读书报告 8.3-8.9

61518426 周之遥

·读书内容

学习第4章Naive Bayes and Sentiment Classification

这一章主要以朴素贝叶斯为例介绍了文本分类问题的各个方面。

文本分类问题细分有情感分析、垃圾邮件筛选、语言检测、话题分类等等，尽管有基于规则的方法，一般采用监督学习的方式。

朴素贝叶斯分类是简单的分类模型。做了两个简单假设，一个是用词袋BOW来表示文本，忽略单词顺序的影响。另一个是朴素贝叶斯假设，即不同单词的分布相互独立。

模型使用贝叶斯法则，在给定文本时，输出P（d|c）条件概率最大的c为分类结果。而P（d|c）则根据之前的假设分解为一连串条件概率之积，这些条件概率也是简单的频率估计，一般使用+1平滑。对于UNK直接忽略，该模型不具备处理UNK的能力。

对于单词的选取，一般会去除出现最频繁的，或者使用停用词表。

在部分情况下，二项贝叶斯比多项贝叶斯有更好的效果。

否定词也需要一定的处理，一个简单的处理方式是对否定词后的所有单词加一个NOT前缀。

在样本较少时，可以使用现成的情感词典，根据一篇文章/句子含有其中单词的数量来判断。

注意有时我们也可以不使用所有的单词作为特征，而人工筛选出一些重要特征。

对一个分类模型的评估有recall、precision、accuracy，当样本不均衡时accuracy不能表现出模型的好坏，这个时候可以使用recall和precision，F1score是一个不错的评估方式。

对于多标签问题，可以看作多个二分类问题；对于多分类问题，如果有C个类别，就有C个recall、precision、F1。可以用macro average简单取平均，这样评估等同地看待不同类别的样本。也可以用micro average，池化之后计算，这样做更看重频率高的类别。