软件工程本科生《机器学习》课程教案（三）

讲解人：李济洪 (教授)、王瑞波 (讲师)

一、授课课题

一元线性回归及均值回归

二、授课时间

2019 年 9 月 17 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 3.1 节

七、教学目标及要求

1. 深入掌握一元线性回归算法的形式，并掌握参数估计方法；
2. 基本掌握一元线性回归算法的 R 实现方法；

八、教学重点

1. 一元线性回归的基本形式；
2. 一元线性回归参数估计的形式；
3. 一元线性回归与均值回归之间的关系；

九、教学难点

1. 一元线性回归参数估计的计算过程；
2. 参数估计的准确性与算法预测准确性间的关系；

十、教学方式

讲授与演示结合

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；）

回顾统计学习算法的一般形式、一些基本概念、建模任务及算法的估计方法；引导学生梳理统计学习的前提及过程。具体地，前提是训练数据（假定为表格数据）已经给定。过程分为三步：

1. 选取算法族，并抽象出算法形式（限定在参数算法）；
2. 基于训练数据，估计算法参数；
3. 基于算法参数估计，根据算法参数形式，对未来的观测进行预测。

2. 导入新课（教学方式：讲授；时间：10 分钟；）

使用广告（Advertising）数据集为例引入新课。首先，让学生回顾广告数据集中的预测变量（电视、广播及报纸媒体的预算）和响应变量（销量）。进而，仅考虑单预测变量（电视媒体），引导学生考虑如下四个问题：

- 电视媒体的预算与产品销量之间有关系么？
- 它们之间是什么关系？关系有多强？
- 如何精确地估计电视媒体的预算与产品销量之间的关系？

- 给定未来电视媒体的预算，如何预测未来销量？该预测的精度如何？

在上述四个问题中，引导学生从线性关系入手，形式化电视媒体预算与产品销量之间的线性回归算法。

3. 介绍一元线性算法的核心概念（教学方式：讲授；时间：10 分钟）

3.1 介绍一元线性回归算法的基本形式

将课程导入例子中的电视媒体预算及销量分别抽取成一般的统计学习概念预测变量及响应变量。然后，给出线性回归算法的一般形式：

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

其中，$\epsilon \sim N(0, \sigma^2)$。

线性回归算法的另一种等价的表示为：

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2). \tag{2}$$

显然，$E[Y|X] = \beta_0 + \beta_1 X$。也就是说，一元线性回归算法是使用一条直线来拟合响应变量 Y 的均值。

3.2 介绍一元线性回归算法的基本概念

预测变量为 X、响应变量为 Y、算法参数为 β_0 和 β_1、误差项为 ϵ。

3.3 以图形的方式讲解线性回归方程中正态分布的含义

- 简单回顾正态分布，并推荐学生课下阅读材料《正态分布的前世今生》(`https://songshuhui.net/archives/76501`)；
- 给出一元线性回归模型中正态分布的随机噪声的图解；

3.4 介绍无截距一元线性回归算法

介绍无截距项的一元线性回归算法形式。具体如下：

$$Y = \beta_1 X + \epsilon. \tag{3}$$

介绍无截距项线性回归算法与一般线性回归算法间的关系，即：无截距项线性回归算法为线性回归算法在 $\beta_0 = 0$ 时的一种特殊情况。进而，介绍 $\beta_0 = 0$ 是一种人为对

数据集的假设。该假设意味着，当 $X = 0$ 时，响应变量 Y 的均值也应为 0。该假设并不总是成立。

4. 介绍一元线性回归算法参数估计方法（教学方式：讲授；时间：15 分钟；）

4.1 介绍参数估计的优化准则

- 从算法与数据集的接近程度的角度，引入算法优化的基本准则：**残差平方和**。
- 从一般的统计学习角度介绍优化的一般准则：**极大似然准则**。
  - 回归“似然”的概念。
  - 介绍“极大似然”的含义。
- 介绍“最小化残差平方和”与“极大似然”两个准则之间的等价关系。

4.2 介绍参数的最小二乘估计

- 明确写出优化问题的形式化表示；
- 给出该优化问题的解（鼓励学生课下自己动手求解）；
- 介绍 β_0 和 β_1 的含义；
- 引入“最小二乘估计”的概念，并给出该术语名的含义；

4.3 以广告例子为例，深入介绍最小二乘估计的含义

给出广告例子的一元回归的估计及其图象，生动分析最小二乘估计的物理含义。

5. 介绍参数估计的准确性（教学方式：讲授；时间：10 分钟）

5.1 介绍总体回归直线与最小二乘线

强调总体回归直线是建立在数据总体之上的概念，因此通常是不知道的理想值，只有一条。最小二乘线和给定的数据集 D_n 有关。即，给定 D_n 的不同取值，可以获得不同的最小二乘线。我们希望最小二乘线可以很好地“靠近”总体回归直线。换句话说，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 要很好地“靠近”其理论值 β_0 和 β_1。

5.2 介绍参数估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的性质

估计值是否可以“靠近”理论值，可通过估计值的均方误差准则来刻画。估计值的均方误差可以分解成偏差的平方和方差之和。因此，估计值的质量可反映在参数估计的偏差和方差上。

- 从偏差角度，通过示例图像直观地讲解参数估计的无偏性。

- 从方差角度，给出参数估计的方差。进而引入标准误的概念。

5.3 介绍参数估计的置信区间

给出每个参数的置信区间及其对应的解释。

5.4 介绍预测变量和响应变量是否有关的检验

- 引入假设检验问题，刻画预测变量和响应变量是否相关的问题。
- 引入 t 检验统计量。
- 给出 p 值的概念及对应的决策。

进而，以教材中表 3-1 为例，介绍参数估计的相关含义。

6. 介绍模型评价的准确性（教学方式：讲授；时间：15 分钟）

首先，回顾模型评价的两个角度：预测和推断。其中，推断主要关注算法的解释性。具体地，当前得到的一元线性回归模型可多大程度上解释（拟合）训练数据？

然后，引入两个准则：**残差标准误**和**拟合优度 R^2 统计量**。对这两个准则分别介绍。

7. 课堂实验：一元回归算法的应用（教学方式：讲授；时间：30 分钟）

7.1 引入实验示例问题及数据集

仍以广告数据集为例，介绍实验所要解决的问题。

给出数据集文件，让学生观察数据集，介绍数据集的存放方式，并给出环境变量的概念来制定数据集的存放路径。

讲解 RCRAN 标准包的目录结构，Rstudio 的项目、工作目录的读取和设置等基本使用方法。

7.2 机器学习建模流程及代码规划

讲解机器学习（实验室）建模的工作流程：

1. 读取实验数据；
2. 给定机器学习算法的超参数；
3. 训练机器学习模型；
4. 评估机器学习模型的解释性（推断）；
5. 使用机器学习模型进行预测；
6. 评估机器学习的预测性能（预测）；

根据上述流程规划实验代码目录结构；

7.3 实现数据文件的读取和格式化

首先，向学生讲解清楚为了让代码适用于不同的数据集，因对数据集在内存中的存储格式进行规范化。本节课使用的数据存储规范为：使用 data.frame 存储数据，最后一列为回归值、前面的列对应于不同的预测变量。

7.4 实现一元线性回归算法文件的代码结构

- 训练函数: train()
- 预测函数：predict()
- 模型评估函数：assess()

进一步，讲解各个函数的参数结构。

7.5 实现一元线性回归算法的训练、测试和评估过程

讲解 R 中 ‘lm‘函数，‘predict‘函数及 ‘summary‘函数的使用方法。特别是，结合 R 中的帮助文件对这些函数的参数结构进行讲解。

7.6 实现整个算法的建模流程

给出一个标准 R 文件，执行了第 7.2 节中各个流程的标准函数。

十三、作业

1. 基于残差平方和准则，给出一元线性回归算法的参数 β_0 和 β_1 的估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的详细求解过程。
2. 思考：从预测角度，应该如何评价一元线性回归算法的准确性？请给出具体思路。

十四、参考资料