Assignment 1

Yabo Ling

September 26, 2019

1. **Answer to question 1:**

   (1) This is a scope ambiguity. a course specifically causes this ambiguity. The first meaning of this sentence is that for every student, there is a course, and each student took a different course. The second meaning of this sentence is that there is one particular course which is taken by every student. The ambiguity is caused by semantics. The first-order logic of these two meanings are represented as:

   $$\forall x.student(x) \to (\exists y.course(y) \land took(x, y))$$

   $$\exists y.course(y) \land (\forall x.student(x) \to took(x, y))$$

   Usually, for human and machine, we address this ambiguity by giving a specific course name, like Every student took a course named Natural Language Processing, or giving a definition, like Every student took a different course.

   (2) This is an anaphora unclear ambiguity. he specifically causes this ambiguity. The first situation is that he refers to John. So the meaning is that John was upset at Kevin but John didnt care. The second situation is that he refers to Kevin. So the meaning is that John was upset at Kevin but John didnt care. The ambiguity is caused by semantics.
   Usually, for human and machine, we address this ambiguity by replacing he with specific person name, like John or Kevin. Human and machine will be more clearly about this sentence.

   (3) The ambiguity is caused by the newspaper. the newspaper could refer to the real paper, and could also refer to the newspaper company. So the first meaning of this sentence is that Sara has the paper which is written news. The second meaning is that Sara owns the newspaper office. The different semantics of the word – newspaper caused this ambiguity.
   For human, in order to disambiguate, we could need some contextual knowledge, like Sara owns the newspaper. But it went bankrupt
   For computer, we could give more specific definition, like Sara owns the newspaper office or Sara owns the todays newspaper.

   (4) The reason of this ambiguity is that ex and to be appear in a sentence at same time. ex-father-in-law means a former father-in-law in a relationship. father-in-law-to-be means a future father-in-law in a relationship. It is a syntax ambiguity. For human and machine, there are two methods to address this ambiguity. Firstly, we can delete ex or to be to make the meaning more clearly. Second, we can add some contextual knowledge to have a clear meaning, like, My ex-boyfriend is John, and this is his father. So he is my ex-father-in-law-to-be.

(5) This ambiguity is caused by abbreviation. The most obviously meaning is Talk to your later, usually use at the end of online conversation. However, there are also some other meanings, like, Thanks to your letter, etc. For human and machine, in order to disambiguate, we can expand this sentence like talk to you later, or add some contextual information. For example, I have to go, ttyl ;)

2. **Answer to question 2**

In Navie Bayes model, we assume that the features are independent of each other given given the class. So we can calculate the probability of each given class as:

$$P(y = c|F_1, F_2, \ldots, F_n) = \frac{P(F_1, F_2, \ldots, F_n|y = c) \cdot P(y = c)}{P(F_1, F_2, \ldots, F_n)}$$

$$= \frac{P(y = c) \cdot \prod_i P(F_i|y = c)}{P(F_1, F_2, \ldots, F_n)}$$

Therefore, we can know:

$$P(y = c|F_1, F_2, \ldots, F_n) \propto P(y = c) \cdot \prod_i P(F_i|y = c)$$

In order to simplify calculations, we are consider the log probability rather than the probability.

$$logP(c) + \sum_i logP(F_i|y = c)$$

Now, We suppose F_i has r different values, from 1 to r. And we add an indicator function $I(F_i = r)$ into the above expression, when $F_i = r$, $I(F_i = r) = 1$, otherwise $I(F_i = r) = 0$. Therefore, the expression will be:

$$logP(c) + \sum_i \sum_r I(F_i = r)logP(F_i = r|y = c)$$

As we know, Naive Bayes model's parameters are the prior probabilities and the probability of feature given class. So we could re-parametrize so that the parameters are the log prior probability and the log probability of the feature given class. Assume the log prior probability $logP(c)$ as b_c, and the log probability of feature given class $logP(F_i = r|y = c)$ as $w_{c,ir}$.
Then the only remaining term in the expression is $I(F_i = r)$, denoted as f_{ir}.
Therefore, the score for the class c becomes:

$$b_c + \sum_i \sum_r f_{ir} w_{c,ir}$$

which is a linear function of the feature f_{ir}

3. **Report of question 3**

i) problem setup

This is a binary classification task about sentiment Analysis. We set the label of positive reviews is 1, and the label of negative reviews is 0. Then I divide the data set as training data (80%) and test data (20%). The goal is to train some different models by given training data with labels, then predict the labels of test data by just given test data without labels. Finally, we calculate the accuracy, precision, recall, F1-score and confusion matrix to compare these models.

ii) experimental procedure

Firstly, I processed the data. According to the movie review dataset, there are 5331 positive data and 5331 negative data. Therefore, I randomly divided these data into two groups: training data (8519 data) and test data (2133 data). Then, I put every letter in lowercase and removed stopwords, one or two-letter words and words contain digitals and words, like 2-year. Secondly, I have implemented 7 models to be compared which are the random baseline, logistic regression, support vector machine, nave bayes, k-nearest neighbors, decision tree and random forest classifier. Secondly, I have implemented 7 models to be compared which are random baseline, logistic regression, support vector machine, nave bayes, k-nearest neighbors, decision tree and random forest classifier.

iii) I tried a range of parameters, for example, alpha in nave bayes model, penalty in logistic regression model, and so on.

iv) results and conclusions

The results chart was given as the following:

Table 1: Evaluation

model	accuracy	precision	recall	f1-score
Random	0.50	0.49	0.51	0.50
Logistic Regression	0.76	0.76	0.76	0.76
SVM	0.76	0.77	0.77	0.77
Navie Bayes	0.74	0.74	0.74	0.73
KNN	0.51	0.50	0.51	0.49
Decision Tree	0.64	0.64	0.64	0.64
Random Forest Classifier	0.69	0.69	0.69	0.69

As we can see, the performance of logistic regression and support vector machine is better than other models, which are the same accuracy: 0.76. Let us see the confusion matrix of these two models.

	Logistic regression	SVM
Confusion Matrix	$\begin{bmatrix} 803 & 287 \\ 218 & 825 \end{bmatrix}$	$\begin{bmatrix} 803 & 287 \\ 215 & 828 \end{bmatrix}$

The true positive(TP) and the false negative(FN) are the same amount, 803 and 287 respectively. As for the negative class, the number of the true negative of SVM is greater than the logistic regression. Therefore, we can assume that SVM is the best performing model.