《数据挖掘与机器学习》课程导论

汪小圈

2025-02-24

内容安排

- 机器学习的基本概念
- 课程介绍
- 机器学习的分类
- 机器学习的一般步骤

机器学习 vs 统计

Machine Learning	Statistics
Network, graphs	Model
Weights	Parameters
Learning	Fitting
Generalization	Test-set performance
Supervised learning	Regression / classification
Unsupervised learning	Density estimation, clustering
Large grant = $1,000,000	Large grant = $50,000
Nice place to have a meeting: Snowbird, Utah, French Alps	Nice place to have a meeting: Las Vegas in August

机器学习 vs 统计术语

机器学习

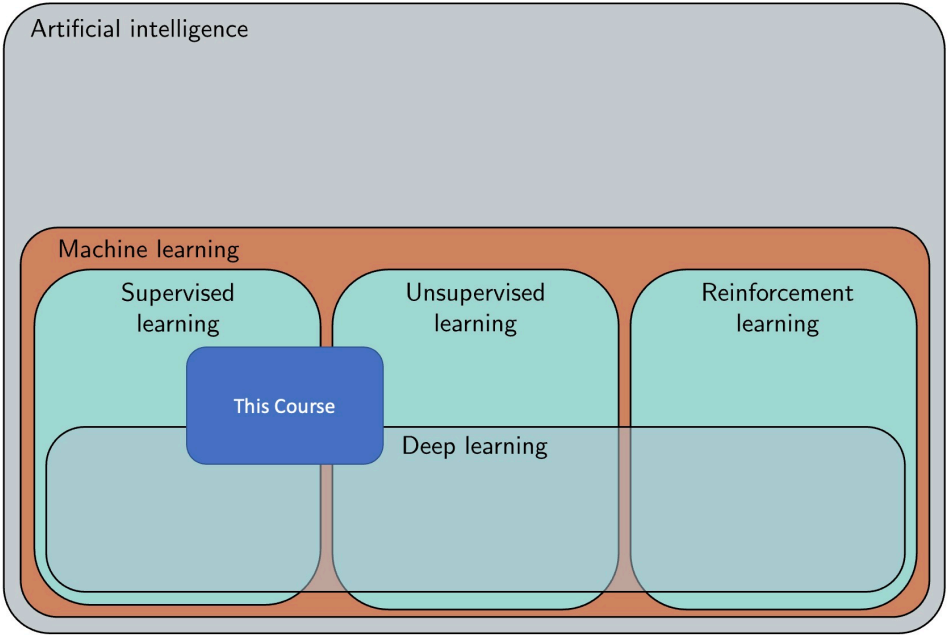
- 训练、学习
- 学习器、算法
- 特征、输入
- 目标、标记、输出
- 样例、示例

统计

- 估计
- 模型、估计量
- 协变量、解释变量、自变量、预测变量
- 因变量
- 数据点、观测数据

什么是机器学习

- **定义与核心思想**：从数据中学习模式，进行预测和决策。
  - 机器学习使计算机无需显式编程即可学习。
  - 核心在于从经验（数据）中自动改进。
- **与传统编程的区别**：
  - 传统编程：明确规则，处理确定性问题。
  - 机器学习：从数据中发现规则，处理不确定性和复杂性问题。
- **主要类型**：
  - 监督学习 (Supervised Learning)：有标签数据，预测结果。
  - 无监督学习 (Unsupervised Learning)：无标签数据，发现模式。
  - 强化学习 (Reinforcement Learning)：通过试错学习策略，优化。


Artificial intelligence
Machine learning
Supervised learning
Unsupervised learning
Reinforcement learning
This Course
Deep learning

课程目标与考核方式

- **课程目标**
  - **理论**：掌握机器学习的基本理论和常用算法。
  - **实践**：熟悉 Python 数据分析和机器学习工具库 (Pandas, Scikit-learn 等)。
  - **应用**：能够运用机器学习方法解决金融领域的实际问题。
  - **效率**：培养利用 AI 辅助工具进行高效编程和问题解决的能力 (Cursor, Copilot)。
- **考核方式**
  - **综合项目 1** (25%)：借贷违约风险评估模型
    - 特征工程 + 模型评估报告 + 复现代码
  - **综合项目 2** (25%)：股票价格预测/财报文本分析
    - 机器学习部分 + 金融预测部分 + 复现代码
  - **期末考试** (50%)：闭卷笔试
    - 理论 80% + 案例分析 20%

综合项目 1：借贷违约风险评估模型

- **项目背景**
  - P2P 借贷平台风险管理至关重要，降低坏账率是平台生存和发展的关键。
- **项目目标**
  - 利用 Lending Club 数据集，构建机器学习模型预测借款人是否违约。
- **Lending Club 数据集**
  - 来源：美国 P2P 借贷平台 Lending Club 公开数据。
  - 规模：包含大量借款人的个人和贷款信息。
  - 特征字段：借款人特征 (年龄、收入、信用评分等)，贷款特征 (贷款金额、利率、期限等)。
- **评价指标**
  - 准确率 (Accuracy), 精确率 (Precision), 召回率 (Recall), F1-score, AUC 等。
  - 强调根据业务场景选择合适的评估指标 (例如，关注坏账率，则 Recall 更重要)。

综合项目 2A：股票价格预测

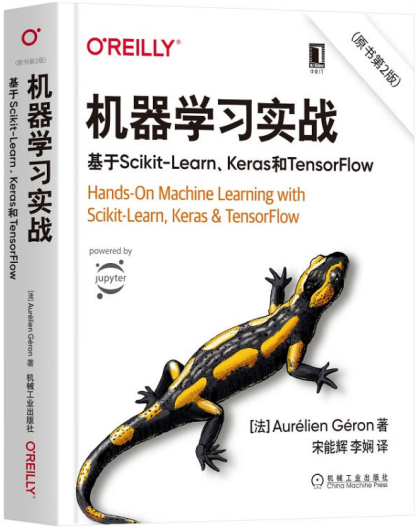
- **项目背景**
  - 量化投资依赖于对未来市场走势的预测。
  - 机器学习为股票价格预测提供了新的工具和方法。
- **项目目标**
  - 利用历史股票数据，构建机器学习模型预测未来股票价格走势。
  - 掌握时间序列数据处理和机器学习模型应用。
- **数据获取与特点**
  - 数据来源：Tushare 金融数据接口, CSMAR, WIND 等。
  - 时间序列数据特点：时间依赖性，自相关性，趋势性，季节性等。
- **评价指标**
  - 均方根误差 (RMSE), 平均绝对误差 (MAE), R-squared 等。
  - 根据预测目标选择合适的指标 (例如，关注预测精度，则 RMSE 更常用)。

综合项目 2B：财报文本分析

- **项目背景**
  - 财务报告包含大量非结构化文本信息。
  - 文本信息蕴含重要的公司经营风险信息。
- **项目目标**
  - 利用上市公司财报数据，进行文本分析，提取关键信息。
  - 运用机器学习方法辅助金融预测，例如：风险预警。
- **上市公司财报数据获取**
  - 数据来源：巨潮资讯网。
- **评价指标**
  - 准确率, 精确率, 召回率, F1-score 等。
  - 根据公司经营风险的定义选择合适的指标。

参考书目


机器学习
周志华 著
清华大学出版社
周志华 著
MACHINE
LEARNING
机器学习
清华大学出版社


O'REILLY
(原书第2版)
机器学习实战
基于Scikit-Learn、Keras和TensorFlow
Hands-On Machine Learning with
Scikit-Learn, Keras & TensorFlow
powered by
jupyter
[法] Aurélien Géron 著
宋能辉 李娴 译
机械工业出版社
China Machine Press
(原书第2版)
机器学习实战
基于Scikit-Learn、Keras和TensorFlow
[法] Aurélien Géron 著
机械工业出版社
O'REILLY

机器学习的一般步骤

数据搜集 → 数据清洗 → 特征工程 → 数据建模

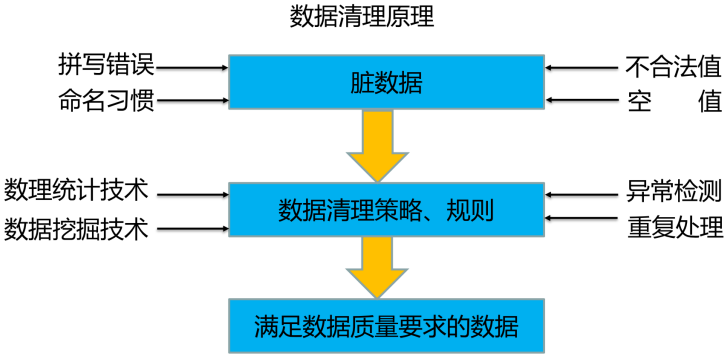
- 网络下载
- 网络爬虫
- 数据库读取
- 开放数据
- ……

- 数据清理和格式化
- 探索性数据分析(EDA)

- 特征工程
- 特征选择

- 基于性能指标比较几种机器学习模型
- 对最佳模型执行超参数调整
- 在测试集上评估最佳模型
- 解释模型结果
- 得出结论

数据清洗


数据清理原理
拼写错误
命名习惯
脏数据
不合法值
空　　值
数理统计技术
数据挖掘技术
数据清理策略、规则
异常检测
重复处理
满足数据质量要求的数据

数据预处理

输入数据标准化和数据转换能大大提高机器学习模型的性能

- 缺失值处理
  - 删除
  - 填补：
    - 单变量：使用均值、中位数、众数、或常数填补，抑或采用聚类填充等
    - 多变量：使用多重填补、回归、随机森林等方法
- 标准化
  - 数值型变量，使自变量均值为 0、方差为 1
  - 分类型变量，先转换为哑变量，再标准化

探索性数据分析

- 目的：找到异常、模式、趋势或关系
- 方法多样：以画图和描述性统计分析为主
  - 单变量：直方图、箱线图、描述性统计量
  - 多变量：相关性分析、Pairs Plot
  - 缺失值分析：缺失情况、填补与否、填补方法

特征工程和特征选择

- 特征工程：从原始数据中提取或创建新特征的过程
  - 主要方法：离散型变量处理、分箱/分区、交叉特征、特征缩放、特征提取
- 特征选择：选择数据中最相关的特征的过程
  - 减少特征数量、降维，使模型泛化能力更强，减少过拟合
  - 增强对特征和特征值之间的理解
  - 主要方法：去除变化小的特征、去除共线特征、去除重复特征、主成分分析
- 重要性：数据和特征决定了机器学习的上限，而模型和算法只能逼近这个上限而已

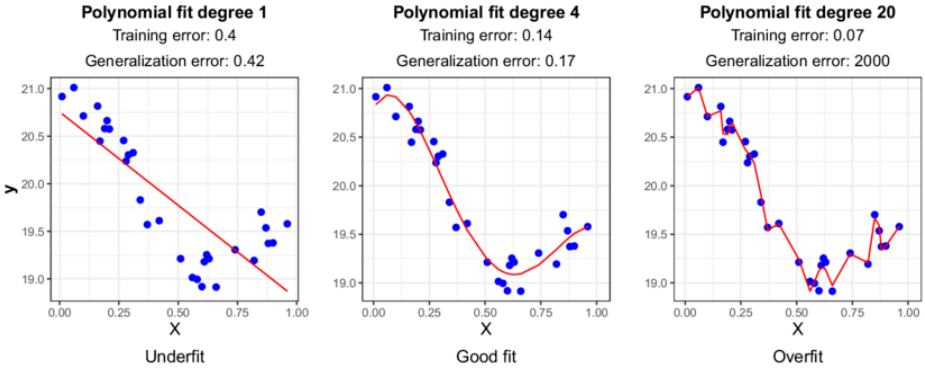
模型优化问题

- 误差 (error)：模型的实际预测输出 $f(x)$ 与样本的真实输出 y 之间的差异
- 经验误差 (empirical error) / 训练误差 (training error)：模型在训练样本上的误差
- 泛化误差 (generalization error) / 测试误差 (test error)：模型在新样本上的误差
- 模型最优化问题是找到泛化误差最小的模型
- 由于新样本未知，模型最优化问题往往是基于训练样本的经验误差最小化

$$\min_f \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

- 其中 $L(.)$ 为损失函数

欠拟合 (underfitting) 与过拟合 (overfitting) 问题



- 欠拟合比较容易克服
- 过拟合是机器学习面临的关键障碍

原理：泛化误差分解

- 假设数据来自于模型 $Y = f(X) + \epsilon$，其中 $E(\epsilon) = 0$，$Var(\epsilon) = \sigma^2$
- 该模型在新数据点 x_0 处的期望预测误差可以分解如下：

$$
\begin{aligned}
Err(x_0) &= E[(Y - \hat{f}(x_0))^2 | X = x_0] \\
&= E[(f(x_0) + \epsilon - \hat{f}(x_0) + E\hat{f}(x_0) - E\hat{f}(x_0))^2 | X = x_0] \\
&= \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \\
&= \sigma_\epsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0))
\end{aligned}
$$

- 第一项是误差的下界
- 第二项为偏差项，即期望预测 $E\hat{f}(x_0)$ 与真实均值 $f(x_0)$ 的偏离程度，刻画了模型本身的拟合能力
- 第三项为方差项，刻画了数据样本变动导致的模型性能的变化

偏差与方差的权衡

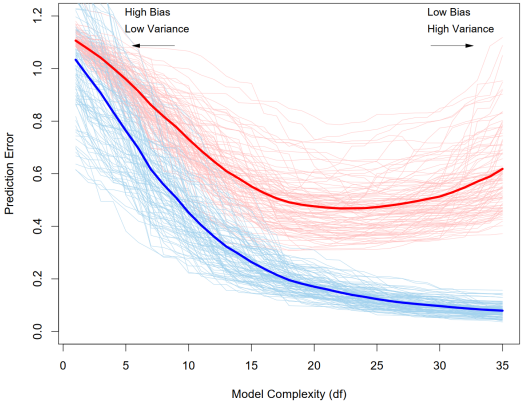


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

上机实验部分

- 使用 Python 作为主要编程软件
- 推荐 VSCode 代码编辑器 + GitHub Copilot 插件进行 AI 辅助编程
- 需要注册 GitHub 账号
  - 校园网络可访问 GitHub（https://github.com ）
  - 用学校邮箱注册为学生（https://github.com/education/students）可获取 Copilot 无限使用权
- 可使用 git 进行代码管理