Pattern Recognition 模式识别

参考： 《模式识别》 上海交通大学出版社

课程《Pattern Recognition》 By Prof Huang

绪论

- 什么是模式识别
    - 从大量信息与数据出发，在专家经验与已有认知的基础上，利用数学推导对模式、数字、图形等自动完成识别的过程
    - A pattern is the opposite of a chaos，模式是对存在于时空中的某种客观事物的描述
    - eg： Regressing, Clustering，Classification， Dimension Reduction and so on
    - todo： Solving a pattern recognition problem is to model an optimization problem, usually a continuous programming, and solve it.
    - 模式识别系统：信息获取->预处理->特征提取->分类决策->输出
- 特征选择
    - 经典传统算子，如SIFT，CANNY etc.
    - 设计特殊特征
    - 学习：神经网络
- 卷积算子(convolution operator)：CNN 如Lenet-5
- 优化 optimization
    - the selection of a best element from some set of available alternatives.
- 线性约束问题(Linear programming)
    - 解一：单纯形法 (Simplex Algorithm)
        - 可行域->多边形
        - 至少存在一个定点为最优解
        - However, NP Hard
    - 解二： 内点法(Interior point method)，不等式约束
        - 使用对偶问题，可行域搜索
- 非线性约束问题 Non-linear Constraint Problem
    - Gauss， GN， LM等

线性回归问题 Linear Regression

- loss function 主要是平方损失函数

    - squared residual
    - absolute residual
    - huber pinball and so on
- 直接求逆 $O(n^3)$
- 梯度下降法
    - 梯度下降：GD
    - 随机梯度下降法SGD: stochastic gradient descent
    - SGD 收敛快且占用内存小
    - 还有SAG等等
- overfitting problem
- Lasso回归
- 非线性： Logistic regression （逻辑回归）解决分类问题
    - 线性回归+sigmoid非线性映射
    - 数据服从伯努利分布情况下，通过极大似然的方法，使用梯度下降法求解参数，从而达到数据二分类

Linear Classification

- Fisher判别分析：LDA模型 Fisher's linear discriminant
    - 将不同类别的点尽可能分离开，投影到低维空间再分类
    - 最优判别准则：同属一类的样本之间的距离越小越好
- Logistic regression
    - 推广: multinomial logistic regression
- Perceptrons
- to solve
    - (generalized) eigen-value problem
    - 一阶：SGD:Stochastic gradient descent
    - 二阶：Newton's method
    - IRIS方法:iterative reweighted least squares

无监督学习

- Unsupervised learning
- only have data ,but no label/target

降维

- 降维常用于其它有监督算法之前

PCA

- 主成分分析 Principal Component Analysis
- Dimensionality reduction
- 组成矩阵->中心化->协方差矩阵及其特征值和特征向量->按特征值排序，取前n大（协方差矩阵分解可基于特征值，也可基于SVD分解）
- 将原本的数据进行映射降维：find a direction that maximizes the data's variance，最小化投影损失，或者说最大化保留投影后数据的方差
- directions with low eigenvalues usually correspond to irrelevant aspects of data， use top K directions to re-represent the data for Denoising/Compression/Correction/Visualization
- 若非线性，可考虑kernel PCA
- 其余线性降维方法：LDA算法:Linear Discriminant Analysis(也称Fisher Linear Discriminant)，降维后的点尽可能分开，有监督

非线性降维方法

- Neighbor embedding, 根据一个点与其临近点关系进行降维分析，又称流形学习(Manifold Learning)
- SNE算法(Stochastic Neighbor Embedding):随机邻近嵌入
    - 依据条件概率估算两点之间的距离
    - 可用联合概率joint probability来替代条件概率conditional probability
    - t-SNE算法，将SNE中使用的Gaussian分布换为t分布
    - 后两种方法考虑了相似数据的聚集，SNE则进一步考虑了不相似数据的分开
- LLE算法(Locally Linear Embedding)：局部线性嵌入
    - 假定每个点都能通过周围数据的线性组合表示，即基于局部线性，且降维后这一关系尽可能的保持不变
- LE算法(Laplacian Eigenmaps):拉普拉斯特征映射
    - 基于graph，相互有关系的点在降维后的空间仍然能保持原有的结构、尽可能接近

Clustering

- 分类
    - 层次聚类：Hierarchical Clustering
        - 自下而上 Bottom-up：凝聚法
        - 自上而下 Top-down：分裂法
    - 扁平算法：开始随机划分，迭代修正，如k-means
- 硬聚类：仅属于一个标签。软聚类：样本可属于多个标签

- k-means聚类
    - 到其所在簇的质心向量的平方和最小
    - 初始化聚类中心(如随机分配）->最小距离分类->重新计算质心向量->再次分配
    - 缺点：k值的确定，非凸数据集难收敛，噪声和离群点敏感，各类别数据失衡影响大

EM algorithm

- Expectation Maximization Algorithm 最大期望算法
- 思想：最大似然估计+迭代，这一思想在很多算法中有着应用
    - Expectation-step: 每个样本计算属于各类的概率，根据概率打标签
    - Maximization-step: 根据重新打好的标签估计模型参数
- GMM：Gaussian mixture model

AutoEncoder 自编码器

特殊的神经网络架构，深度学习早期的特征提取的重要方法

应用于降维、异常值检测等

半监督与无监督领域

度量学习 Metric Learning

度量->如距离，一个具有度量的集合可以称为度量空间，按这个说法，大部分基于度量或者相似度的算法都可以说是度量学习

目前在人脸识别等领域有着广泛应用

存在有监督与无监督的

- 度量：反身性、对称性、三角不等式
- 欧式距离(明考夫斯基距离在p=2时的特例)，绝对值距离，切比雪夫距离等

半监督学习

- 无监督学习->先验知识，有监督学习->标签，半监督->部分数据有标签
- 先验
    - continuity assumption： 相近的点，标签相同
    - cluster assumption： 聚类假设，同一类的点标签相同
    - Manifold assumption：输入空间由多个低维manifold组成，同一manifold上标签相同
    - Low-density assumption
- 可融入EM算法思想

- 伪标签学习，或者称简单自训练：simple self-training
- 方法
    - 半监督SVM 或者叫S3VM
    - 半监督机器学习

监督学习

Ensemble Learning 集成学习

- 有监督学习->完美的模型，实际上->多个有偏好的模型(弱监督模型)->集成学习：组合弱监督学习模型，即便某一模型出错，也不会造成很大损失
- 一种思想，在其它算法中有着广泛应用

1. Bagging

- 思想：模型多势重
- 如：分类问题->vote, 回归问题->平均，应用如随机森林

2. Boosting

- 思想：迭代进步
- Adaboost(Adaptive boosting算法)：带权重训练多次，对出错的训练例给予更大的权重，即注意学习出错的示例
- GBDT：Gradient Boost Decision Tree，多个树，每个树学的是之前所有树的残差
    - 每次迭代都学习一颗CART树来拟合之前n-1颗树的残差，初值敏感
- XGBOOST，GBDT优化：目标函数加入正则项（相当于预剪枝防过拟合）、损失函数二阶泰勒展开(相比于一阶更精确逼近真实损失)等

3.Stacking

- 训练多个分类器，将多输出接一个模型(如knn、随机森林、朴素贝叶斯分类器)进行输出

Decision tree决策树

- 有监督，分类器或回归
- 常用的
    - CART:Classification and Regression Trees 分类回归树(最常用)
        - 使用信息增益 Gain index
            - 信息增益是以某特征划分数据集前后的熵的差值，熵可以表示样本集合的不确定性，熵越大，样本的不确

定性就越大。

        - 可分类可回归
    - ID3
        - 使用信息增益 Information Gain
        - 分类问题
        - 只能处理离散数据，对缺失值敏感
    - C4.5
        - 使用信息增益比 Gain Ratio（=信息增益*加权系数，对取指过多的特征进行惩罚，避免过拟合）
        - 分类问题
    - 以上都是自上而下的贪心算法，度量方式不同
- 剪枝：前剪枝和后剪枝
- 基本只在小数据集上使用，容易过拟合，受数据量影响大
- 改进：随机森林，GBDT(梯度提升决策树)

Random forest随机森林

决策树不鲁棒不稳定->随机森林：训练多个决策树，集成多个树的结果，bagging典型应用

- 两个随机保证：每棵树数据集随机、特征随机采样
- 缺点：噪声大的数据易过拟合

SVM

- 支持向量机：Support Vector Machine
- find a hyperplane to separate two classes
- Lagrange duality
- kkt condition
- optimal margin classifier
- 超参：C，gamma，即正则系数与支持向量的数目
- 缺点：对缺失数据敏感，对参数和核函数敏感
- 非线性SVM
    - kernel tricks 核函数代替内积
        - Polynomial、RBF、Mercer kernel

KNN

- K-Nearest Neighbor
- 测试点与已有label点计算距离，取前k个点的标签，最多的即为预测
- 与kmeans有点像，knn是监督分类问题，kmeans是无监督聚类