605.744: Information Retrieval
Problem Set (Module 7)

Sabbir Ahmed

October 17, 2022

1. (40%) Briefly explain the following concepts in your own words and be sure to indicate how they are relevant to text classification:

   (a) bias-variance tradeoff

   Answer: Bias is the difference between the average prediction of the model and the expected value, while variance is the measure of the spread of the data. High bias oversimplifies the model which leads to consistent errors, while high variance leads to overfitting models. The bias-variance tradeoff refers to the tradeoff between those 2 metrics such that they do not underfit or overfit the model.

   (b) k-fold cross-validation

   Answer: K-fold cross-validation is typically required in models involved with limited data. It is a method where the data is shuffled and resampled into K-groups to evaluate the model.

   (c) macro-averaging vs. micro-averaging

   Answer: Macro-averaging gives equal weight to each class by computing their average, while micro-averaging prioritizes each per-document classification decision by pooling them across classes to create contingency tables.

   (d) soft margin

   Answer: Soft margins refer to the margins in SVMs being more relaxed so as to allow some wiggle room in the separating lines for the training data points. This is due to real data being unable to perfectly separate with SVM margins.

2. (10%) Why is 3-Nearest-Neighbor (3-NN) almost always a better choice than 1-NN for a binary (i.e, two-class) text classification problem.

   Answer: A 1-Nearest-Neighbor may introduce high bias to the model. Using 3-NN will allow for a test data point to be more accurately classified by observing the 3 nearest neighbors and being categorized with the majority class.

3. (50%) Naïve Bayes using the Binomial (also known as Bernoulli) model. First calculate estimates of P(c) and P(w | c) given the following training sentences. There are only three classes: Travel, Business, and Health. You should not use any smoothing. For features you should only use the following seven vocabulary terms: {denver, employers, florida, hospital, jobs, nurses, vacation} and you should ignore all other words and any punctuation. Next compute the probability of each class for the two test documents A & B below. Finally,

indicate which is the predicted (i.e., the most likely) class for each test document. Please read these directions thoroughly, count carefully, and do show all of your work. Report probabilities using scientific notation (e.g., 1.563×10^{-5}) with three digits after the decimal point.

Training data:

1) Travel: denver hospital administrator takes vacation in florida
2) Travel: nurses plan a trip to florida
3) Travel: employers offering more jobs with vacation benefits
4) Business: employers see growth in information science
5) Business: hospital nurses in denver say high paying jobs are vanishing
6) Business: employers say florida is nice vacation spot and there are good jobs
7) Health: study: more hospital nurses need to take a vacation
8) Health: local doctors attend florida conference on diabetes
9) Health: hospital trains maternity ward nurses
10) Health: denver hospital says local employers have jobs for nurses

Test documents:

(a) florida nurses take skiing vacation in denver
(b) jobs available for experienced nurses at florida hospital

Answer: The probabilities, P(c) and P(w|c), are calculated with the following equations:

$$P(c_i) = \frac{count(c_i)}{N},\ N = 10$$

$$P(w_j|c_i) = \frac{count(w_j = 1 \in c_i)}{count(c_i)}$$

Computing the probabilities for the classes, Travel, Business, and Health:

$$P(Travel) = \frac{3}{10},\quad P(Business) = \frac{3}{10},\quad P(Health) = \frac{4}{10}$$

Computing the probabilities for the terms, {denver, employers, florida, hospital, jobs, nurses, vacation}:

$$P(denver|Travel) = \frac{1}{3},\quad P(denver|Business) = \frac{1}{3},\quad P(denver|Health) = \frac{1}{4}$$

$$P(employers|Travel) = \frac{1}{3},\quad P(employers|Business) = \frac{2}{3},\quad P(employers|Health) = \frac{1}{4}$$

$$P(florida|Travel) = \frac{2}{3},\quad P(florida|Business) = \frac{1}{3},\quad P(florida|Health) = \frac{1}{4}$$

$$P(hospital|Travel) = \frac{1}{3},\quad P(hospital|Business) = \frac{1}{3},\quad P(hospital|Health) = \frac{3}{4}$$

$$P(jobs|Travel) = \frac{1}{3},\quad P(jobs|Business) = \frac{2}{3},\quad P(jobs|Health) = \frac{1}{4}$$

$$P(nurses|Travel) = \frac{1}{3},\quad P(nurses|Business) = \frac{1}{3},\quad P(nurses|Health) = \frac{3}{4}$$

$$P(vacation|Travel) = \frac{2}{3},\quad P(vacation|Business) = \frac{1}{3},\quad P(vacation|Health) = \frac{1}{4}$$

Summarizing the probabilities into a table:

Table 1: Probabilities of terms per documents

Term	**P(w|Travel)**	**P(w|Business)**	**P(w|Health)**
denver	0.33	0.33	0.25
employers	0.33	0.66	0.25
florida	0.66	0.33	0.25
hospital	0.33	0.33	0.75
jobs	0.33	0.66	0.25
nurses	0.33	0.33	0.75
vacation	0.66	0.33	0.25

To find the best class, the following equation can be used:

$$\text{BestClass} = P(c_i) \prod_{j=1}^{\#\text{ words}} P(w_j|c_i)$$

After normalization, the first test document becomes:

$$D_A = \{florida, nurses, vacation, denver\}$$

$$\begin{aligned} P(Travel|D_A) &= P(Travel) \cdot P(denver|Travel) \cdot P(employers|Travel) \cdot P(florida|Travel) \\ &\quad \cdot P(hospital|Travel) \cdot P(jobs|Travel) \cdot P(nurses|Travel) \cdot P(vacation|Travel) \\ &= (0.30)(0.33)(1-0.33)(0.66)(1-0.33)(1-0.33)(0.33)(0.66) \\ &= 4.390 \times 10^{-3} \end{aligned}$$

$$\begin{aligned} P(Business|D_A) &= P(Business) \cdot P(denver|Business) \cdot P(employers|Business) \\ &\quad \cdot P(florida|Business) \cdot P(hospital|Business) \cdot P(jobs|Business) \\ &\quad \cdot P(nurses|Business) \cdot P(vacation|Business) \\ &= (0.30)(0.33)(1-0.66)(0.33)(1-0.33)(1-0.66)(0.33)(0.33) \\ &= 2.743 \times 10^{-4} \end{aligned}$$

$$\begin{aligned} P(Health|D_A) &= P(Health) \cdot P(denver|Health) \cdot P(employers|Health) \cdot P(florida|Health) \\ &\quad \cdot P(hospital|Health) \cdot P(jobs|Health) \cdot P(nurses|Health) \cdot P(vacation|Health) \\ &= (0.40)(0.25)(1-0.25)(0.25)(1-0.75)(1-0.25)(0.75)(0.25) \\ &= 6.592 \times 10^{-4} \end{aligned}$$

Therefore, the best class for D_A is “Travel”.

After normalization, the second test document becomes:

$$D_B = \{jobs, nurses, florida, hospital\}$$

$$
\begin{aligned}
P(Travel|D_B) &= P(Travel) \cdot P(denver|Travel) \cdot P(employers|Travel) \cdot P(florida|Travel) \\
&\quad \cdot P(hospital|Travel) \cdot P(jobs|Travel) \cdot P(nurses|Travel) \cdot P(vacation|Travel) \\
&= (0.30)(1-0.33)(1-0.33)(0.66)(0.33)(0.33)(0.33)(1-0.66) \\
&= 1.097 \times 10^{-3}
\end{aligned}
$$

$$
\begin{aligned}
P(Business|D_B) &= P(Business) \cdot P(denver|Business) \cdot P(employers|Business) \\
&\quad \cdot P(florida|Business) \cdot P(hospital|Business) \cdot P(jobs|Business) \\
&\quad \cdot P(nurses|Business) \cdot P(vacation|Business) \\
&= (0.30)(1-0.33)(1-0.66)(0.33)(0.33)(0.66)(0.33)(1-0.33) \\
&= 1.097 \times 10^{-3}
\end{aligned}
$$

$$
\begin{aligned}
P(Health|D_B) &= P(Health) \cdot P(denver|Health) \cdot P(employers|Health) \cdot P(florida|Health) \\
&\quad \cdot P(hospital|Health) \cdot P(jobs|Health) \cdot P(nurses|Health) \cdot P(vacation|Health) \\
&= (0.40)(1-0.25)(1-0.25)(0.25)(0.75)(0.25)(0.75)(1-0.25) \\
&= 5.933 \times 10^{-3}
\end{aligned}
$$

Therefore, the best class for D_B is “Health”.