软件工程本科生《机器学习》课程教案（五）

讲解人：李济洪 (教授)、王瑞波 (讲师)

一、授课课题

算法预测性能估计方法

二、授课时间

2019 年 10 月 8 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

第五章

七、教学目标及要求

1. 掌握统计模型的训练错误率和测试错误率的区别，及其估计方法；
2. 掌握 Hold-out 验证方法；
3. 掌握常用的交叉验证方法；
   - 留一交叉验证方法

- Repeated Learning-testing 方法
- K 折交叉验证方法
- $m \times 2$ 交叉验证方法
- 块正则化 $m \times 2$ 交叉验证方法
- 块正则化 Repeated Learning-testing 交叉验证方法
- 自助（bootstrap）法

4. 使用 R 语言实现模型预测性能的各种验证方法；

八、教学重点

1. 掌握数据切分的基本原则；
2. 掌握各种交叉验证方法的切分方法；

九、教学难点

1. “块正则化”的基本思想；
2. “块正则化”交叉验证的数据切分方法；

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；）

回顾多元线性回归算法的形式、参数估计方法、参数估计的准确性以及模型拟合准确性的评价方法。

回顾哑变量处理方法、非线性关系处理方法以及线性回归算法中存在的一些问题。

2. 引入新课（教学方式：讲授；时间：10 分钟；）

1. 介绍 Auto 数据集及相应的任务；

- 介绍 Auto 数据集中各个预测变量的含义、每一条观测对应的对象以及预测变量的含义；
- 本讲仅关心马力（horsepower）与油耗（mpg）之间的关系：**给定一种车型的马力，如何预测该车的油耗？**
- 介绍对应的多项式回归算法及多项式阶数的选择问题。

2. 从**推断**和**预测**两个角度来给出机器学习领域关心的两个问题，并以 Auto 数据集介绍两个问题的含义。
3. 介绍基于训练数据集和未来观测计算的算法的性能之间的差别。
4. 介绍算法预测性能估计与多项式算法阶数选择之间的关系。

3. 介绍算法预测性能估计的一些基本概念及记号（教学方式：讲授；时间：10 分钟；）

介绍如下一些基本概念。

1. 数据分布：$\mathcal{F}$；
2. 数据集 $D_n = \{z_i : (x_i, y_i)\}_{i=1}^n$；
3. 测试样本 z；
4. 机器学习算法：$\mathcal{A}$；
5. 平方损失函数：$L(y, \hat{y}) = (y - \hat{(y)})^2$；
6. 泛化误差：$\mu = E[L(y, \hat{y})]$；

进一步，介绍计算算法的泛化误差的难处：无法获得数据总体，只能获得有限大小的数据集 D_n。

要解决的问题：**如何基于有限大小的数据集 D_n，准确地估计出算法 $\mathcal{A}$ 的预测性能？**

解决问题的一个基本思路：**使用额外的数据集，来估计算法的预测性能**。进而，引入算法预测性能的估计符号 $\hat{\mu}$，并引入评估算法预测性能估计的一个准则：均方误差。

4. 介绍 Hold-out 验证方法（教学方式：讲授；时间：15 分钟；）

1. 介绍 hold-out 验证方法的数据切分方式、及相应的概念：训练集、验证集（保留集）。
2. 介绍 hold-out 验证方法中涉及的一些问题：
   (a) 训练集大小、测试集大小应该如何选择？
   (b) 数据集应该如何切分？

3. 介绍泛化误差的 Hold-out 估计的形式及含义。
4. 以 Auto 数据集为例，介绍 hold-out 验证估计的优点及缺点。

5. 介绍交叉验证方法（教学方式：讲授；时间：45 分钟；）

1. 解决 Hold-out 缺点的一个方法时：多次切分，取平均。
2. 介绍“平均”的思想和优点。

进而，详细介绍如下几种交叉验证切分方法。

5.1 留一交叉验证（Leave-one-out Cross-validation）

- 介绍留一交叉验证的数据切分特点、给出泛化误差的留一交叉验证估计。
- 介绍留一交叉验证在线性回归时的快速计算方法。
- 介绍留一交叉验证的优缺点：渐进无偏、方差大，计算量大。

5.2 Repeated Learning-testing 交叉验证

- 介绍将验证集大小增大的优点：压缩方差。
- 介绍穷尽交叉验证（Exhaustive Cross-validation）。
- 介绍穷尽交叉验证的替代方法：Repeated Learning-testing。
  - 给出 Repeated Learning-testing 交叉验证的数据切分特点。
  - 给出泛化误差的 Repeated Learning-testing 交叉验证估计。
  - 给出 Repeated Learning-testing 交叉验证的控制参数：训练集大小、切分次数。
  - 给出 Repeated Learning-testing 交叉验证估计的优缺点。

5.3 K 折交叉验证

- 介绍 K 折交叉验证的数据切分方式及控制参数。
- 介绍泛化 K 交叉验证估计的形式。
- 介绍 K 折交叉验证估计的优良性质。
- 介绍 K 折交叉验证估计的方差偏差权衡问题。
- 介绍 K 折交叉验证估计的缺点。

5.4 $m \times K$ 交叉验证

- 介绍 $m \times K$ 交叉验证的数据切分方式及控制参数。
- 介绍 $m \times K$ 交叉验证估计的形式。
- 介绍 $m \times K$ 交叉验证的优良性质。

- 介绍 $m \times K$ 交叉验证的偏差和方差。
- 介绍 $m \times K$ 交叉验证估计的缺点。

5.6 块正则化 $m \times K$ 交叉验证

- 介绍正则化 $m \times K$ 交叉验证的数据切分方式及控制参数。
- 介绍正则化 $m \times K$ 交叉验证估计的形式。
- 介绍正则化 $m \times K$ 交叉验证的优良性质。
- 介绍正则化 $m \times K$ 交叉验证的偏差和方差。
- 介绍正则化 $m \times K$ 交叉验证估计的缺点。

5.5 块正则化 Repeated Learning-testing 交叉验证

- 介绍正则化 Repeated Learning-testing 交叉验证的数据切分方式及控制参数。
- 介绍正则化 Repeated Learning-testing 交叉验证估计的形式。
- 介绍正则化 Repeated Learning-testing 交叉验证的优良性质。
- 介绍正则化 Repeated Learning-testing 交叉验证的偏差和方差。
- 介绍正则化 Repeated Learning-testing 交叉验证估计的缺点。

5.7 进一步探讨交叉验证切分数据的原则、特点和待解决的问题

- 给出数据切分的三个原则。
- 交叉验证切分的特点在于：“切”和“反复”。
- 待解决的问题：如何找到 MSE 更小的泛化误差估计？切分次数受限、训练集大小和验证集大小受限。

6. 介绍自助法（教学方式：讲授；时间：10 分钟；）

- 对本节简要介绍。
- 重点突出：自助法与交叉验证方法的不同。
- 给出 Bootstrap 的数据抽样方式。
- 给出泛化误差的 Bootstrap 估计。
- 简要介绍 Bootstrap 方法的优缺点。

7. 本讲总结（教学方式：讲授；时间：5 分钟；）

- 总结算法预测性能估计要解决的问题。
- 简要回顾数据切分的原则。

- 回归交叉验证方法；
- 回顾 Bootstrap 方法。

十三、作业

1. 【计算】两个对半切分的 Hold-out 验证中，训练集重叠样本个数等于 $n/4$ 的概率是多大？
2. 【思考】对于含有 J 次切分的 Repeated Learning-testing，给定重叠样本个数 k，如何切分数据，可以使 J 个训练集间的重叠样本个数均为 k？
3. 【思考】对于数据集 D_n，又放回地抽取出两组数据集 $D_n^{(1)}$ 和 $D_n^{(2)}$，请问 $D_n^{(1)}$ 和 $D_n^{(2)}$ 中重叠样本个数是如何分布的？

十四、参考资料

1. Ruibo Wang, Yu Wang, Jihong Li, Xingli Yang, and Jing Yang. 2017a. Block-regularized $m\times 2$ cross-validated estimator of the generalization error. Neural Computation, 29(2):519–554.
2. Ruibo Wang, Jihong Li, Xingli Yang, and Jing Yang. 2019. Block-regularized repeated learning-testing for estimating generalization error. Information Sciences, 477:246–264.