COMP 550 Natural Language Processing

Assignment #1 - Christopher Zheng - 260760794

christopher.zheng@mail.mcgill.ca

September 25, 2019

1. Q1

   (a) Every student took a course.

   The ambiguity is the type/token ambiguity in the sentence's semantics. One interpretation can be: all students took only one course, and some other interpretation can be each student took one course, respectively (might be different courses). The cause is a course. Some knowledge of common-sense reasoning might help in this case by reading more contexts beforehand which will make this issue less likely to happen.

   (b) John was upset at Kevin but he didn't care.

   The ambiguity is the pronoun ambiguity in the sentence's semantics. One interpretation can be: Kevin didnt care, and another interpretation can be that John didnt care. Specifically, the cause is he. Knowledge about the opposite meaning of two parts of this sentence as well as the context or real-world knowledge about the relationship between John and Kevin may help.

   (c) Sara owns the newspaper.

   The ambiguity is the lexical ambiguity in the sentence's semantics. One interpretation can be: Sara owns one specific piece of newspaper, and another interpretation can be that Sara own a newspaper company, e.g., Wall Street News. The cause of this is the multiple meanings of "the newspaper". Knowledge about the description of this newspaper can be helpful or knowledge about Sara's career might also help.

   (d) He is my ex-father-in-law-to-be.

   The ambiguity is the structural ambiguity in the sentence's morphology. One interpretation can be: he just got married (ex "father-in-law-to-be"). One interpretation can be: he is on the brink of being divorced ("ex-father-in-law" to-be). Another interpretation is that he just finished his engagement (ex "father-in-law-to-be"). The cause of the ambiguity is the complicated morphological structure of "ex-father-in-law-to-be". Basic knowledge of engagements is helpful and knowing the exact narrator also helps.

   (e) ttyl ;) [text message]

   The ambiguity is the orthographic ambiguity in the sentence's word spelling. One interpretation can be talk to you later. Another interpretation can be that this is not even a grammatical sentence. The cause originates from "ttyl" which is seemingly not a word but capital letters of a message sentence. Knowing this is a text message in the first place helps quite a lot. Plus, some knowledge in Internet slangs or youth culture is helpful for analyzing texts like this.

2. Q2

In terms of the Naive Bayes classifier, we utilize the categorical distributions for the prior and the features in this case. From the result we gained in the lecture slides, we can estimate each class's probability conditioned on the features as follows:

$$P(y|f_1, f_2, ..., f_{n-1}, f_n) \rightarrow P(y)P(f_1, f_2, ..., f_{n-1}, f_n|y) \rightarrow P(y)\prod_i P(f_i|y) \tag{1}$$

where $y = c$, the class, and f_i's are all the features in the feature space ϕ. The first arrow follows the Bayes' Rule, and the second arrow follows the a priori of Naive Bayes that all features are mutually independent. To maximize the likelihood, we need to figure out the class that produces the largest number of Equation 1. Since the differences among the values of all the classes are not always obvious, we naturally take the logarithm and transform 1 into:

$$logP(c) + \sum_i logP(f_i|y) \tag{2}$$

where $y = c$, the class, and f_i's are the features.

On the contrary, the Logistic Regression has the distribution as we have proved in class that

$$P(y|\{f_1, f_2, ..., f_{n_1}, f_n\}, \theta) = \frac{1}{Z} e^{a_1 f_1 + a_2 f_2 + ... + a_n f_n + b} \tag{3}$$

where the hyper-parameter $\theta = \{a_1, a_2, ..., a_n, b\}$, and f_i's are all the features in the feature space ϕ. To maximize the conditional likelihood

$$L^{LR}(\theta) = \prod_c P(y|f_1, .., f_n; \theta) = \prod_i P(y)P(f_i|y)/P(f_1, ..., f_n) \tag{4}$$

$$\rightarrow P(y)\prod_i P(f_i|y) \tag{5}$$

The arrow follows that the probability distributions of all features are categorical and fixed a priori, and thus the probability values are constant. Similar to what we do to the Naive Bayes classifier, we take the logarithm and transform 5 into

$$logP(y) + \sum_i logP(f_i|y) - logP(f_1, ..., f_n) \tag{6}$$

Since the hyper-parameter θ is arbitrary, there exists a class c such that $logP(y = c) + \sum_i logP(f_i|y) - logP(f_1, ..., f_n)$ equals $logP(c) + \sum_i logP(f_i|y)$ (Equation 2). Therefore, the Naive Bayes classifier is linear just like the Logistic Regression classifier. □

3. Sentiment Analysis Report

1. Introduction

In this work, we experiment on a modern movie review dataset to evaluate the performance of three common classifiers (Logistic Regression, Support Vector Machine with linear kernel, and Naive Bayes) by letting them classify comments with attitudes. We also briefly examine to what extent our various pre-processing strategies can boost the three's classification accuracies.

2. Experiment Procedure

On the pre-processing stage, we generate unigrams and remove stop words preset by NLTK. (We later improve this step by refining this set of stop words.) Then, we shall accordingly choose to stem, lemmatize the tokens or simply do nothing. Next, we delete infrequent words, which may also prevent us from non-English words. Finally, we vectorize the filtered tokens and put them into a multi-dimensional numpy array for training.

For training, we shuffle and split the input into a training set and test set w.r.t a reasonable ratio. Then, we feed the training data into the classifiers and evaluate them on the test set. The result of accuracy is as follows. (Please note that the following result is not complete and only the averaged best scores after parameter-tuning and 5-fold cross-validation are displayed.)

Classifier	stopword	revised stopword	stem	lemma	infreq removal	revised stop/stem /lemma/infreq rm	revised stop /infreq rm	revised stop/ stem/infreq rm	revised stop/ lemma/infreq rm
Random	0.50586	0.49648	0.49976	0.49835	0.51242	0.50164	0.49460	0.50023	0.49132
Log. Reg.	0.70464	0.73183	0.65635	0.66338	0.66994	0.71729	0.71495	0.72011	0.71917
SVM	0.70604	0.73277	0.66526	0.66920	0.67229	0.71589	0.71401	0.72386	0.72620
N. Bayes	0.70588	0.72823	0.65588	0.69092	0.69995	0.75293	0.74543	0.76699	**0.77449**

The key parameters and their ranges in our experiment are listed:

(a) Infrequent words: words that appear less than 2 to 50 times in the whole corpus.

(b) Revised stop words: NLTK stop words excluding those that may reflect attitudes. (Manually scrutinized)

(c) Train-test split ratio: 70% to 90% for training and 30% to 10% for testing, respectively.

(d) K-fold: the times of cross-validation. We choose k=4 or 5 since the training time is overly long above 5.

3. Analysis and Conclusion

We analyze the pros and cons of three classifiers on our dataset. Logistic Regression can have low variance and handle tokens being correlated but may introduce high bias if our dataset itself is poor in quality. SVM is commonly used in text classification (high-dimensional space) but is very memory-intensive and inefficient to train, which renders SVM not suitable for a large dataset. Naive Bayes converges more quickly than discriminative models (e.g., Logistic Regression) do under the conditional independence assumption which is also its drawback since such an independence cannot be always guaranteed. We, however, uphold this assumption in experiments.

		Predicted	
		Yes	No
Actual	Yes	842	199
	No	282	810

In conclusion, the best model we pick is the Naive Bayes classifier fed by processed dataset where we revise the NLTK stopwords, lemmatize the tokens and remove infrequent tokens. It achieves an accuracy of **0.77449** and its confusion matrix is shown above. The relative success of this model may depend on the independence of tokens in the provided dataset and appropriate data pre-processing.

4. References

Jackie Cheung. *Fall 2019 - COMP 550 - Lecture Slides - Lecture 3.*

Bharath Hariharan. *Why Naive Bayes a linear model* on Quora.

Jianqing Fan, Yingying Fan, and Yichao Wu. *High-dimensional Classification.*

Bo Pang and Lillian Lee. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales.*