机器学习

杨雨禾

Weekly Seminar

2025 年 9 月 30 日

1.1 机器学习的基本定义

- 将训练数据输入计算机，让计算机通过学习和训练，然后利用训练后的算法进行结果的预测或者分类。
- 基本任务：分类（离散数据）与预测（连续数据）
- 涉及三个关键概念：
  - 数据
  - 模型
  - 学习

1.2 机器学习的分类

- 监督学习（Supervised Learning）：**分类与回归**
  已知数据 X 和标签 Y，学习一个函数 f 使得 $f(X) \approx Y$。
  - 分类任务：结果是离散类别。
    示例：判断贷款是否违约（是/否）、识别邮件是否为垃圾邮件
  - 回归任务：预测结果是连续数值。
    示例：预测股价、房价、GDP 增长率
- 无监督学习（Unsupervised Learning）：**聚类与降维**
  已知数据 X，无标签 Y。主要任务是发现数据的潜在结构。
  - 聚类：聚类的目标是把没有标签的数据分组，使得同一组数据的相似度高，不同组之间差异大。
    示例：K-means 聚类细分客户群体。
  - 降维：将高维数据映射到低维空间，同时保留数据的主要结构和信息。
    示例：主成分分析（PCA）比较学生的学习水平差异。
- 半监督学习：已知少量的标记数据与大量的未标记数据
- 强化学习（Reinforcement Learning）：智能体与环境交互

2.1 关键模块

- 数据：特征 & 标签
  Tips: 特征、属性、自变量、预测变量同义；标签、因变量、被预测变量同义
- 模型：数据到输出结果的映射函数 $f: X \to Y$。
- 学习：参数估计。通过数据优化模型参数，使模型能更好地拟合训练数据并推广到新数据。

2.2 参数估计的目标

- 设 $x_n \in \mathrm{R}^D$ ，因变量 $y_n \in R$，$\mathrm{n} = 1, 2, \cdots, \mathrm{N}$
- 估计一个预测器 $\mathrm{f}(\cdot, \theta) : R^D \to R$ (θ 为参数)
- 期望能够找到一个 θ^* $(\mathrm{n} = 1, 2, \cdots, \mathrm{N})$，使得损失函数最小：

$$f(x_n, \theta^*) \approx y_n$$

- 假设数据集 $(x_1, y_1), \ldots, (x_N, y_N)$ 是独立同分布的
- 平均损失：

$$R_{erm}(f, X, y) = \frac{1}{N} \sum_{n=1}^{D} \ell\left(y_n, \hat{y}_n\right)$$

- 这种学习策略被称为经验风险最小化 (ERM)
- 最小二乘法：$\min_\theta \frac{1}{N} \sum_{n=1}^{N} \left(y_n - \theta^T x_n\right)^2$

2.2 参数估计的目标

- ERM 的缺点：容易过拟合
- 结构风险最小化（SRM）：经验风险与模型复杂度之间取得平衡

$$R_{srm}(f) = R_{erm}(f) + \lambda \cdot \Omega(f)$$

  - $R_{erm}(f)$：经验风险（训练误差）
  - $\Omega(f)$：模型复杂度的惩罚项（regularization term），如参数范数
  - λ：平衡参数，决定经验风险与复杂度约束的权重

2.3 正则化

- L1 正则化（LASSO）：惩罚项为参数绝对值之和

$$\Omega(\theta) = \sum_j |\theta_j|$$

  - 产生稀疏解（部分参数为 0），用于特征选择。
- L2 正则化（Ridge）：惩罚项为参数平方和

$$\Omega(\theta) = \sum_j \theta_j^2$$

  - 不会使参数为 0，更适合参数相关性高的情况。

2.4 模型评估

- 过拟合：在学习过程中，不仅学到了数据的普遍特征，而且学习到了训练数据集的独特特征
- 欠拟合：在学习过程中，没有学到数据的普遍特征

(拟) 第 4 讲: 模型评估、训练技巧与可解释性大纲

- 模型训练与评估:
  - 数据划分：训练、验证、测试
  - 模型评估指标：AUC、F1、混淆矩阵
- 模型可解释性:
  - 线性模型与 LASSO：解释简洁
  - 树模型路径可视化

计划

1. 机器学习基础框架与关键概念
2. 监督学习与集成学习
   - 惩罚项线性回归
   - 分类问题与树模型
   - 集成学习概况
   - 实操环节
3. 神经网络与深度学习
   - 神经网络基础
   - 深度学习结构演进
   - 实操环节
4. 模型评估、训练技巧与可解释性
   - 模型训练与评估
   - 模型可解释性
   - 实操环节

计划

1. 多模态学习与经济金融应用
   - 多模态学习深度模型构建
   - 多模态在经济金融中的应用场景
   - 实操环节
2. 机器学习与因果推断融合
   - 机器学习与因果推断结合
   - 集成学习之随机森林
   - Causal Forest 与异质性估计
   - 复刻文献

- 共享 Github 仓库：https://github.com/arlionn/weekshare
- 预读文献池