- 机器学习7天速通指南
  - Day 1: 基础核心
    - 机器学习定义与目标
    - 三大学习范式
    - 关键概念
    - 评估指标
  - Day 2-4: 核心模型
    - 监督学习模型
      - 线性回归
      - 逻辑回归
      - 决策树与随机森林
    - 无监督学习模型
      - K-Means聚类
      - PCA降维
    - 深度学习基础
      - 神经网络基础
      - CNN卷积神经网络
      - RNN与LSTM
  - Day 5-6: 快速实践

机器学习7天速通指南

Day 1: 基础核心

机器学习定义与目标

- **核心思想**：从数据中自动发现模式并进行预测/决策
- **核心流程**：数据收集 → 预处理 → 模型训练 → 评估 → 部署
- **关键区别**：与传统编程不同，ML是从数据中学习规则而非显式编程规则

三大学习范式

- **监督学习**：使用有标注数据（回归/分类问题）
  - 回归：预测连续值（如房价）

    - 分类：预测离散类别（如垃圾邮件识别）
- **无监督学习**：使用无标注数据（聚类/降维问题）
    - 聚类：发现数据内在分组（如客户细分）
    - 降维：减少数据维度保留关键信息
- **强化学习**：智能体与环境交互获取奖励（了解即可）
    - 适用于序列决策问题（如游戏AI、机器人控制）

关键概念

- **过拟合**：模型过于复杂，记忆训练数据而非学习通用模式
    - 解决方案：正则化(L1/L2)、交叉验证、Dropout(深度学习)
    - 检测方法：训练误差远低于验证误差
- **欠拟合**：模型过于简单，无法捕捉数据中的基本模式
    - 解决方案：增加特征、使用更复杂模型、减少正则化
    - 检测方法：训练和验证误差都很高
- **No Free Lunch定理**：没有万能模型，需根据问题选择合适算法
    - 实际意义：模型选择取决于数据特征和问题背景

评估指标

- **回归问题**：均方误差(MSE)、平均绝对误差(MAE)、R^2分数
- **分类问题**：准确率、精确率、召回率、F1分数、AUC-ROC曲线
- **聚类问题**：轮廓系数、Calinski-Harabasz指数

Day 2-4: 核心模型

监督学习模型

线性回归

- **公式**：$y = w \cdot x + b$（可扩展为多元：$y = w_1x_1 + w_2x_2 + ... + b$）
- **损失函数**：最小二乘法（最小化预测值与真实值的平方差）
- **优化方法**：梯度下降法迭代更新权重
- **正则化变体**：
    - 岭回归(L2正则化)：防止过拟合，适用于特征较多的情况

    - Lasso回归(L1正则化)：可产生稀疏模型，适用于特征选择
- **应用**：房价预测、销量预测、经济指标分析

逻辑回归

- **核心函数**：Sigmoid函数 $\sigma(z) = 1/(1+e^{-z})$ 将线性输出映射到(0,1)区间
- **决策边界**：通常以0.5为阈值进行二分类
- **损失函数**：交叉熵损失（对数损失），更适合概率评估
- **多分类扩展**：通过Softmax函数实现多类别分类
- **应用**：垃圾邮件检测、用户流失预测、疾病诊断

决策树与随机森林

- **决策树**：通过特征问答构建树形结构，易于解释
    - 分裂标准：信息增益、基尼不纯度
    - 优点：无需特征缩放、处理混合数据类型
    - 缺点：容易过拟合、对数据微小变化敏感
- **随机森林**：多个决策树的集成算法（Bagging方法）
    - 核心思想：通过构建多棵树并投票提高泛化能力
    - 超参数：树的数量(n_estimators)、最大深度(max_depth)
    - 优点：减少过拟合、处理高维数据、提供特征重要性
    - 应用：信用评分、疾病诊断、客户细分

无监督学习模型

K-Means聚类

- **算法流程**：随机初始化中心点 → 分配点到最近簇 → 重新计算中心点 → 迭代至收敛
- **距离度量**：通常使用欧氏距离，也可用余弦相似度等
- **K值选择**：肘部法则（SSE随K变化曲线）、轮廓系数
- **优缺点**：简单高效但需预设K值且对初始中心敏感
- **变体**：K-Medoids（对异常值更鲁棒）、Mini-Batch K-Means（大数据集）
- **应用**：客户分群、图像分割、文档聚类

PCA降维

- **数学原理**：通过线性变换将数据投影到方差最大的方向（主成分）
- **核心步骤**：数据中心化 → 计算协方差矩阵 → 特征值分解 → 选择主成分

- **方差解释**：每个主成分保留的原数据方差比例
- **应用场景**：高维数据可视化、特征提取、数据压缩
- **注意事项**：PCA是线性方法，核PCA可处理非线性关系

深度学习基础

神经网络基础

- **感知机**：最基本神经网络单元，加权求和后通过激活函数
- **激活函数**：引入非线性，常用ReLU、Sigmoid、Tanh
- **反向传播**：通过链式法则计算梯度，使用梯度下降更新权重
- **优化器**：SGD、Adam、RMSprop等，加速收敛并避免局部最优

CNN卷积神经网络

- **核心结构**：卷积层(特征提取) → 池化层(降维) → 全连接层(分类)
- **卷积操作**：使用滤波器提取局部特征，参数共享大幅减少参数量
- **典型架构**：LeNet、AlexNet、VGG、ResNet等
- **应用**：图像识别、目标检测、语义分割

RNN与LSTM

- **RNN问题**：处理序列数据但存在梯度消失/爆炸问题
- **LSTM创新**：引入门控机制(输入门、遗忘门、输出门)控制信息流
- **变体**：GRU(简化版LSTM)、双向LSTM(捕捉前后文信息)
- **应用**：文本生成、时间序列预测、机器翻译

Day 5-6: 快速实践