Student Name: _____

Midterm Exam
Text Data Mining (INLS 613)
October 19th, 2016

Answer all of the following questions. Each answer should be thorough, complete, and relevant. Points will be deducted for irrelevant details. Use the back of the pages if you need more room for your answer.

The points are a clue about how much time you should spend on each question. Plan your time accordingly.

Good luck!

Question	Points
1	10
2	10
3	20
4	10
5	10
6	20
7	20
Total	**100**

1. Inter-annotator Agreement [10 points]

Predictive analysis of text often requires annotating data. In doing so, one important step is to verify whether human annotators can reliably detect the phenomenon of interest (e.g., whether a product review is *positive* or *negative*).

Suppose that two annotators (A and B) independently annotate 100 product reviews and produce the following contingency table. Answer the following questions.

		Annotator B	
		Positive	Negative
Annotator A	Positive	30	20
	Negative	20	30

(a) What is the inter-annotator agreement between A and B based on *accuracy* (i.e., the percentage of times both annotators agreed)? **[5 points]**

(b) What is the inter-annotator agreement between A and B based on Cohen's *Kappa* assuming unbiased annotators—each annotator makes annotations randomly with equal (i.e., 50/50) probability. **[5 points]**

2. Training and Testing [10 points]

The goal in predictive analysis is to train a model that can make predictions on new data. When a model fails to perform well on new data, it is often because it "catches on" to regularities in the training data that are not meaningful (i.e., do not hold true in general).

(a) Suppose we increased the size of the training set. Would this likely improve or deteriorate the performance of the model on new data? Why? **[5 points]**

(b) Suppose we reduced the feature representation by ignoring terms that appear only once in the training set. Would this likely improve or deteriorate the performance of the model on new data? Why? **[5 points]**

3. Evaluation Metrics [20 points]

Suppose we train a model to predict whether an email is **Spam** or **Not Spam**. After training the model, we apply it to a test set of 200 new email messages (also labeled) and the model produces the contingency table below.

		True Class	
		Spam	Not Spam
Predicted Class	Spam	60	120
	Not Spam	0	20

(a) Compute the precision of this model with respect to the **Spam** class and with respect to the **Not Spam** class. **[5 points]**

(b) Compute the recall of this model with respect to the **Spam** class and with respect to the **Not Spam** class. **[5 points]**

(c) Suppose we have two users (Emily and Simon) with the following preferences.

> **Emily** hates seeing spam messages in her inbox! However, she doesn't mind periodically checking the "junk" folder for messages incorrectly marked as spam.
>
> **Simon** doesn't even know where the "junk" folder is. He would prefer to see spam messages in his inbox than to miss genuine messages without knowing!

Would Emily like this classifier? Justify your answer based on the precision and recall values of this classifier with respect to the **Spam** class. **[5 points]**

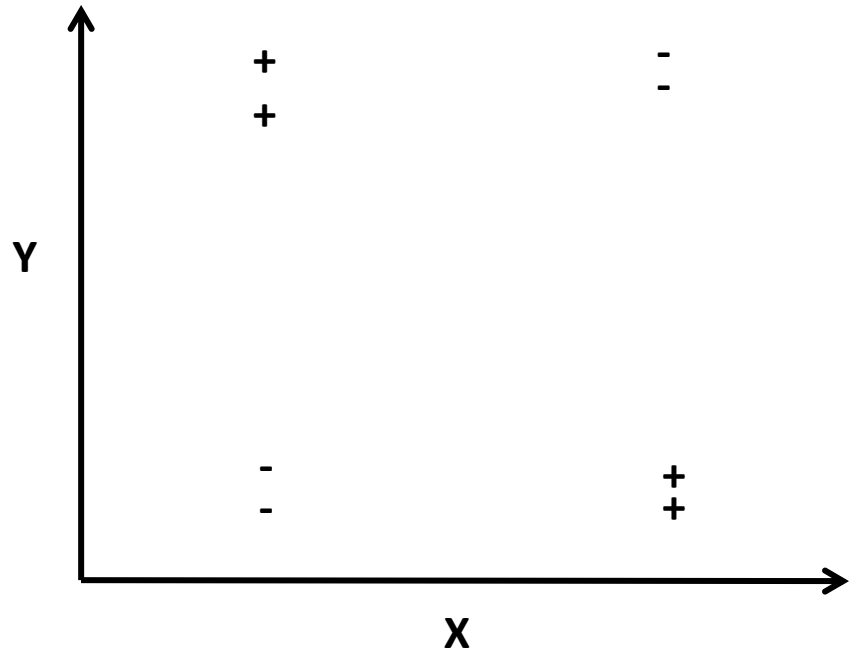
(d) Would Simon like this classifier? Justify your answer based on the precision and recall values of this classifier with respect to the **Spam** class. **[5 points]**

4. Instance-Based Classification and Parameter Tuning [10 points]

Given a new instance, a KNN classifier predicts the majority class associated with the K nearest neighbors. K is a parameter that needs to be set using training data.

Suppose we have the following training set of *positive* (+) and *negative* (-) movie reviews.

All instances are projected onto a vector space of two real-valued features (X and Y) and the distance between instances is computed using the Geometric (or Euclidean) Distance.

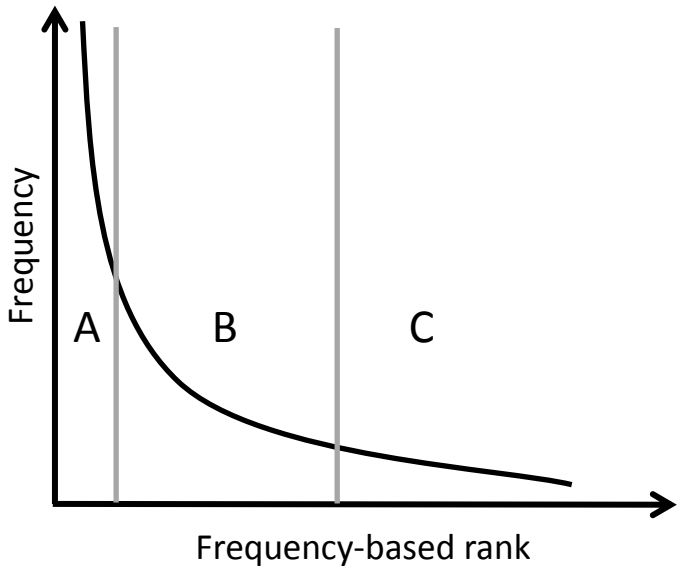


(a) What value of K would maximize <u>accuracy</u> on this training set? What is the training set accuracy of a classifier that uses this <u>optimal</u> value of K? **[5 points]**

(b) Is the training set accuracy of a classifier that uses the optimal value of K a good indicator of how the classifier will perform on new data? Why or why not? **[5 points]**

5. Zipf's law and Feature Representation [10 points]

Zipf's law tells us that in most collections of text, a few terms will occur very frequently and many terms will occur very infrequently. In other words, if we were to plot term frequency (Y-axis) as a function of frequency-based rank (X-axis), we get the graph below with three general regions: A, B, and C.



In predictive analysis of text, we often use individual terms as features. It is oftentimes beneficial to use only those terms from region B as features and to ignore those terms from regions A and C. Answer the following questions.

(a) What is the reasoning behind ignoring terms from region A? **[5 points]**

(b) What is the reasoning behind ignoring terms from region C? **[5 points]**

6. Naïve Bayes [20 points]

Suppose we have the following training set of *positive* (+) and *negative* (-) movie reviews. There are 5 training instances and 3 features.

great	fine	terrible	class
1	0	0	+
0	1	1	--
0	1	1	--
0	0	0	+
1	0	1	+

Suppose we train a Naïve Bayes classifier on this training set without doing any sort of smoothing. Answer the following questions.

(a) What is the prior probability of *positive*, denoted as P(+)? **[5 points]**

(b) What is the prior probability of *negative*, denoted as P(-)? **[5 points]**

(c) What class (*positive* or *negative*) would the model predict for a movie review that just says “great!” and what would be the confidence value associated with the predicted class? **[10 points]**

7. Prediction Confidence and Precision vs. Recall [20 points]

Suppose we train a Naïve Bayes Classifier to predict *positive* vs. *negative* movie reviews. At test time, a Naïve Bayes Classifier estimates the probability that a review is positive, **P(+|D)**.

Suppose we apply our Naïve Classifier to a test set of 20 instances and obtain the following ranking:

Rank	P(+|D)	True Category
1	0.99	+
2	0.97	+
3	0.91	+
4	0.89	--
5	0.80	--
6	0.78	--
7	0.60	+
8	0.55	+
9	0.41	+
10	0.39	--
11	0.22	--
12	0.19	--
13	0.10	--
14	0.09	--
15	0.06	--
16	0.05	--
17	0.04	--
18	0.03	--
19	0.02	--
20	0.01	--

As it turns out, we can apply a threshold *T* to **P(+|D)** in order to favor precision over recall (or vice-versa).

Answer the following questions.

(a) With respect the *positive* class, **Thomas** cares more about precision than recall. In fact, he would like precision to be higher than 80% and recall to be higher than 45%. What value of T would you use for Thomas. Explain your answer in terms of expected level of precision and recall for you chosen value of T. **[10 points]**

(b) With respect the *positive* class, **Sarah** cares more about recall than precision. In fact, she would like recall to be 100% and precision to be higher than 50%. What value of T would you use for Sarah. Explain your answer in terms of precision and recall for you chosen value of T. **[10 points]**