Sheetal Parikh
Problem Set – Module 7
605.744.81 Information Retrieval
Fall 2021

1.) In your own words explain the following concepts and mention how they are relevant to text classification. Each response should not be more than a few sentences.

(a) bias-variance tradeoff

The bias-variance tradeoff concept explains the importance of having a good balance between bias and variance demonstrating why there isn't a universally optimal method that can be used to model all types of data. Our goal in text classification is finding a classifier that can best organize the text documents into one or more categories. It will be unlikely a model can perfectly classify the data so it's important to understand the type of text classification problem and the data, so you know whether it's better to have higher variance or bias. Understanding the data helps you determine what model would give you the best balance between bias and variance so that the total error is minimized, and you can avoid underfitting and overfitting the data as much as possible.

References:
Pages 308-314: https://nlp.stanford.edu/IR-book/pdf/14vcat.pdf
https://towardsdatascience.com./understanding-the-bias-variance-tradeoff-165e6942b229

(b) feature selection

Feature selection is the process of selecting a subset of relevant features from the features that are present in the training set. This subset is then the only features that are used in text classification. Feature selection is an important concept in text classification as it should remove the redundant or irrelevant features, making the process of training and applying a classifier more efficient as the size of the data would be reduced. Also, a smaller dataset means that time required for training the model may also decrease. Furthermore, classification accuracy may also improve as the data that may have mislead the model in training has been removed and now we only have the data that we believe is most representative of the text classification problem.

References:
Pages 212 – 272: https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

(c) linearly separable

Data is linearly separable if when graphed, there is a hyperplane that can perfectly separate the data. This concept is important in text classification because knowing whether data is linearly separable can help determine what machine learning model to apply to the data for the given problem. Linearly separable data can be considered a more "simple" dataset as many machine learning techniques can be used. The support vector machine has become a very popular technique for text classification and is based on finding the decision boundary to separate different classes to maximize the margin. If we know that the data is not linearly separable,

then techniques such as the soft margin or kernel trick can be used when applying the SVM to a dataset.

References:
Pages 331 & 327: https://nlp.stanford.edu/IR-book/pdf/15svm.pdf
Pages: 304-305: https://nlp.stanford.edu/IR-book/pdf/14vcat.pdf

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe

(d) microaveraging

Microaveraging is a method that can be applied to machine learning metrics in which we give equal weight to each individual classification decision. Because each individual decision is included, microaveraging favors topics with a large number of text examples and so it weighs the metric towards the larger class/topic. Microaveraging is an important concept in text classification because it is a method that can be applied to a reported metric depending on your goal. For example, if you have an imbalanced dataset and your goal is to maximize the number of correct predictions that the classifier makes for the majority class (which a very common text classification goal), microaveraging is a good method to apply to F-1 Score, precision, recall etc.

References:
Lecture 7C
Pages 280-284: https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf
https://datascience.stackexchange.com/questions/36862/macro-or-micro-average-for-imbalanced-class-problems

2.) List three differences between Multinomial Naïve Bayes and Bernoulli Naïve Bayes.

a.) When classifying a text document, the Bernoulli model only uses binary occurrence information, meaning that the Bernoulli model only considers whether a word occurs in the document. However, the multinomial model keeps track of whether a word occurred in a document as well as the term frequency in the document.

b.) Both methods have different approaches of calculating the conditional probability, $P\hat{}(t|c)$, which is the relative frequency of a term in a document belonging to a specific topic/class. The Bernoulli method is calculated by taking the fraction of documents of topic c_j in which word the term t appears. However, the Multinomial method is calculated by taking the fraction of times in which word w appears across all documents of topic c_j.

c.) Both methods have different approaches of how nonoccuring terms are used in text classification. In the multinomial model, nonoccuring terms do not impact classification decisions. However, the Bernoulli model considers the non-occurrence of terms and so includes the probability of a term not occurring when calculating the probability of a document d being in class c, or P(c|d).

References:
Lecture 7B

Pages 263-264: https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

3.) This is a problem about Naïve Bayes classification using the Bernoulli (also known as Binomial) model. First calculate estimates of P(c) and P(w|c) given the following training data. When computing estimates of P(w|c) I want you to use add-one smoothing (also known as Laplace smoothing) -- see the example in Section 13.3. There are three disjoint classes: Business, Health, and Travel. For features you should use only the following six terms: {chicago, employers, hawaii, jobs, nurses, vacation} and you should completely ignore all other terms. Next calculate and report scores for each class for the two input documents A & B below. Finally, indicate which class is predicted for each test document.

Count carefully, show how you arrive at your estimates, and include any intermediate work. Report scores using scientific notation (e.g., 1.563 x 10-5).

Training data

1 Travel: chicago mayor takes vacation in hawaii
2 Travel: nurses plan a trip to hawaii
3 Travel: employers offering more jobs with vacation benefits
4 Business: employers see growth in computer sector
5 Business: high paying retail jobs in chicago
6 Business: nurses find employers spending more on hospital jobs
7 Business: nice vacation spot but no jobs in hawaii
8 Health: nurses need to take a vacation
9 Health: doctors attend hawaii conference
10 Health: chicago employers have jobs for nurses

Test documents:

A: nurses take golf vacation in hawaii
B: top employers move jobs to Chicago

After removing all the terms that are not in the selected feature set, we get the following for the training data and test documents:

Training Data:

1 Travel: chicago vacation hawaii
2 Travel: nurses hawaii
3 Travel: employers jobs vacation

Total words in Travel: 8

4 Business: employers
5 Business: jobs chicago
6 Business: nurses employers jobs
7 Business: vacation jobs hawaii

Total words in Business: 9

8 Health: nurses vacation
9 Health: hawaii
10 Health: chicago employers jobs nurses

Total words in Health: 7

We can see above that the Travel documents have 8 words total, the business documents have 9 words total and the health documents have 7 words total. The number of unique words in the training data equals the number of selected features which is 6.

Test Documents
A: nurses vacation hawaii
B: employers job Chicago

Priors = P(c):
P(Business) = number business documents/total documents = 4/10 = 2/5

P(Health) = number of health documents/total documents = 3/10

P(Travel) = number of travel documents/total documents = 3/10

Test Document A:

	P(w|Travel)	P(w|Business)	P(w|Health)
nurses	$\frac{1+1}{8+6} = \frac{2}{14}$	$\frac{1+1}{9+6} = \frac{2}{15}$	$\frac{2+1}{7+6} = \frac{3}{13}$
vacation	$\frac{2+1}{8+6} = \frac{3}{14}$	$\frac{1+1}{9+6} = \frac{2}{15}$	$\frac{1+1}{7+6} = \frac{2}{13}$
hawaii	$\frac{2+1}{8+6} = \frac{3}{14}$	$\frac{1+1}{9+6} = \frac{2}{15}$	$\frac{1+1}{7+6} = \frac{2}{13}$

P(nurses vacation hawaii|Travel) = $\frac{2}{14} * \frac{3}{14} * \frac{3}{14} = 0.0065597 = 6.560 \times 10^{-3}$

P(nurses vacation hawaii|Business) = $\frac{2}{15} * \frac{2}{15} * \frac{2}{15} = 0.0023704 = 2.370 \times 10^{-3}$

P(nurses vacation hawaii|Health) = $\frac{3}{13} * \frac{2}{13} * \frac{2}{13} = 0.0054620 = 5.462 \times 10^{-3}$

Test Document A belongs to the Travel class

Test Document B:

	P(w|Travel)	P(w|Business)	P(w|Health)
employers	$\frac{1+1}{8+6} = \frac{2}{14}$	$\frac{2+1}{9+6} = \frac{3}{15}$	$\frac{1+1}{7+6} = \frac{2}{13}$
jobs	$\frac{1+1}{8+6} = \frac{2}{14}$	$\frac{3+1}{9+6} = \frac{4}{15}$	$\frac{1+1}{7+6} = \frac{2}{13}$
chicago	$\frac{1+1}{8+6} = \frac{2}{14}$	$\frac{1+1}{9+6} = \frac{2}{15}$	$\frac{1+1}{7+6} = \frac{2}{13}$

P(employers jobs chicago|Travel) = $\frac{2}{14} * \frac{2}{14} * \frac{2}{14} = 0.002915 = 2.915 \times 10^{-3}$

P(employers jobs chicago|Business) = $\frac{3}{15} * \frac{4}{15} * \frac{2}{15} = 0.007111 = 7.111 \times 10^{-3}$

P(employers jobs chicago|Health) = $\frac{2}{13} * \frac{2}{13} * \frac{2}{13} = 0.003641 = 3.641 \times 10^{-3}$

Test Document B belongs to the Business class.