CMPT 413/825 Natural language processing

Homework 1, Fall 2020

Due: Sep 30th, 2020

Instructions Answers are to be submitted by 11:59pm of the due date as a PDF file through Coursys. Go to Coursys, select the **Homework 1** activity and submit your answer as **answer.pdf**. This assignment is to be done individually.

Question 1 [12 pts]

Suppose we have collected the following Bigram and Unigram frequencies from a text corpus. Also assume that all other observed counts are 0. In the bigram table, rows represent w_{i-1} and columns represent w_i : e.g. C(computer, keyboard) = 2. The total number of words in our vocabulary V=24.

	computer	keyboard	monitor	store
computer	0	2	4	4
keyboard	1	0	0	1
monitor	0	1	1	1
store	2	0	0	0

Table 1: Bigram Frequencies

computer	keyboard	monitor	store
10	3	5	6

Table 2: Unigram Frequencies

Consider the following sentence fragment S: “I shopped at the computer _____”
You need to determine whether the sentence is more likely to end with “computer store” or “computer monitor”.

a. Compute the raw bigram probabilities for the candidate words {store, monitor} to complete the sentence S, i.e. P(store|computer) and P(monitor|computer). Is one word more likely than the other, and if so which one? [4 pts]
b. Compute the smoothed bigram probability of the candidate words {store, monitor} using Add-one Smoothing. Is one word more likely than the other, and if so which one? [4 pts]
c. Compute the smoothed bigram probability of the candidate words {store, monitor} using Jelinek-Mercer Smoothing as a mix of unigram and bigram probabilities. Use $\lambda_1 = 0.6$ as the bigram weighting factor. Note: $\sum_i \lambda_i = 1$. Also there is no need to consider unigram smoothing (P_{JM}). Is one word more likely than the other, and if so which one? [4 pts]

Question 2 [10 pts]

We will build a log-linear model that generates a probability distribution p(pos | word) using a log-linear model. Given a word, we want to predict the part-of-speech (pos) tags for that word. The variable "pos" can take any one of three values, *A* (article), *N* (noun), *V* (verb). The variable "word" can be any value from a set S of possible words. Assume that the set S contains the words *a*, *bear*, *eats*, as well as additional words (i.e., |S| > 3). The distribution should give the following probabilities:

p(A | a) = 0.7
p(N | bear) = 0.7
p(V | eats) = 0.7
p(A | word) = 0.4 for any word other than *a, bear or eats*
p(N | word) = 0.4 for any word other than *a, bear or eats*
p(V | word) = 0.2 for any word other than *a, bear or eats*

Values for probabilities like p(N | a), p(V | a), p(A | bear), p(V | bear), p(A | eats), p(N | eats) are left undefined and could take any values such that Σ_{pos} p(pos | word) = 1 is satisfied for word = *a*, *bear*, *eats*, or any other word in S.

a. Define the features for a log-linear model that can model this distribution p(pos | word) perfectly. Each feature should be an indicator function: i.e., each feature f_j (x, y) can take only the values 0 or 1 depending on the values of x and y. Your model should make use of as few features as possible. [3 pts]
   The first feature is given as an example:
   f_1(word, pos) = 1 if word = *a* and pos = *A*, 0 otherwise

b. Write an expression for each of the probabilities
   p(A|cat)
   p(N|walks)
   p(A|bear)
   as a function of the parameters in your model. (Assume that these words are members of the set S.) [3 pts]

c. What value do the parameters in your model take to give the distribution described above? [4 pts]

Question 3 [15 pts]

	Document	Sentence	Class
Training	1	This is my book	statement
	2	They are novels	statement
	3	Have you read this book	question
	4	Who is the author	question
	5	I like the story of the book	statement
	6	What is your favourite book	question
	7	What are the characters	question
Test	8	What is the title of the book	?

Consider the above table with documents classified either as ‘question’ or as ‘statement’. Find the class of the document (d8) in the Test set, i.e. “What is the title of the book”. Use a multinomial naive Bayes classifier. Use Add-one (Laplace) smoothing for missing/unknown words in a class, using the following equation:

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}(count(w,c)+1)} = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right)+|V|}$$

Question 4 [7 pts]

We train two different clinical text classification models from medical reports of patients which can predict if a tumour is “malignant” or “benign”. Our next task is to evaluate the two models and select one of them as our final classifier. Below are two tables with Ground truth and Predicted number of cases for each model. Will accuracy be a sufficient evaluation metric to choose one of the models? If not, what other metric can be used for evaluation? Justify your answer.

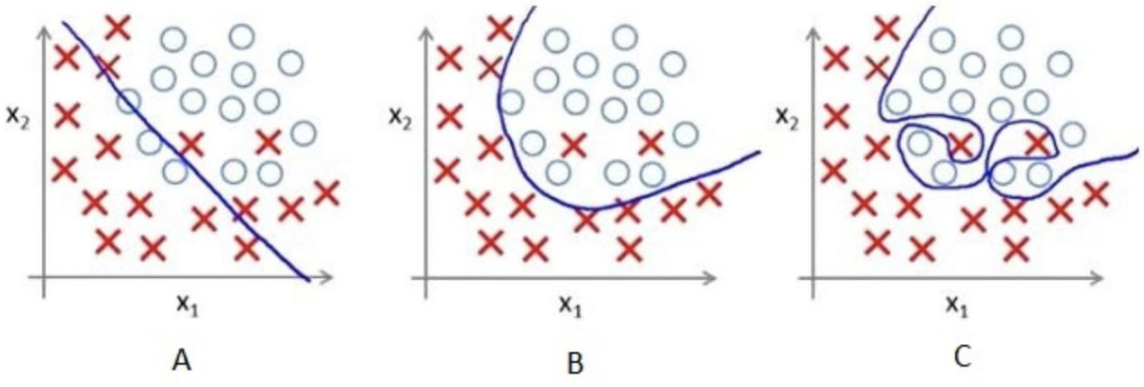
		Truth	
		Malignant	Benign
Predicted	Malignant	5	0
	Benign	5	990

Table 1: Model 1

		Truth	
		Malignant	Benign
Predicted	Malignant	7	2
	Benign	3	988

Table 2: Model 2

Question 5 [6 pts]



Consider we have trained three different logistic regression models on a restaurant review dataset to predict whether the review is positive or negative. The above three scatter plots (A,B,C) show the decision boundaries for those three models.

a. What do you conclude after seeing the visualization?
   Justify your answer. [3 pts]
   i. The training error in the first plot is maximum as compared to the second and third plot.
   ii. The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
   iii. The second model is more robust than first and third because it will perform best on unseen data.
   iv. The third model is overfitting more as compared to first and second.
   v. All will perform the same because we have not seen the testing data.

   A) i and ii
   B) i and iii
   C) i, iii and iv
   D) v

b. Suppose, above decision boundaries were generated for the different values of regularization. Which of the above decision boundaries shows the maximum regularization?
   Justify your answer. [3 pts]
   A) A
   B) B
   C) C
   D) All have equal regularization