CS 180: Introduction to Artificial Intelligence

Machine Exercise 04 - NAIVE BAYES

Introduction In this machine exercise, you will apply the Naive Bayes to solve a classification problem.

Text Document Classification One application of the Naive Bayes is the classification of text documents to a set of topics. Each document is represented by a feature vector whose attributes correspond to a defined dictionary with values 0 or 1 denoting the absence or presence of the word in the document.

Dataset The dataset is composed of 18,846 documents under 20 topics. For this machine exercise, partition the dataset into two - one for training and one for testing. Let the first 60% of the documents comprise the training set and the remaining comprise the test set. The correct label for the documents can be found inside the file *index*.

Building the Dictionary You may build the dictionary from the set of words that could be obtained from the training set. Additionally, you may obtain the stem of each word to remove suffixes like *-ing*, *-s*, *-es*, *-ed*. Alternatively, you may use available word corpora online.

Conditional Probabilities and Lambda Smoothing After building the dictionary, obtain the conditional probabilities for each word in the dictionary given each of the classes. As some of these may yield zero values, use lambda smoothing. Experiment on these values: *{ 0.01, 0.1, 0.2, 0.5, 1 }*.

Classification To classify a document, perform a *Maximum A Posteriori* Decision. Pick the classification that maximizes $p(classification|document)$.

Performance and Accuracy Since every document is labelled, you can compare its true classification with what you got using the Naive Bayes model. Construct a confusion matrix to measure the performance of your model.

Deliverables For this machine exercise, submit the following deliverables:

- All source files
- A text file containing the dictionary of words
- A README file containing instructions to build and run your program
- A documentation with complete details of your implementation, experiments and analysis

Place all deliverables in a *zip* archive with your student number (without the dash) as filename.

Deadline The deadline for this machine exercise is on **15 October 2016, 11:59pm**. Submit the archived deliverables to **kedelaspenas@up.edu.ph** with the subject **CS 180 ME4** <**Section**> <**Surname**>.

Notes

- Work on this individually.
- You cannot use any library that directly implements the Naive Bayes.
- You might need to use a library to extend the floating-point precision available for your program.
- As always, if you have any questions, consult with your instructor.

References

[1] 20_Newsgroups Dataset. Available Online: *http://qwone.com/~jason/20Newsgroups/*.