机器学习

KNN：通过最近的K个邻居来做决策，可以分类（多数投票）也可以回归（特征方程均值）

- 距离测度：
  - Manhatan Distance：用于非常高维度的情况，且用于离散数据

$$d(X,Y)=\sum_{i=1}^{n}|x_i-y_i|$$

  - Euclidean DIstance：用于连续变量

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(y_i-x_i)^2}$$

  - Hamming Distance：计数对于两个二进制串中不同bit的数量，用于字符串或DNA
  - Cosine Similarity：用于文档向量，文本数据

$$cos(\theta)=\frac{A\cdot B}{||A||||B||}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}}$$

- 超参数K：通常用奇数以免于投票相等
- **（变种）权重投票Weighted voting：**考虑邻居的距离对投票的影响

$$y = \sum \frac{1}{dist(k,p)} \times k_{label}$$

- （变种）Spherical-KNN

Naive Bayes：

- 假设：特征是相互独立的，即$P(X_1...X_n|Y) = \prod_i P(X_i|Y)$
- 已知：
  - Prior $P(Y)$
  - 给定类型Y，n个条件独立的特征X
  - 对于每个X_i，有likelihood$P(X_i|Y)$
- 决策规则：

$$y^* = arg\max_y P(y)P(x_1,...,x_n|y) = arg\max_y P(y)\prod_i P(x_i|y)$$

Bag of Words Model词袋模型：假设文字在文本中的位置不重要，且相互独立

$$P = P(y) \prod_{i=1}^{\text{LengthDoc}} P(x_i|y)$$

- 对上式进行Log防止数字溢出
- Add-k Laplace Smoothing：平滑以防止训练集部分单词在某些类型中未出现，导致概率为零。其中$|V|$是整个词典的大小

$$P(x_i|y) = \frac{\text{count}(x_i,y)+1}{\text{sum}(\forall i, \text{count}(x_i,y)) + |V|}$$

- 未知单词：若测试单词在词典中未出现，则忽略
- Stop Words：无意义的常用词

决策树 Decision Tree：

- 组成部分：

    - Decision Node：储存数据测试的特征值，并根据该值选择分支
    - Root Node：树顶部的第一个Decision Nod
    - Branch/Sub-Tree：在树中Decision Node之间的连接
    - Layer：同一层的Decision Node
    - Leaf：叶节点，不再按照特征决策而是给出结果
    - Depth：决策树中Layer的数量
- Ockham's Razor：目标是找到最小符合数据的决策树
- ID3：类似贪心，每一步找到局部最优的特征进行分类
- Leaf Node处理：
    - 输出特定的标签
    - 计算Nearest Neighbout并输出该标签
    - 将输出输入到其它模型
- Purity的检测指标Entropy：用于衡量数据的混乱程度

$$H(S) = -\sum_{c \in C} p_c \log_2 p_c$$

- Information Gain：通过计算根据特征分类后得到了多少信息量来决定特征选取，即将分类前的Entropy减去分类后的Entropy
- 停止树增长：当标签一致、将所有特征都用完、或决策树开始过拟合时，需要停止树增长。通常有如下停止标准
    - 最大深度
    - 叶节点剩下的样本少于一定数量
    - 只剩下一个标签
- 减少过拟合：
    - PrePruning：在某些结点停止树增长，能过提早预防过拟合并减少树的大小，但可能会缺失重要信息
    - PostPruning：在构建整个树后根据证据和精度移除节点，能够提取有效信息并减少过拟合，但需要计算量更大且无法保证有效性