第一章 机器学习概论

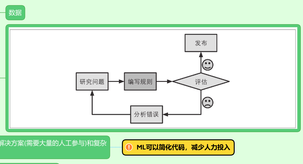
机器学习是什么？

机器学习是一门通过编程让计算机从数据中进行学习的科学(和艺术)

为什么使用机器学习？

和传统学习方法的区别

数据



ML适用于

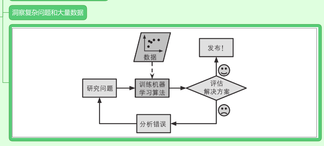
有复杂的解决方案(需要大量的人工参与)和简单的code

ML可以简化代码，减少人力投入

传统方法根本解决不了的问题

环境有波动，ML可以适应新数据

洞察复杂问题和大量数据



机器学习的应用示例

CV

- 图像分类
  - 生产产品分类
- 语义分割
  - 医学影像处理

NLP

- 新闻分类
- 标记不良发言
- 对长文本进行总结

- 创建小冰和客服
  - NLU+问答模块
- 利用多个指标来预测下一年的收入
  - 回归问题
  - 线性、多项式、SVM、随机森林，人工神经网络
- 语音识别
- 异常检测
  - 检测欺诈
- 用户分类
  - 聚类
- 利用图表表示高维数据集
  - 数据可视化
- 推荐产品
  - 推荐系统
- 创造人机
  - 玩

机器学习系统的类型

分类方式(互不干扰)

- 有监督
  - 有无半强
- 是否增量
  - 在线学习
  - 批量学习
- 基于What
  - 基于实例
    - 类似度量
  - 基于模型
    - 常见

有无监督

有监督学习

- 带标签
- 垃圾邮件分类器
- $(x^{(i)},y^{(i)})$

无监督学习

- 不带标签
- 聚类
- $(x^{(i)})$

半监督学习

主要用于难于给数据标记的情况

会将部分已标记样本作为正样本(三角形)和负样本(正方形)

正样本负责将未分类的样本(圆点)拉过来

负样本负责将未分类的样本推过去

强化学习

智能体根据环境来取步骤，获得奖励或者惩罚，学习目的是随着时间的推移获得最大的回报

批量学习和在线学习

- 区别
  - 数据集是否变化(增加)
- 批量学习
  - 无法进行增量学习(只能基于传入的数据流进行学习)
  - 缺点是训练测试集和未来的实用集可能有不一致，缺点是可以尽量实现自动化
- 在线学习
  - 可以进行增量学习
  - 缺点是容易受到噪声干扰，优点是实用集和训练测试集分布高度一致；而且适用于超大数据集

基于实例或模型

基于实例

- 关查表
- k临近，计算新样本和已知的k个样本的距离，根据结果来分类

基于模型

- 最普遍的情况
  - 研究数据分布
  - 选择模型
  - 一般步骤
  - 使用训练数据进行训练(通过学习算法搜索模型参数值，使得代价函数值最小)
  - 应用模型对新样本进行预测，用于衡量模型的泛化结果

机器学习的主要挑战

主要类型

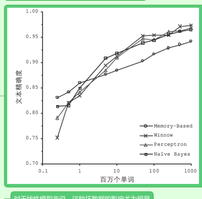
坏算法

坏数据

坏数据

训练数据的数量不足

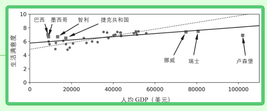
大部分机器学习算法都需要大量的数据才能正常工作



在数据足够充足时，很多算法的性能都近似

训练数据不具有代表性

对于线性模型来说，这种坏数据的影响尤为明显



如图所示，圆点是我们模型训练时并未估计到方点的收据集，实线是将方点加入数据集之后训练所得模型

低质量数据

- 训练集离群值、异常值和噪声，使得模型难以检测到底层特征
- 解决办法
  - 直接丢弃明显错误样本，尝试手动修复错误
  - 若样本缺少部分特征，可以：整体忽略这些特征，忽略缺失特征的样本，将缺失值补充完整(中位数)

无关特征

- 特征选择(最有用的特征)
- 特征提取(对现有特征进行整合，产生更有用的特征)
- 通过收集新数据创建新特征

坏算法

过拟合训练数据

- 产生原因
  - 模型过于复杂
    - 高次项权重过大，过拟合就是为了减少高次项的权重贡献
  - 数据集过少
  - 训练Epoch过多
  - 数据集中噪声过大
- 解决办法
  - 构建更小模型
  - 收集更多数据
  - 早停止
  - 减少训练集中噪声
- 方差(测试/验证集)很大，误差(训练集)相对较小

欠拟合

- 模型过于简单
  - 线性模型难以拟合二次分布的特征
- 产生原因
  - 数据集过多(一般不嫌多)
  - 训练Epoch过少
  - 模型约束过多
- 解决办法
  - 换一个更复杂的模型
  - 利用特征工程削减特征集
  - 减少模型约束
- 方差(测试/验证集)和误差(训练集)都很大

测试和验证

我们的模型最终会运用在新场景(新数据)，为了了解模型性能，我们不能只通过基于训练集的表现来看(毕竟存在过拟合)，所以我们需要分割出部分数据来模拟“新场景”，这便是训练集。

超参数调整和模型选择

单个模型可以通过基于测试集上面的表现来衡量，多个模型如何衡量呢？我们可以将这多个模型都训练来看基于测试集的泛化效果。

但是在我们多次基于测试集挑选出效果最好的模型之后，该模型可能对测试集过拟合而导致在实用集上面表现不好，所以我们仍需要另外一个集合用于和合在实用集上面的性能，这个集合被称为开发集。

三个集合总结

- 训练集
  - 用于训练模型
- 开发集
  - 用于调整超参数和选择模型
- 测试集
  - 用于测试模型的最终性能

数据不匹配

- 实用集和训练测试集分布不同的情况
- 我们可以将训练集分出一部分用于开发(train-dev)，表现不好则来自数据不匹配