Assignment 2: Text Classification with Naive Bayes

Name: Min Wu.

Introduction

In this assignment, Naive Bayes were studied and to see how a non-trivial statistical model is implemented in practice using Python together with Numpy and Pandas. The common statistical experimental techniques such as the calculation of interval estimates and p-values for comparisons were studied in this assignment.

Part I. Estimating parameters for the Naive Bayes classifier and Classifying new documents

The whole data in this assignment was split the data into 80% for training and 20% for evaluation part. The function that uses the training set (80%) of documents to estimate the probabilities in the Naive Bayes model is used to classified the evaluation set (20%) of documents.

The code for Navie Bayes are attached in the Assignment2.py.

Sanity checks for Naive Bayes classifier

Sanity check 1. The probability of a positive document containing just the word "great" is about 0.00135 higher than the probability of a negative document 0.00054. Conversely,the probability of a positive document containing just the word "bad" is about 0.00018 lower than the probability of a negative document 0.00047.

Sanity check 2. The probability of the document `['a', 'top-quality', 'performance']` is 3.78e-12.

Evaluating the classifier

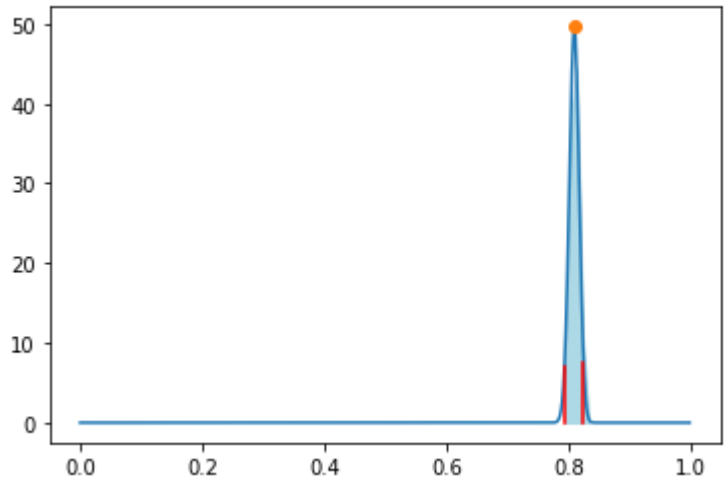
20% of documents is used to evaluate the classifier using Naive Bayes. The accuracy is about 0.81 based on the number of correctly classified documents divided by the total number of documents.

Part 2: Statistical analysis of the experimental

Computing an interval estimate for the accuracy

Compute a 95% interval estimate for the accuracy using the Bayesian method.
The number of documents in the test part is 2383. 1936 is right predicted(Nright) results and 447 is wrong predicted results (Nwrong). The accuracy is 0.81. To calculate the credible

interval using Bayesian approach with a Beta distribution. Assume it is a symmetric distribution with non specific prior a=1 and b=1. Because Beta is a conjugate prior, Bernoulli model is used, the posterior is a + Nright and b + Nwrong. The 95% credible interval is between 0.8 and 0.83 in this case.



Cross-validation

The cross-validation method is used in order to get a more reliable estimate and tighter interval. In the cross-validation, the data is divided into N parts (4-10) of equal size. Each part once becomes a test set, then the other parts form the training set. The results of the *N* different evaluations are obtained which is for the whole dataset, not just a certain test set. In this assignment, the dataset was divided into 10 parts.

95% credible interval is between 0.8 and 0.81 which is more tighter than the results from previous experiment.

Comparing the accuracy to a given target value

The results (the number of tested samples (2383) and the right results (1936)) are got from previous experiment. The results with the hypothesis classifier's accuracy 0.80 gives a *p*-value 0.13 indicating that the hypothesis accuracy is accepted at the 5% level of significance because the returned p-value is greater than the critical value of 5%.

Comparing two classifiers

Two different classifiers are generated by using different values for the smoothing parameter, the first one is based on the number of english words appeared in the

documents, the second is based on the number of english words appeared in the dictionary such as the Second Edition of the 20-volume Oxford English Dictionary contains full entries for 171,476 words in current use [1].

McNemar test and compare the two classifiers on the 20% test set.

	first correct	first incorrect
second correct	1931	5
second incorrect	10	437

The *p*-value is 0.08 for the the second one indicating that the hypothesis accuracy is accepted at the 5% level of significance because the returned p-value is greater than the critical value of 5%.

Reference:

[1] How many words are there in the English language?, *lexico* [website], https://www.lexico.com/en/explore/how-many-words-are-there-in-the-english-language, (accessed 2019)