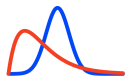


Naive Bayes Classifier

邹颖

北京大学 硕士研究生

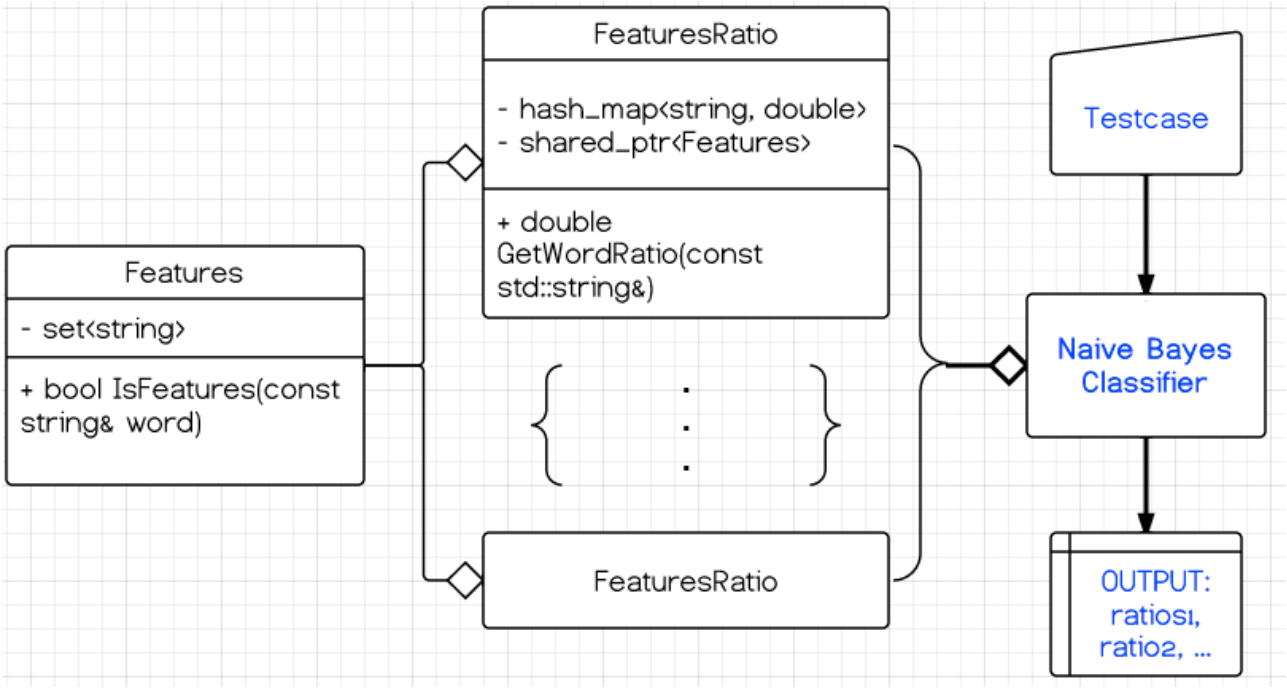
xpzouying@gmail.com
18612210096

Naive Bayes Classifier

总体

朴素贝叶斯分类器，对分类器模型进行训练后，可以根据传入测试用例的内容对其分类。

结构示意图



项目组织结构

bin/：应用程序生成结果目录，包括程序，训练模型的结果、训练过程中的日志。

data/：训练样本、特征字典存放目录。

traindata.h：存放训练样本的数据结构，从配置文件中读出每一类的训练样本文件名及其对应的类型。

features.h/cpp：从特征字典的文件中读入特征词保存在set<>中。

featuresratio.h/cpp：模型训练的统计数据。训练时，统计特征出现的次数。测试时，输出相对应特征的条件概率，如果查询特征在训练样本中没有出现，则使用拉普拉斯修正。

naivebayesclassifier.h/cpp：贝叶斯分类器，使用朴素贝叶斯进行分类。提供接口，传入测试样本，输出每一类对应的概率，输出结果按照降序排列。分类个数和分类名称在配置文件中设置。

traindata.conf：配置文件，可以设置分类的个数，训练样本所在目录、特征字典、是否进行新的训练还是使用上次训练的结果。

log.h：生成日志使用的类。

tools.h/cpp：常用的工具函数。

timer.h：计时类。

main.cpp：入口函数。

运行

可执行程序为nbc.o，保存在bin/目录下。

根据配置文件的设置，选择根据上一次训练结果进行执行还是重新模型训练，上一次的模型的训练结果以.tr为后缀名存放在bin/目录下。

运行命令为：./nbc.o -c ./traindata.conf -t ./Testcase

-c： 指定配置文件。

-t： 指示测试用例。