Assignment 2

Due on November 22, 2019 (23:59:59)

Click here to accept your Assignment 2

Instructions. There are two parts in this assignment. The first part involves a series of theory questions and the second part involves coding. The goal of this problem set is to make you understand and familiarize with Naive Bayes algorithm.

Part I: Theory Questions

MLE

1. Assume that you have a data consisting of $x_1, x_2, ..., x_m$ where each x_i represents a single real value, which means you have m instances in data and each instance has an single real-valued attribute. Assume also that the given data has random uniform distribution between $-w$ and w. You are expected to find the maximum likelihood estimate of w with respect to the given data.
   - Specify a likelihood function $F(w)$.
   - Specify the maximum likelihood estimate for w. Consider your answer based on the likelihood function you state.
   - Assume that this time you are given a labelled data (x_i, y_i), where y_i is 1 or 0. Remember that a generative classifier will try to model $P(y)$ and $P(x|y)$. Define an example dataset the generative classifier utilizing the model you defined above for each $P(x|y)$ could not perform well on.
   - Remember that a discriminative classifier will try to model $P(y|x)$. State that whether you can classify the labelled data given in previously part using such a discriminative classifier or not. If your answer is answer is yes, then please also show that what your suggested classifier looks like.
2. Fill the blanks with T (True) or F (False) for the statements given above:
   - Maximum likelihood estimation provides not only point estimation, but a distribution information of the parameters estimated. (_)
   - Maximum likelihood and Bayesian approachs for parameter estimation perform well with low-dimensional dataset with many training examples while their performance is bad on high-dimensional dataset with few training examples. (_)

Naive Bayes

x	y	z	C
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

1. Consider that you are given the dataset in the table above consisting of boolean variables x, y and z and a single boolean output variable C. Suppose that the Naive Bayes classifier is going to be used.

   - Specify the value of $P(C = 1|x = 1, y = 1, z = 0)$. Show your solution step by step.
   - Specify the value of $P(C = 0|x = 1, y = 1)$. Show your solution step by step.

   Now suppose that the Joint Naive Bayes classifier is used for the options below:

   - Specify the value of $P(C = 1|x = 1, y = 1, z = 0)$. Show your solution step by step.
   - Specify the value of $P(C = 0|x = 1, y = 1)$. Show your solution step by step.

2. Assume that you have three variables, which are A, B and C.

   - Suppose that you have the following informations $P(C|A) = 0.7$ and $P(C|B) = 0.4$. State that whether you can compute $P(C|A, B)$ with the informations given previously or not. Besides show your solution if you can and explain the reason if you can not.
   - Suppose that besides two informations above, $P(A) = 0.3$ and $P(B) = 0.5$ informations are given. State that whether you can compute $P(C|A, B)$ with the informations given previously or not. Besides show your solution if you can and explain the reason if you can not.
   - Finally assume that you have only informations, which are $P(C, A) = 0.2$, $P(A) = 0.3$ and $P(B) = 1$. State that whether you can compute $P(C|A, B)$ with the informations given previously or not. Besides show your solution if you can and explain the reason if you can not.

PART II: Detection of Fake News

In this part of the assignment, you will try to determine whether a headline is real or fake news[1] (see Table 1). You will implement a Naive Bayes classifier and verify its performance on Fake News dataset [1]. As you learned in class, Naive Bayes is a simple classification algorithm that makes an assumption about the conditional independence of features, but it works quite well in practice.

id	title	author	text	label
1	FLYNN: Hillary Clinton, Big Woman on Campus - Breitbart	Daniel J. Flynn	Ever get the feeling your life circles the roundabout rather than heads in a straight line toward the intended destination?...	0
12773	New Leaked Clinton Emails Came from the Devices of Anthony Weiner	Dean Daniels	New Leaked Clinton Emails Came from the Devices of Anthony Weiner 6 shares by Dean Daniels...	1

Table 1: Some real/fake examples from the dataset

Dataset
Fake News is a dataset provided to determine when an article is fake or real. It includes the following features:

- **id:** unique id for a news article
- **title:** the title of a news article
- **author:** author of the news article
- **text:** the text of the article, could be incomplete
- **label:** a label that marks the article as potentially unreliable
    1. 1: unreliable
    2. 0: reliable

Training and test dataset will provided later and be announced from Piazza group.

[1]This assignment is adapted from https://www.teach.cs.toronto.edu// csc411h/winter/projects/proj3/

Approach

1. **Part 1: Understanding the data**
   You will be predicting whether a headline is real or fake news from words that appear in the headline. Is that feasible? Give 3 examples of specific keywords that may be useful, together with statistics on how often they appear in real and fake headlines.

2. **Part 2: Implementing Naive Bayes**
   You will represent your data with listed approaches and use them to learn a classifier via Naive Bayes algorithm. You have to implement your own Naive Bayes algorithm.

   - Features: You will use Bag of Words (BoW) model which learns a vocabulary from all of the documents, then models each document by counting the number of times each word appears. You will use BoW with two options:
     - Unigram: The occurrences of words in a document(frequency of the word).
     - Bigram: The occurrences of two adjacent words in a document.

   Note: You should compute the log probabilities to prevent numerical underflow when calculating multiplicative probabilities.
   You may encounter words during classification that you havent during training. This may be for a particular class or over all. Your code should deal with that. Hint: You can use Laplace smoothing.
   You have to use a dictionary for BoW representation. You can implement your own method to obtain BoW model or you can use Count Vectorizer function (`https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html`).

3. **Part 3:**

   (a) Analyzing effect of the words on prediction

   - List the 10 words whose presence most strongly predicts that the news is real.
   - List the 10 words whose absence most strongly predicts that the news is real.
   - List the 10 words whose presence most strongly predicts that the news is fake.
   - List the 10 words whose absence most strongly predicts that the news is fake.

You can narrow down your dictionary by choosing specific words for real and fake news. In other words, your classification results can be improved by selecting a subset of extremely effective words for the dictionary. TF-IDF (`https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html`) and Information Theory are good places to start looking. Reimplement the part2 and see the effect of using specific words on the task.

State how you obtained those in terms of the the conditional probabilities used in the Naive Bayes algorithm. Compare the influence of presence vs absence of words on predicting whether the headline is real or fake news.

(b) Stopwords
You may find common words like a, to, and others in your list in Part 3(a). These are called stopwords. A list of stopwords is available in sklearn here. You can import this as follows:
`from sklearn.feature_extraction.text import ENGLISH_STOP_WORDS`
Now, list the 10 non-stopwords that most strongly predict that the news is real, and the 10 non-stopwords that most strongly predict that the news is fake.

(c) Analyzing effect of the stopwords

Why might it make sense to remove stop words when interpreting the model? Why might it make sense to keep stop words?

4. **Part 4: Calculation of Accuracy**
You will compute accuracy of your model to measure the success of your classification method:

$$\textbf{Accuracy} = 100 * (\frac{\textbf{number of correctly classified examples}}{\textbf{number of examples}}) \quad (1)$$

Submit

You are required to submit all your code (*all your code should be written in Jupyter notebook* long with a report in ipynb format (should be prepared using Jupyter notebook). The codes you will submit should be well commented. Your report should be self-contained and should contain a brief overview of the problem and the details of your implemented solution. You can include pseudocode or figures to highlight or clarify certain aspects of your solution. Finally, prepare a ZIP file named **name-surname-pset2.zip** containing

- report_and_code.ipynb (Jupyter notebook file containing your report and code)

The ZIP file will be submitted via Github Classroom. Click here to accept your Assignment 2

NOTE: To enter the competition, you have to register kaggle in Class with your department email account. The webpage of the competition will be announced later. Top 5 assignment will earn extra points.

Grading

- Code (50): Part1: 5p, Part2: 25p, Part3: 15p, Part4: 5p
- Report(50): Theory part: 20p, Analysis of the results for prediction: 30p.

  Notes for the report: Preparing good report is important as well as your solutions! You should explain your choices (Unigram, Bigram or both of their use for Bow, or constraints on data) and their effects to the results.

Late Policy

You may use up to four extension days (in total) over the course of the semester for the three problem sets you will take. Any additional unapproved late submission will be weighted by 0.5. You have to submit your solution in (rest of your late submission days + 4 days), otherwise it will not be evaluated.

Academic Integrity

All work on assignments must be done individually unless stated otherwise. You are encouraged to discuss with your classmates about the given assignments, but these discussions should be carried out in an abstract way. That is, discussions related to a particular solution to a specific problem (either in actual code or in the pseudocode) will not be tolerated. In short, turning in someone else's work, in whole or in part, as your own will be considered as a violation of academic integrity. Please note that the former condition also holds for the material found on the web as everything on the web has been written by someone else.

References

[1] This reference is hidden due to the obvious reasons.