机器学习与数据挖掘课堂作业

特征选择

数据科学与计算机学院　17大数据与人工智能

17341015　陈鸿峥

问题 1. 针对本课程的人格分析课程大作业，你会如何进行特征选择？请描述具体做法。

解答. 无论是采用词袋模型(Bag of Words, BoW)还是TF-IDF模型生成出来的词向量维度都非常的大，要处理这么高维度的数据显然对于机器学习模型来说难度太大了。注意到原始词向量中具有大量的零元素，因此可以对这些特征进行合理选择。

分别采用今天课堂上所讲的三种特征选择思路，有以下方法：

- 过滤型方法：
  - 皮尔逊相关系数：设词向量第i维的特征为X_i，预测值（即人格类型）为Y，那么可算出每个特征与预测值之间的线性关系

    $$r_i = \frac{cov(X, Y)}{\sqrt{var(X)\, var(Y)}}$$

    取相关性绝对值大的那些特征即可（但这里只能作用于线性特征，对于词向量这种非线性特征可能不好用）。
  - 信息增益：类似于决策树的方法，选择出信息增益最大的特征子集，但这个计算量会比较大（如果词向量是上万维度的话）

    $$\mathrm{Gain}(D, A) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

- 封装型方法：
  - Las Vegas Wrapper：在循环的每一轮随机产生一个特征子集，在该特征子集上交叉验证得到准确率，多次循环后选择误差最小的特征子集作为最终解。但这种方法随机性太强，选出的特征很有可能大部分是零，对于人格分类任务来说也不是很合适。
- 嵌入型方法：
  - 正则化方法：这种方法是非常好用的，在最终的优化项上添加L1或L2范数即可，但这里需要采用回归来完成分类任务。

– 基于树的模型：这一点在第一次实验中已经采用决策树和随机森林完成，本质上就是每轮迭代选择最优划分特征。

至于无监督的Laplacian score的方法，要计算上万维数据的p近邻计算量还是相当大的，而且构造图之后还要进行大规模的矩阵运算，这个开销过于庞大，因此对于人格分类这一任务来说这种无监督选择特征的方式可能并不合适。

更优的特征工程方法可能是下节课会讲到的降维，包括PCA和LDA等。