软件工程本科生《机器学习》课程教案（四）

讲解人：李济洪 (教授)、王瑞波 (讲师)

一、授课课题

线性回归基础（2）：多元线性回归

二、授课时间

2019 年 9 月 24 日星期二 8:00am-10:00am

三、课时安排

2 课时

四、授课类型

理论课

五、教材

加雷斯 • 詹姆斯, 丹妮拉 • 威滕, 等.《统计学习导论: 基于 R 应用》[M]. 机械工业出版社, 2015.

课程网站：`http://www-bcf.usc.edu/~gareth/ISL/`

作业网站：`https://github.com/nguyen-toan/ISLR`

六、对应章节

第三章第 3.2 节 -3.5 节

七、教学目标及要求

1. 掌握多元线性回归的基本形式、参数估计方法及预测方法；
2. 基本多元线性回归的 R 函数：‘lm’ 的基本用法；

八、教学重点

1. 多元线性回归的基本形式；
2. 多元线性回归参数估计的形式；
3. 多元线性回归中模型拟合准确性的评价；

九、教学难点

1. 多元线性回归中定性变量的处理方式；
2. 多元线性回归中潜在的若干问题；

十、教学方式

讲授

十一、教学手段

课件讲解 + 课间讨论

十二、教学过程

1. 上讲回顾（教学方式：讲授；时间：5 分钟；）

回顾上讲中线性回归算法的基本概念、算法形式、参数估计的准确性用于评价模型准确性的两个准则。

特别是，要回顾一元线性回归算法中的基本假设：①iid 假设；②线性假设；③随机性与预测变量无关的假设。

2. 引入新课（教学方式：讲授；时间：10 分钟；）

基于广告数据的例子，让学生思考当多个预测变量同时影响响应变量时，该如何建立模型？

可选的方案有两种：

1. 基于一元线性回归算法，每一个预测变量与响应变量均建立一个模型。然后，基于多个一元线性回归模型，对未来观测进行预测。
   - 这种方法的**问题**在于：在对未来观测进行预测时，如何融合多个模型的预测值，形成一个统一的预测值？其关键在于由于数据相同，是的多个预测值存在相关性，如何分析这些相关性对于预测值的影响？

- 要强调这种方案不是不可取，本讲只是从另外一种方案入手。

2. 建立多元线性回归算法。

给出广告数据集上多元线性回归算法的具体形式，及其蕴涵的基本概念。

3. 介绍多元线性回归算法的理论知识（教学方式：讲授；时间：30 分钟）

给出一般的多元线性回归算法的基本形式和相关概念。介绍该形式与一元线性回归算法的区别之处。

给定多元线性回归与正态分布之间的关系。

给出多元线性回归的预测形式，并介绍其使用方法。

具体为 5 大问题：

1. 给定训练数据 D_n，如何估计多元线性回归算法的参数 $\beta_1, \ldots, \beta_p$？
2. 预测变量 $X_1, \ldots, X_p$ 中是否至少有一个预测变量可以用来预测响应变量？
3. 所有预测变量都有助于解释响应变量 Y 吗？或仅仅是其一个子集对预测有用？
4. 模型对数据的拟合程度如何？
5. 给定一组预测变量的值，预测的精度如何？

4. 介绍定性预测变量的处理方法（教学方式：讲授；时间：10 分钟）

引入信用数据 (Credit) 数据，让学生清楚定性数据的表示方法，进而引入 “水平” 的概念，并介绍二水平、多水平之间的区别。

4.1. 介绍二水平定性预测变量的处理方法

针对二水平的定性变量，介绍哑变量的处理技术。

比较多种哑变量的形式化方法，并用实例说明它们之间的等价性。

4.2. 介绍多水平定性预测变量的处理方法

以信用数据集中的 ethnicity 变量为例，介绍多水平定性变量处理方法。

强调哑变量个数与预测变量水平数之间的关系。

5. 介绍交互作用及非线性关系在线性回归算法中的建模（教学方法：讲授；时间：20 分钟）

5.1 预测变量的交互作用对回归算法的影响

以工厂的生产力为例，介绍预测变量的交互作用对响应变量的影响，进而给出含交互作用的回归算法形式；

在广告数据集及信用数据集上，给出含有交互作用的回归模型实例。

5.2 非线性关系对回归算法中的影响

以 Auto 数据集为例，分析预测变量 ‘horsepower’ 与响应变量 ‘油耗’ 之间的关系。

6. 介绍多元线性回归中的一些潜在问题（教学方法：讲授；时间：10 分钟）

本节内容简要介绍，每种问题点到即止。

主要介绍如下几点问题：

- 数据非线性
- 误差项自相关
- 误差项方差非恒定
- 离群点
- 高杠杆点
- 共线性

7. 介绍多元线性回归算法的 R 实现（教学方法：讲授；时间：15 分钟）

主要讲解多元线性回归情形下，‘lm’ 函数、‘predict’ 函数及 ‘summary’ 函数的使用方法和输出格式。

8. 本讲总结（教学方法：讲授；时间：5 分钟）

对本讲内容进行总结：

- 多元线性回归的形式；
- 参数估计；
- 参数准确性的相关检验；
- 定性变量的处理方式；
- 非线性关系的处理方式；

- 算法潜在的一些问题；

十三、作业

1. **推导**：多元线性回归算法参数的最小二乘估计的形式。
2. **思考**：在处理多水平的定性预测变量时，为何不将其形式化为取多值得单个变量，而要形式化为多个取二值的哑变量？
3. **应用**：分析清楚 Letter 数据集的预测变量和响应变量，及任务定义。进而，使用 R 语言的多元线性回归算法对 Letter 数据集进行建模。
   - Letter 数据集地址：`http://archive.ics.uci.edu/ml/datasets/letter+recognition`

十四、参考资料