

机器学习

线性判别分析 (LDA)

- LDA思想总结
- LDA算法原理
- LDA算法流程
- LDA与PCA对比

主成分分析 (PCA)

损失函数

- 常见损失函数
 - 0-1损失函数
 - 绝对值损失函数
 - 平方损失函数
 - 对数损失函数
 - 指数损失函数
 - Hinge损失函数

大数据与深度学习之间的关系

- 数据挖掘
 - 从数据中提取模式的特定算法的应用
- 机器学习
 - 如何构建使用经验自动改进的计算机程序
- 大数据
 - 超出常用软件工具捕获、管理和处理能力的数据集

学习方式

- 监督学习
 - 支持向量机 Support Vector Machine
 - 朴素贝叶斯 Naive Bayes
 - 逻辑回归 Logistic Regression
 - k-近邻 K-Nearest Neighborhood
 - 决策树 Decision Tree
 - 随机森林 Random Forest
 - AdaBoost
 - 线性判别回归 Linear Discriminant Analysis
- 非监督学习
 - Apriori
 - K-Means
- 半监督学习
 - 图论推理 Graph Inference
 - 拉普拉斯支持向量机 Laplacian SVM
- 弱监督学习

监督学习步骤

- 数据集创建和分类
- 数据增强
- 特征工程
 - 特征提取
 - 特征选择
- 构建预测模型和损失
- 训练
- 验证和模型选择
- 测试及应用

模型评估

- 模型评估常用方法
 - 分类模型
 - 回归模型
- 误差、偏差、方差
- 过拟合与欠拟合
- 混淆矩阵
- ROC与AUC
- 代价敏感错误率与代价曲线

贝叶斯分类器

- 极大似然估计原理
- 贝叶斯分类器基本原理
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器

决策树

- 决策树基本原理
- 决策树三要素
- 决策树学习基本算法
- 决策树算法优缺点
- 熵
- 信息增益

降维与聚类

EM算法

支持向量机

逻辑回归

- 平方误差代价函数的主要思想就是将实际数据给出的值与拟合出的线的对应值做差，求出拟合出的直线与实际的差距。
 - 代价函数作用原理
 - 最优解即为代价函数的最小值
- 目标函数存在一个下界，若优化算法能够使目标函数不断缩小，根据单调有界原理，可证明优化算法收敛有效。故代价函数非负更方便
 - 为何代价函数要非负
- 常见代价函数
 - 二次代价函数
 - 交叉熵代价函数
 - 对数似然代价函数

- 用于概率预测
- 用于分类
- 适用性

- 各特征之间不需要满足条件独立假设，但各个特征的贡献独立计算
- 逻辑回归是判别模型，朴素贝叶斯是生成模型
- 朴素贝叶斯是贝叶斯，逻辑回归是最大似然
 - 朴素贝叶斯需要条件独立假设
- 逻辑回归与朴素贝叶斯区别
 - 逻辑回归需要特征参数之间是线性的
 - 目的：线性回归是预测，逻辑回归是分类
 - 输出：线性未知，逻辑 (0, 1)
 - 函数：线性拟合函数，逻辑预测函数
 - 线性回归与逻辑回归区别
- 参数计算方式：线性最小二乘法，逻辑极大似然估计

分类算法

常用分类算法比较

- 常用术语
 - 被正确地划分为正例的个数
 - True positives
 - 被错误地划分为正例的个数
 - False positives
 - 被错误地划分为负例的个数
 - False negatives
 - 被正确地划分为负例的个数
 - True negatives
- 评价指标
 - 正确率 accuracy
 - $accuracy = (TP + TN) / (P + N)$
 - 被分对的样本数在所有样本数中的占比
 - 错误率 error rate
 - $error\ rate = (FP + FN) / (P + N) = accuracy - 1 - error\ rate$
 - 描述被分类器错分的比例
 - 灵敏度 sensitivity
 - $sensitivity = TP / P$
 - 衡量了分类器对正例的识别能力
 - 特异性 specificity
 - $specificity = TN / N$
 - 衡量了分类器对负例的识别能力
 - 精度 precision
 - $precision = TP / (TP + FP)$
 - 精度是精确性的度量
 - 召回率 recall
 - $recall = TP / (TP + FN) = TP / P = sensitivity$
 - 召回率是覆盖面的度量
 - 鲁棒性
 - 处理缺失值和异常值的能力

分类算法评估方法

是以灵敏度（真阳性率）为纵坐标，以1减去特异性（假阳性率）为横坐标

Receiver Operating Characteristic Curve

图例

以recall为横坐标，precision为纵坐标绘制的曲线

该曲线的所对应的面积AUC实际上是目标检测中常用的评价指标平均精度（Average Precision, AP）。AP越高，说明模型性能越好

代表意义