ML概述

机器学习基础

数据

数据驱动：基于客观的量化数据，通过主动数据的采集分析支持决策。(相对的是经验驱动)

模型

模型：基于数据X做决策Y的假设函数。

算法

算法：学习模型的具体计算方法。(通常是**最优化问题**)

机器学习的核心技术

分类

用分类数据训练模型，利用模型对新样本分类和预测。

> 应用领域：构建用户画像、情感分析、用户行为预测、图像识别分类

聚类

从数据中识别数据的相似性、差异性，按照最大共同点聚合为多个类别

> 应用领域：市场细化、模式识别、空间数据分析、图像处理与分析

异常检测

对数据分布规律进行分析，识别与正常数据差异较大的离群点

> 应用领域：日常运行监控、风险识别、舞弊检测

回归

根据对已知属性值数据的训练，为模型寻找最佳拟合参数，基于模型预测新样本输出值

> 应用领域：趋势预测、价格预测、流量预测

机器学习流程

数据预处理

1. 输入：未处理数据+标签
2. 处理过程：特征处理&幅度缩放+特征选择+维度约减+采样
3. 输出：训练集+测试集(+验证集)

模型学习

- 选择模型
- 交叉验证
- 评估
- 选择超参

模型评估

模型对数据集的得分

新样本预测

预测测试集

机器学习专有名词

1. 监督学习：训练集有标记信息(人工标注)，可以用**分类和回归**的方式学习
2. 无监督学习：训练集没有标记信息，可以用**聚类和降维**的方式学习
3. 强化学习：有延迟和稀疏的反馈label的学习方式

> 样本/示例、属性/特征、属性空间/样本空间/输入空间、特征向量(空间中每个点对应的一个坐标向量)、**标记(关于示例结果的信息)**、分类、假设、真相(潜在规律自身)、学习过程、**泛化能力(学得模型适用于新样本的能力，一般来说训练样本越大，越有可能获得泛化能力强的模型)**

机器学习算法分类

分类问题

1. 二分类
2. 多类分类
3. 多标签分类：每个样本一系列目标标签

回归问题

聚类问题

降维问题

模型评估

> 模型评估的目标是选出泛化能力强的模型
>
> 分为离线和在线实验方法：(通常是指离线方法)
>
> 1. 离线方法：
>   - 使用历史数据训练模型
>   - 对模型进行验证(Validation)和离线评估(Offline Evaluation)
>   - 通过评估指标选择较好的模型
> 2. 在线方法：比如A/B test，制定两个方案让一部分用户使用A方案，一部分用户使用B方案。评估指标通常为Customer Lifetime Value(用户生命周期值)、Click Through Rate(广告点击率)、Customer Churn Rate(用户流失率)

过拟合问题

过拟合：在训练集表现很好，但是在交叉验证集合测试集上表现一般。泛化能力较差

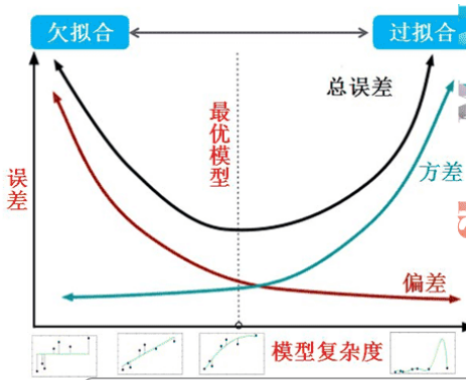
防止过拟合的方法：early stopping、data augmentation(数据集扩增)、正则化(在目标函数后添加正则化项)、dropout(修改神经网络本身的结构实现)

偏差

偏差：模型拟合的偏差程度，是真实模型与平均模型(给定大量训练集而期望拟合出来的模型)的差异

方差

方差：模型的平稳程度



评估指标

1. 回归问题：MAE(平均绝对误差)、MAPE(平均绝对百分误差)、MSE(均方误差)、RMSE(均方根误差)、决定系数

> $MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$，标签值与预测值偏差的绝对值的平均，缺点是不能反映预测的无偏性。
>
> $MAPE = \frac{100\%}{n}\sum_{i=1}^{n}|\frac{\hat{y}_i - y_i}{y_i}|$，对$MAE$的改进，考虑了绝对误差相对真实值的比例。
>
> $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$，绝对值的存在导致函数不光滑，而平方不会。
>
> $RMSE = \sqrt{MSE}$
>
> **总平方和SST：** $SST = \sum_i (y_i - \overline{y})^2$
>
> **回归平方和SSR：** $SSR = \sum_i (\hat{y}_i - \overline{y})^2$
>
> **残差平方和SSE：** $SSE = \sum_i (\hat{y}_i - y_i)^2$
>
> $R^2 = \frac{SSR}{SST}$，比例越接近1表示当前回归模型对数据的解释越好，表征因变量y的变化中有多少可以用自变量x来解释。

2. 分类问题：Error Rate(错误率，分类错误样本数占样本总数的比例)、Accuracy(精确率acc，分类正确的样本数占样本总数的比例)、Precision(准确率)、Recall(召回率)、F_1、F_β、ROC(受试者工作特性曲线)、AUC曲线、PR曲线、R^2

   混淆矩阵如下：TP：True Positive，其余类似。每一列代表预测类别，每一行代表数据的真实归属类别。

(Confusion Maxtrix)	Prediction Positive	Prediction Negative
Condition Positive	TP	FN
Condition Negative	FP	TN

> $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$
>
> $Precision = \frac{TP}{TP+FP}$，真正正确的个数占预测中正确的比例
>
> $Recall = \frac{TP}{TP+FN}$，真正正确的个数占整个数据集中真正正确的比例
>
> $F_1 = \frac{2}{\frac{1}{P}+\frac{1}{R}}$
>
> $F_\beta = (1+\beta^2)\cdot\frac{P\cdot R}{\beta^2\cdot P+R}$，当$\beta=1$时就是$F_1$；当$\beta<1$时更关注$Precision$；当$\beta>1$时更关注$Recall$
>
> ROC纵轴为**真正例率**$TPR=\frac{TP}{TP+FN}$横轴为**假正例率**$FPR=\frac{FP}{FP+TN}$，曲线越接近左上角，分类器性能越好。
>
> AUC是ROC下面积，物理意义是正样本的预测结果大于负样本的预测结果的概率，本质就是分类器对样本的排序能力。
>
> PR横坐标是R，纵坐标是P

3. 评估指标选择技巧：

Accuracy	适用正负样本相当的情况
Precision	注重"准"，适于正负样本差异很大情况
Recall	注重"全"，适于正负样本差异很大的情况
ROC	对不平衡数据不敏感
AUC	对排序敏感，对预测分数不敏感
PRC	适于负样本数量远大于正样本数量的数据集

评估方法

留出法(Hold-out)

1. 从训练数据中保留**验证集**，不用于训练而用于模型评估。随即划分不一定有效(可以采用分层抽样等方式)。
2. 通常采用**划分-训练-测试求误差**的方式，最后求出误差的平均值(单次划分不一定能得到合适的测试集)。

交叉验证法(Cross Validation)

K折交叉验证：对K个不同分组训练的结果进行平均以减少方差。当K为样本总数时，叫做**留一法**。

Tip：数据量小则K设大点；数据量大则K设小点。

自助法(Bootstrap)

用小样本估计总体值的非参数方法(适用于数据量少的情况)。通过有放回抽样生成**伪样本**，通过计算伪样本获得统计量分布，从而估计数据的整体分布。

e.g. m个样本进行m次有放回抽样得到Training Set，剩下未出现在训练集中的作为Test Set

样本均衡与采样

长尾现象：多数样本数量多信息量大，模型充分识别；少数样本数量少信息量小，模型没有充分学习到特征。

解决方式：数据采样和样本加权

1. 数据采样：
    - 欠采样/下采样：将数据从原始数据集中移除(多数类集合中筛选样本集移除)
    - 过采样/上采样： e.g.随机过采样 →首先在少数类集合中随机选中少数类样本；通过复制生成样本集合E；添加到少数类集合中扩大原始数据集得到新的少数类集合
2. 样本加权：对不同类别分错的代价不同。 e.g. 小众样本分错了造成更大的损失