值是在样本空间内与真值的偏差最小。但是，这两类方法的共同点均是基于归纳的思想方法，利用数据构造模型，根据模型得到结论，发掘数据中隐藏的规律。因此，本课程融合常用的统计方法和经典的机器学习方法，从数据科学的角度，系统介绍如何根据不同的数据类型，构造合适的模型，并以此获得正确的结论。

这门课程是数据科学与大数据技术专业三年级下学期本科生的核心专业必修课。本课程主要囊括了方差分析、回归分析、变量选择、聚类方法、判别与分类方法、支持向量机、决策树、随机森林、人工神经网络、支持向量机和强化学习方法。通过本课程的学习，学生能够具备以下能力：1）能够掌握数据分析的一般过程；2）对于不同数据类型，能够选择合适模型分析数据；3）掌握运用适当的指标评判模型的优劣；4）能够解释模型的拟合或预测结果，并以此得出结论。本课程作为数据分析模块中的核心课程，为学生进一步学习后续课程打下基础。

二、课程目标

通过为本科生开设《统计方法与机器学习》这门课程，将达到如下目标：

目标1：了解统计学与机器学习中的方法论体系，熟练掌握统计方法与机器学习的一般过程（支撑毕业要求7）

目标2：理解统计学与机器学习中不同方法的在理论上的优势与局限，为分析问题并实现问题求解奠定理论基础，为将来从事科学研究和职业发展提供有力支撑（支撑毕业要求13、支撑毕业要求14）

目标3：在充分理解数据特点的基础上，针对不同的数据类型，能够选择适合数据建模方法，提升学生分析问题的能力（支撑毕业要求5）

目标4：能够运用如变量选择、交叉验证和正则化等方法改善模型预测或拟合效果，提高学生解决问题的能力（支撑毕业要求6）

目标5：深刻理解衡量模型拟合和预测效果的指标，依据不同的指标，多维度能够评价和解释模型的优劣，为进一步优化模型奠定基础（支撑毕业要求6）

目标6：熟练运用多种数据分析和机器学习开源软件和工具，能够正确解读模型的运行结果，并据此进行科学决策（支撑毕业要求10、支撑毕业要求12）

三、课程目标与毕业要求的对应关系

|  |  |  |
| --- | --- | --- |
| **毕业要求** | **指标点** | **课程目标** |
| 专业技能扎实 | 掌握数据科学与工程的基础知识，包括相关的计算机、统计与应用数学、信息系统的基础知识 | 目标4 |
| 掌握数据分析和机器学习的基本模型和算法 | 目标4、5 |
| 学科理念先进 | 深刻理解数据的获取、建模、管理、利用的全生命周期，深刻理解数据科学与工程相关技术发展与社会经济发展的关系 | 目标1 |
| 工程能力全面 | 了解大数据应用中需求分析、数据和应用建模、系统选型、应用设计、开发和实施的过程，具备合作进行系统 和应用研发能力 | 目标6 |
| 具备参与数据系统或数据应用设计、开发、运维工程所需的沟通交流与协作能力，掌握基本的工程管理知识与能力 |
| 研究能力突出 | 了解“数据科学与工程”学科领域，以及相关应用 领域的技术发展前沿 | 目标2 |
| 具有初步的从事数据科学与工程研究工作的科学训 练，具有从事相关学科科学研究、教学或工程开发的技术工 作的能力 |

三、教学内容与学时安排

（重难点请标注🞛）

第〇章 绪论（支撑课程目标1）

学时：3（理论）+2（实验）

1. 课程简介

2. 对高等数学、线性代数、概率论与数理统计等先修知识回顾与补充

**要求**：熟悉数学工具；

第一章 方差分析（支撑课程目标1、2、3、6）

学时：3（理论）+2（实验）

1. 单因素方差分析🞛

2. 多重比较

**要求：**掌握方差分析的步骤；理解多重比较的问题定义及Tukey方法的应用。

第二章 线性回归分析（支撑课程目标1、2、3、5、6）

学时：6（理论）+4（实验）

1. 多元线性回归🞛

**要求**：掌握线性模型的假设、模型结构、参数估计方法、参数检验方法、预测方法。

第三章 变量选择（支撑课程目标1、2、3、4、5、6）

学时：3（理论）+2（实验）

1. 过拟合和欠拟合🞛

2. 变量选择的评价指标

**要求**：理解过拟合与欠拟合对模型预测效果的影响；了解常用的变量选择的评价指标；

第四章 多重共线性及其解决方案（支撑课程目标1、2、3、5、6）

学时：9（理论）+6（实验）

1. 多重共线性的定义🞛

2. 主成分回归🞛

3. 岭回归🞛

**要求：**理解多重共线性的定义及常用量化的判断方法；理解主成分回归的一般步骤及其优势；理解岭回归的一般步骤及其优势。

第五章 聚类分析（支撑课程目标1、2、3、5、6）

学时：3（理论）+2（实验）

1. 相似度测量

2. 常用的聚类方法🞛

3. 聚类方法评价方式

**要求：**了解聚类的思想与常见的相似度测量方法；理解四种常用的聚类方法；了解聚类方法的评价方式。

第六章 统计学习方法概论（支撑课程目标1、2、4）

学时：3（理论）+2（实验）

1. 统计学习方法的定义和分类

2. 模型评估、正则化、泛化能力

**要求：**了解统计学习方法的定义与分类方法。理解模型评估、正则化、泛化能力等概念。

第七章 判别与分类方法（支撑课程目标1、2、3、5、6）

学时： 6（理论）+4（实验）

1. 感知机🞛

2. K近邻🞛

3. 朴素贝叶斯🞛

4. 逻辑斯蒂回归🞛

**要求：**掌握常用的判别与分类方法，并掌握分类模型的评价方法。

第八章 支持向量机（支撑课程目标1、2、3、5、6）

学时：6（理论）+4（实验）

1. 理论基础（经验风险最小化\*、最大间隔超平面\*、对偶问题与支持向量）🞛

2. 非线性支持向量机

3. 序贯最小化实现方法

**要求：**理解支持向量机的理论；了解非线性支持向量机的模型机理；了解序贯最小华实现方法。

第九章 决策树与集成学习（支撑课程目标1、2、3、5、6）

学时：6（理论）+4（实验）

1. 决策树🞛

2. 集成学习

**要求**：掌握决策树的构建、剪枝等过程；理解不同的集成算法。

第十章 隐马尔可夫模型（支撑课程目标1、2、3、5、6）

学时：3（理论）+2（实验）

1. 概率计算🞛

2. 学习算法

3. 预测方法

**要求**：掌握隐马尔可夫模型的学习过程与预测方法。

第十一章 计算学习理论（支撑课程目标1、2）

学时：3（理论）+2（实验）

1. PAC学习

2. VC维

**要求**：了解PAC学习理论与VC维等基本概念。

实验内容：

通过在公开数据集上完成特定目标任务，使学生逐步掌握使用机器学习解决问题的流程和方法，学会使用开源工具，并通过深入的研究逐步完善目标任务。实验的设计遵循从简单到复杂、难度系数逐渐增大、任务类型多样等原则。拟采用的项目内容包括但不限于以下几个：

实验1:析因试验分析

数据：饮食与减重实验数据集

描述：24个实验样本数据，1个因素(饮食方式)

任务：判断不同的饮食方式是否对前后体重的差异产生影响

知识点：方差分析、多重比较

实验2:多元线性回归分析

数据：糖尿病人数据集

描述：共442名糖尿病，10个特征变量，1个响应

任务：利用回归分析的方法建立合适的模型，分析哪些因素会影响糖尿病人一年后的病情发展。

知识点：回归分析、变量选择、多重共线性

实验3:模拟实验

数据：4类人造数据集

描述：给定构造人造数据集的模型

任务：利用聚类分析，验证在不同数据集上不同聚类方法的优劣

知识点：聚类分析

实验4： 基于UCI数据集练习分类算法的使用。

数据：UCI数据集上的分类任务数据

任务：任选一种分类算法，采用交叉验证法获得预测准确率。

知识点：分类、性能评估、模型选择和评估

实验5：新闻分类

数据：Twenty Newgroups数据

任务：预测新闻文档所属类别

知识点：逻辑回归、决策树

实验6：字符识别

数据集：MNIST手写识别数据集

任务：使用神经网络或深度学习识别字符

知识点：支持向量机

实验7：人脸识别

数据：CMU Machine Learning Faces

http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/faces.html

描述：20人，每人32张脸图(含表情)

任务1：使用机器学习进行人脸分类识别，给出识别准确率

任务2：使用聚类或分类算法发现表情相似的脸图

知识点：集成学习、降维

四、教学方法

本课程采用以讲授法为主的教学手段，并结合PPT等资料辅助教学。在课程中采用提问的方式引导学生共同思考，让学生进行头脑风暴，锻炼学生批判性思考的能力；同时，采用案例教学，推托具体数据分析案例，让学生深刻理解统计方法与机器学习方法的原理，并让学生掌握其在Python中的实现过程。

五、考核方式

本课程采用五级评分制，并通过过程性评价对学生进行考核。具体考核指标如下：

考核指标1:出勤与课堂表现，占10%。

考核指标2:统计方法的理论作业，占10%。

考核指标3:统计方法的实验作业，占10%。

考核指标4:机器学习的理论作业，占10%。

考核指标5:机器学习的项目作业，占10%。

考核指标6:期末考试，占50%。

六、推荐教材和参考资料

**（一）推荐教材**：

自编讲义

**（二）参考资料：**

1、《Design and Analysis of Experiments》（第8版），Douglas C. Montgomery编著，John Wiley & Sons, Inc.出版社，2013年。

2、《Regression Analysis by Example》（第5版），Samprit Chatterjee，Alis. Hadi编著，Wiley出版社，2012年。

3、《The Elements of Statistical Learning-Data Mining, Inference, and Prediction》（第2版），Trevor Hastie，Robert Tibshirani，Jerome Friedman编著，Springer出版社，2009年。

4、《统计学习方法》（第2版），李航编著，清华大学出版社，2019年。

5、《机器学习》，周志华编著，清华大学出版社，2016年。

6、《Pattern Recognition and Machine Learning》，Christopher M. Bishop编著，Springer出版社，2006年。

## 八、评分标准【请按照本门课程采用的课程考核方式选择下表之一填写】（具体分段可以根据实际情况调整）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **课程目标** | **评分标准** | | | | |
| **优（A）** | **良（B+）** | **中（B-）** | **及格（C）** | **不合格（F）** |
| 学生能够了解统计学与机器学习中的方法论体系，熟练掌握统计方法与机器学习的一般过程；能够理解统计学与机器学习中不同方法的在理论上的优势与局限，为分析问题并实现问题求解奠定理论基础；在充分理解数据特点的基础上，针对不同的数据类型，能够选择适合数据建模方法；能够运用如变量选择、交叉验证和正则化等方法改善模型预测或拟合效果；能够深刻理解衡量模型拟合和预测效果的指标，依据不同的指标，多维度能够评价和解释模型的优劣，为进一步优化模型奠定基础能够熟练运用多种数据分析和机器学习开源软件和工具，能够正确解读模型的运行结果，并据此进行科学决策。 | 学生能够全面地了解统计学与机器学习中的方法论体系，并熟练掌握统计方法与机器学习的一般过程；能够深刻地理解统计学与机器学习中不同方法的在理论上的优势与局限，为分析问题并实现问题求解奠定理论基础；在充分理解数据特点的基础上，针对不同的数据类型，能够选择适合数据建模方法；能够灵活地运用如变量选择、交叉验证和正则化等方法改善模型预测或拟合效果；能够深刻理解衡量模型拟合和预测效果的指标，依据不同的指标，多维度能够评价和解释模型的优劣，为进一步优化模型奠定基础；能够熟练运用多种数据分析和机器学习开源软件和工具，能够正确解读模型的运行结果，并据此进行科学决策。 | 学生能够较为全面地了解统计学与机器学习中的方法论体系，并较为熟练掌握统计方法与机器学习的一般过程；能够较为深刻地理解统计学与机器学习中不同方法的在理论上的优势与局限，为分析问题并实现问题求解奠定理论基础；在充分理解数据特点的基础上，针对不同的数据类型，能够选择较为适合数据建模方法；能够运用如变量选择、交叉验证和正则化等方法改善模型预测或拟合效果；能够较为深刻理解衡量模型拟合和预测效果的指标，依据不同的指标，多维度能够评价和解释模型的优劣，为进一步优化模型奠定基础；能够较为熟练运用多种数据分析和机器学习开源软件和工具，能够正确解读模型的运行结果，并据此进行科学决策。 | 学生能够了解统计学与机器学习中的方法论体系，并掌握统计方法与机器学习的一般过程；能够较为深刻地理解统计学与机器学习中不同方法的在理论上的优势与局限，为分析问题并实现问题求解奠定理论基础；在充分理解数据特点的基础上，针对不同的数据类型，能够选择较为合适的数据建模方法；能够运用如变量选择、交叉验证和正则化等方法改善模型预测或拟合效果；能够理解衡量模型拟合和预测效果的指标，依据一个或多个指标，多维度能够评价和解释模型的优劣；能够运用多种数据分析和机器学习开源软件和工具，能够较为正确解读模型的运行结果，并据此进行科学决策。 | 学生能够基本了解统计学与机器学习中的方法论体系，并理解统计方法与机器学习的一般过程；能够理解统计学与机器学习中不同方法的在理论上的一些优势与局限，为分析问题并实现问题求解奠定理论基础；在理解数据特点的基础上，针对不同的数据类型，能够选择较为合适的数据建模方法；能够运用如变量选择、交叉验证和正则化等方法改善模型预测或拟合效果；能够理解衡量模型拟合和预测效果的指标，依据单个指标，能够评价和解释模型的优劣；能够运用一种数据分析和机器学习开源软件和工具，能够解读模型的运行结果，并据此进行较为正确的决策。 | 学生无法了解统计学与机器学习中的方法论体系，或无法理解统计方法与机器学习的一般过程；无法理解统计学与机器学习中不同方法的在理论上的局限；针对不同的数据类型，无法使用数据建模方法；无法理解衡量模型拟合和预测效果的指标；无法运用数据分析和机器学习开源软件和工具，或无法解读模型的运行结果。 |