西瓜书备忘录

Li Liang*

1 绪论

1.1 No Free Lunch Theorem

总误差与学习算法无关，但针对特定问题时，算法表现不同。脱离实际问题，空谈学习算法毫无意义。

2 模型评估与选择

2.1 评估方法

- 留出法（hold-out）

  常见做法是大约 2/3—4/5 的样本用于训练，剩余样本用于测试。

- 交叉验证法（cross validation）

  k 折交叉验证通常要随机使用不同划分重复 p 次，常见的有 10 次 10 折交叉验证。

  - 留一法（LOO, LEVAE ONE OUT）

- 自助法（bootstrapping）

  数据集较小时很有用，又称包外估计（out of bag estimate），但是改变了初始数据集的分布，可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用。

2.2 最终模型

在学习算法和参数配置已选定后，需要利用完整数据集重新训练模型，得到最终模型。

*https://github.com/leeliang/

2.3 性能度量

- 查准率、查全率、F1

  查准率为在预测为正例的结果中有多少是真正的正例，查全率为真实结果为正例的样本中有多少预测结果为正例。

  $$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}$$

  F1 是基于查准率和查全率的调和平均数定义的，用于综合考虑查准率和查全率的度量。若对查准率和查全率的重视程度不同，则采用更一般的加权调和平均。

  $$F_\beta = \frac{1}{1+\beta^2}(\frac{1}{P} + \frac{\beta^2}{R})$$

  $\beta = 1$ 时退化为 F1，$\beta < 1$ 时查全率影响更大，反之亦然。

- ROC

  我们可根据学习器的结果对样例进行排序，排在最前面的是最可能是正例的样本，按此顺序逐个把样本作为正例进行预测，则得到了查准率和查全率的序列，以查准率为纵轴，查全率为横轴的图为 P-R 曲线。以 True Positive Rate 为纵轴，False Positve Rate 为横轴的图为 ROC 曲线。TPR 是真实情况为正例中有多少真正例，FPR 是真实情况为反例的样本中假正例。

  $$TRP = \frac{TP}{TP+FN}, FPR = \frac{FP}{TN+FP}$$

  ROC 曲线下的面积为 AUC。

2.4 比较检验

- 二项检验、t 检验 (一个算法检验)
- 交叉验证 t 检验（一个数据集两个算法）
- McNemar 检验（一个数据集两个算法、二分类问题）
- Friedman 检验和 Nemenyi 后续检验 (多数据集多算法)

2.5 偏差与方差

偏差-方差分解是解释学习器泛化性能的一种重要工具。泛化误差可以分解为偏差、方差和噪声之和。

$$E(f;D) = bias^2(x) + var(x) + \epsilon^2$$

偏差度量了偏离程度；方差度量了同样大小的训练集的变化导致的学习性能变化，刻画了数据扰动造成的影响；噪声为当前任务在任何学习算法下所能达到的泛化误差下届，刻画了问题的难度。偏差-方差分解说明泛化性能由学习算法的能力、数据的充分性以及任务的难度共同决定。

3 线性模型

对于离散属性，如果其存在顺序关系，可通过连续化将其转化为连续值。

3.1 LDA

LDA 算法见 PDF 文件。

3.2 类别不平衡问题

对决策规则进行缩放：再缩放策略。再缩放策略的难度在于训练集的比例并不一定是总体样本的比例。现在技术上大体有三类做法：

- 欠采样：去除一些反例使得正反例数目相近；
- 过采样：增加正例使得相近；
- 阈值移动：基于原始数据训练，在分类器预测时，根据正反例比例缩放阈值。

4 决策树

4.1 决策树原理

决策树算法见 PDF 文件。

4.2 连续与缺失值

对连续值进行排序，选取相邻点的中间点作为候选划分点，计算每个划分点的增益，选取增益最大的候选点作为划分点。

对于缺失值问题，属性划分可使用没有缺失值的样本子集；若树已经生产，对于属性值缺失的样本以不同的概率划入到不同的子结点。

4.3 多变量决策树

不再是每个属性独立划分的，而是多个属性的加权确定分界线，减少开销。

5 神经网络

5.1 感知机

见 PDF 文件。

5.2 BP 算法

见 PDF 文件。

5.3 全局最小与局部极小

常采用跳出局部极小的策略：

- 从不同初始点开始搜索；
- 使用模拟退火技术，模拟退火在每一步都以一定概率接受比当前更差的解；
- 使用随机梯度下降。

上述跳出局部极小的技术多为启发式，理论上缺乏保障。

6 支持向量机

6.1 SVM

见 PDF。

6.2 损失函数

- hinge 损失；
- 指数损失；
- 对率损失。

软间隔支持向量机的最终模型仅与支持向量有关，即通过采用 hinge 损失函数仍保持了稀疏性。如果采用对率损失函数，即支持向量机变成了对率回归模型。通常情况下，svm 和对率回归的优化目标相近，性能相当，但对率回归优势在于输出具有自然的概率意义，除此以外，对率回归可用于多分类任务。支持向量机的解具有稀疏性，而对率回归没有类似概念，依赖于更多的训练样本。

关于结构风险和经验风险的讨论。

6.3 支持向量回归

传统回归基于模型输出和真实输出之间的差别计算损失，支持向量回归假设能容忍一定的偏差，只有偏差大于容忍值时才计算损失。

7 贝叶斯分类器

7.1 贝叶斯决策和朴素贝叶斯分类器

见 PDF。

7.2 半朴素贝叶斯分类器

为了降低估计后延概率的困难，朴素贝叶斯假设所有样本属性相互独立，但在现实任务中这个假设往往很难成立。半朴素贝叶斯分类器则是适当放松各属性相互独立的假设。

- 独依赖估计（One Dependent Estimator）

  假设每个属性仅依赖一个其他属性，即：

  $$arg\max p(c)\ \prod_{i=1}^{d} p(x_i \mid C_i, p_i) \tag{1}$$

  p_i 为属性 x_i 依赖的属性，称为 x_i 的父属性。现在问题的关键在于怎么确定每个属性的父属性。

  - SPODE：所有属性依赖于同一属性。
  - AODE：尝试每个属性当成父属性，然后将所有结果集成成最终结果。
  - TAN：

7.3 贝叶斯网

略。

7.4 EM 算法

见 PDF。

8 集成学习

弱学习器：泛化性能略优于随机猜测的学习器。

9 聚类

9.1 聚类性能度量指标

聚类性能度量指标大致分为两类：

- 外部指标：与某个参考模型比较
  - Jaccard 系数：$JC = \frac{a}{a+b+c}$
  - FM 指数：$FMC = \sqrt{\frac{a}{a+b}\frac{a}{a+c}}$
  - Rand 指数：$RI = \frac{a+d}{a+b+c+d}$
- 内部指标：
  - DB 指数：具体内容见西瓜书。
  - Dunn 指数

对无序属性可采用 VDM (Value Difference Metric) 计算距离。

9.2 原型聚类

算法先对原型初始化，然后进行迭代更新。

9.2.1 k 均值算法

目标为最小化簇内平方误差：

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2 \tag{2}$$

9.2.2 LVQ 学习向量量化

9.2.3 高斯混合聚类

见 PDF 文件。

9.3 DBSACN 密度聚类

参看北理工视频。待续。

9.4 AGNES 层次聚类

先将数据集中的每个样本看做一个初始簇内，然后找出距离最近的两个簇内合并，该过程不断重复，直至到达预设的聚类簇个数。

若分别采用簇内最小距离、最大距离和平均距离，则分别对应单链接、全链接和均链接 AGNES 算法。

10 降维与度量学习

10.1 MDS 多维缩放

根据任意两个样本的欧式距离在两个空间中相等，推导出降维后样本的内积矩阵 $B = Z^T Z$ ，对矩阵 $B = V\Lambda V^T$ 进行特征值分解，选取其中 d' 个最大特征值构成的特征值矩阵 $\widetilde{\Lambda}$，则 $Z = \widetilde{\Lambda}^{1/2}\,\widetilde{V}^T$。

10.2 PCA

对协方差矩阵进行特征值分解 $XX^TW = \lambda W$ 求取 W，$Z = W^T X$。

10.3 核化线性降维

10.4 流形学习

11 特征选择与稀疏学习

11.1 子集搜索与评价

子集搜索分为前向、后向和双向搜索。前向搜索为在上一轮选定的特征集中加入一个特征，选定最优特征集合，后向从完整特征集开始，每次减少一个特

征，选定最优。

子集评价可采用信息增益作为准备，与决策树类似。

11.2 特征选择

将子集搜索与评价结合起来，即可得到特征选择方法。常见的特征选择方法分为过滤式、包裹式和嵌入式。

- 过滤式：根据度量特征重要性的相关统计量 (Relief)，过滤特征，然后再训练学习器，特征选择过程与后续学习无关。
- 包裹式：将学习器性能作为特征子集的评价准则。
- 嵌入式：特征选择过程与学习训练过程融为一体，如 L1（LASSO 回归）和 L2 正则化 (岭回归)。若 w 取得稀疏解意味着初始特征中部分特征的权值为 0。

字典表示、稀疏学习和压缩感知

12 计算学习理论

略。

13 半监督学习

利用标记样本和未标记样本学习。利用未标记样本最常见的假设有聚类假设和流形假设。聚类假设假设数据存在簇结构，流形假设假设数据分布在流形结构上，邻近样本具有相似输出，流形假设可以看成聚类假设的扩展。半监督学习可分为纯半监督学习和直推学习，纯半监督学习是预测其他样本，直推学习预测未标记样本。

14 概率图模型

概率图模型是用图形表达变量相关关系的概率模型。

14.1 隐马尔科夫模型 HMM

见 PDF。

15 规则学习

略。

16 强化学习

略。