Lab 4 Naïve Bayes

COMP4901K and MATH 4824B

Fall 2018

Prerequisites

- You need to have some background knowledge about Naïve Bayes (NB). If not, you can check out: Lecture 8 and `https://en.wikipedia.org/wiki/Naive_Bayes_classifier`
- You need to install the NLTK, Pandas, Numpy, Scipy, and scikit-learn packages:

```
pip3 install --upgrade nltk pandas numpy scipy scikit-learn

python
>>> import nltk
>>> nltk.download('punkt')
>>> nltk.download('stopwords')
```

1 Assignment

You need to download the following file(s) from canvas: `lab4_skeleton.zip`, including:

```
lab4_skeleton
├── lab4_skeleton.py
└── data
    ├── answer.csv
    ├── test.csv
    └── train.csv
```

Q1 Preprocess the training set.

1. Use pandas to read data from `data/train.csv`
2. Use nltk to tokenize text into words
3. Turn words into Bag-of-words representation.

Q2 Write code to compute the probabilities.

1. Design the Laplace Smoothing
2. Compute $P(Y = y_i)$
3. Compute $P(x_j|Y = y_i)$

Q3 Write code to predict labels

1. Compute $P(Y = y_i) \prod_j^V P(x_j|Y = y_i)$
   (hint: $P(Y = y_i) \prod_j^V P(x_j|Y = y_i) = \exp(\log(P(Y = y_i)) + \sum_j^V \log(P(x_j|Y = y_i))))$
2. Compute $P(Y = y_i|x_1, \cdots x_V)$.
3. Choose labels with the highest probability

2 Submission

You need to submit three files, program output, `submission.csv`, and your python script. After you finished the assignments, make sure you include the header information in the beginning of your code

```
# author: Your_name
# student_id: Your_student_ID
```

Copy all the program output in to a text file named `StudentID_lab4_output.txt`, and submit your .csv file named `StudentID_lab4.csv` and python script solution named `StudentID_lab4.py` to Canvas.