Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 5

Due: February 22, 2024(100 points)

YOUR NAME
YOUR GIT USERNAME
YOUR E-MAIL

Naïve Bayes, Bayes Rules

The original performance of acoustic classification for Parkinsons Disease leverages speech recordings from controlled subject responses from variety of questions. The task in the competition was to detect whether or not a person X had Parkinsons disease from a sampling of data. As of 2018, the state of the art classifiers have achieved 90% correct classification on a held out dataset, both for subjects who had Parkinsons and those who did not (at equal rates). So, when classifier Y sees person X, it works correctly 90% of the time.

1. Let's say that we run a clinic. This clinic leverages this classifier, which has 90% accuracy. Also, let us say that we know that our current patient load is that 10% of the population have Parkinsons and 90% of the population do not. Let's also say that we're seeing patient X, and the classification algorithm has detected that they have Parkinson's disease. **What's the probability that indeed X has Parkinson's disease?**

Come up with the numerical solution, and show your written work.

The Sum of Conditional Probabilities

In class, we reviewed three main rules in Bayesian probability inference:

- Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Bayes Theorem:

$$P(A|B)P(B) = P(B|A)P(A)$$

- Total Probability:

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

A well-known outcome of the three sets of rules is the fact that the sum of all the conditional probabilities equals one.

2. Prove that:

$$\sum_i P(A_i|B) = 1$$

Mining Reviews in Small-ish Datasets

ATTN: The below section is *not* due with homework 5 per teams thread. You can get a head start, but *only questions 1 and 2* need to be turned via Gradescope on February 22.

The question in this section is best started with the starter kit Colab. You can copy and past the cells to a `*.py` Python3 file or or simply use the Colab in your own Google Drive.

We'll process datasets that are small enough that they can fit into CPU memory. We'll review the Amazon purchasing reviews dataset. Amazon has collaborated with Universities to aggregate an extensive set of buyer reviews data. Most of the datasets (including the subcategory datasets) cannot fit into CPU RAM, and an even smaller subset of machine learning algorithms can actually process them.

With this dataset, you will need to do the following for question number 2. Typically we would split the dataset into training / validation / testing dataset, but we will only split into training / testing dataset in this assignment.

3. With the Amazon magazines dataset, do the following. (Feel free to use the `scikit-learn` library).

    - Plot a histogram of the data with the number of positive and negative reviews.
    - Balance the data so that you are training with equal probabilities, $P(\text{Liked}) = P(\text{Not Liked}) = 0.5$. We will use the original distribution for the evaluation dataset.
    - Use machine learning models to predict whether or not a review as written will result in a "Liked" rating. Try changing different parameters, including the maximum vocabulary size. Also try normalization.

- Try out Naïve Bayes, Decision Trees, Random Forests, and Logistic Regression machine learning algorithms
- Print out your best accuracy, precision, and recall numbers for each algorithm.

How do they compare? Which algorithms are overfitting?

Tips and Tricks

You can preprocess the data using files in `cs6220hw5.py`, which you can download here. This will remove *stop words*, punctuation, and ensure that the texts are all lowercase. Caution that this implementation is exceedingly slow and un-optimized. (Extra credit to anyone who implements a faster version.)

Featurize your inputs to the algorithm with `CountVectorizer()`. Make sure to set the maximum vocabulary size / number of features, `max_features`.