Louisa Cho
260672142

Analysis of Naive Bayes Classifier

Abstract

The program I wrote calculates the probability of a certain Jeopardy! category being geography-related using a Naive Bayes algorithm. The two classes defined for this algorithm are "contains a geography word" and "does not contain a geography word". The program reads from a database containing the names of all U.S. states, as well as all countries in the world, and determines whether a question contains one of these words. The features defined for this Bayes analysis are all the categories of Jeopardy! questions contained in the dataset.

The goal of this program is to be able to classify a question as "geography related" based on its category, using the given Jeopardy! question database as a training set.

Analysis

One thing to note initially: the runtime of this program is quite inefficient, due to the need to compare each word from the database of geography words with each word in each question (which gives $\Theta(n * m + c)$ with n as the number of questions, m as the number of geography words, and c as the number of words in each question, a comparatively small constant). It takes approximately 2 minutes to complete the program using the given datasets.

Furthermore, its accuracy is somewhat unpredictable, due to the nature of such a dataset. For example, a question may contain a geography-related word, despite it not being relevant to the answer.

Conclusions

From the results, it a classifier should be able to determine which categories are 100% geography-related, as they give a 1.0 probability of having a geography word in them. For instance, the categories "ATHLETES' COUNTRIES OF BIRTH" and "COUNTRY DEMOGRAPHICS" give a probability of 1.0, and are in fact always geography-related. However, there are some categories, such as "20th CENTURY FICTIONAL CHARACTERS", that give a probability of 1.0 but are not obviously necessarily geography-related. Therefore, there would be a fair amount of error using this classifier.