Student Name: Rees Braam

Final Exam
Text Data Mining (INLS 613)

Please answer all of the following questions. Each answer should be thorough, complete, and relevant. Points will be deducted for irrelevant details.

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	10
4	15
5	20
6	20
7	15
Total	**100**

1. Inter-annotator Agreement [10 points]

Predictive analysis of text often requires annotating data. In doing so, one important step is verifying 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 200 product reviews and produce the following contingency matrix.

Answer the following questions.

		Annotator B	
		Positive	Negative
Annotator A	Positive	80	40
	Negative	0	80

(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]**

$$\frac{\mathrm{A}+\mathrm{D}}{\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{D}}=\frac{80+80}{80+40+0+80}=\frac{160}{200}=0.8$$

Answer: 0.8

(b) What is the inter-annotator agreement between A and B based on Cohen's *Kappa* assuming unbiased annotators (i.e., each annotator has a 50/50 chance of saying the review is positive/negative) **[5 points]**

P(a) can be calculated from part (a). *P(e)* can be calculated by using an alternative chart where we assume that each annotator randomly assessed each of the 200 reviews with a 50/50 chance, meaning A=B=C=D=50.

$$P(a)=\frac{\mathrm{A}+\mathrm{D}}{\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{D}}=0.8, \qquad P(e)=\frac{50+50}{200}=0.5$$

$$K=\frac{P(a)-P(e)}{1-P(e)}=\frac{0.8-0.5}{1-0.5}=\frac{0.3}{0.5}=0.6$$

Answer: 0.6

2. Training and Testing [10 points]

The goal in predictive analysis is to train a model that can make accurate predictions on new data. When a model fails to do well on new data, it is often because it "catches on" to regularities in the training data that 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]**

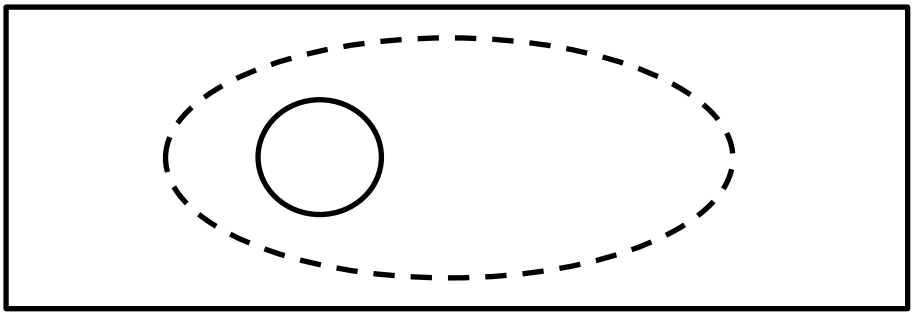
Increasing the size of the training set would likely improve the performance of the model on new data. Adding more data will give more example instances so that the model can understand more edge cases and have a better understanding of which features are important and which are not. For example, in the case where the training set is of size one, the model would do very poorly because it would only have one example. As the training set increases in size, more variation is accounted for, allowing the model to better handle new data that it has never seen before.

(b) Suppose we decide to omit all features 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]**

This is also likely to improve the performance of the model on new data. Assuming that our training set is appropriately large, any feature that occurs only once is not likely to be important or useful as a predictor. However, the model may end up using these features anyways, resulting in a model that uses features that may be misrepresentative of the underlying truth. This can be confirmed from our experience with homework 2 and the usage of LightSIDE, where we saw that increasing the occurrence threshold above one improved out model performance on the test data.

3. Evaluation Metrics [10 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 following result. The dashed oval denotes the test set instances that are actually **Spam** and the solid circle denotes the test set instances that were predicted to be **Spam.**



(a) With respect to Spam, Is this result a high precision or a high recall result? Please explain. **[5 points]**

With respect to Spam, this result is a high precision. Precision is the percentage of positive predictions that are truly positive. Of our positive predictions, 100% are truly positive. On the other hand, recall is the percentage of true positive that were correctly predicted as positive. In this case, we can see that the dashed oval is much larger than the solid circle, meaning that our recall is relatively low.

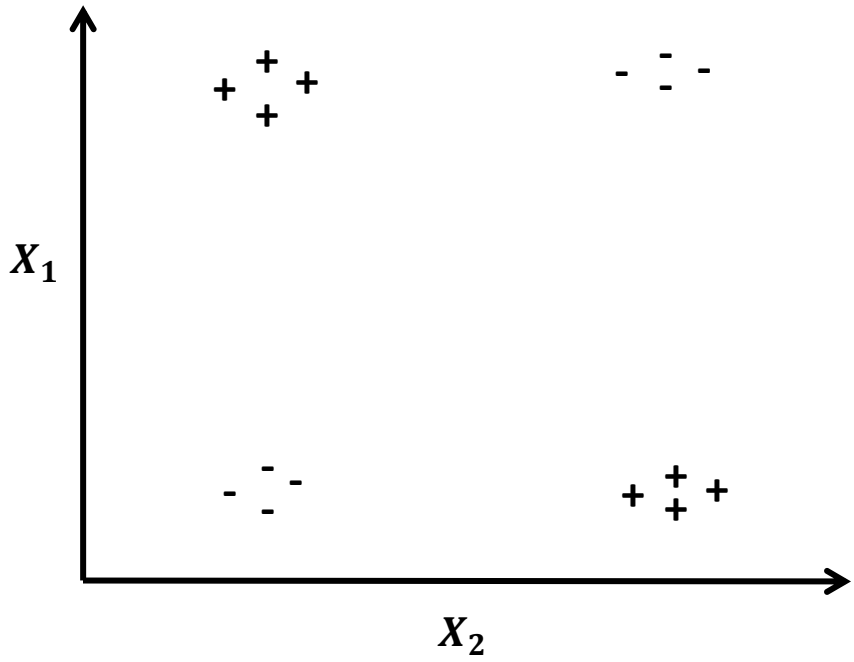
(b) Emily hates seeing spam messages in her inbox! However, she doesn't mind periodically checking the "Junk" directory for non-Spam messages incorrectly marked as spam. Would Emily like this classifier? Justify your answer based on the classifier's performance in terms of precision and recall with respect to **Spam**. **[5 points]**

With the current classifier, Emily will see a lot of spam in her inbox but will almost never have to check the junk directory for non-Spam marked incorrectly. Essentially, this classifier has high precision and low recall, but Emily wants one with low precision and high recall. As the current classifier is the exact opposite of what Emily wants, it is safe to say that Emily will not like this classifier.

4. Logistic Regression and K Nearest Neighbor [15 points]

Suppose you want to train a classifier to predict whether a movie review is *positive* (+) or *negative* (-) based on two real-valued features: X_1 and X_2.

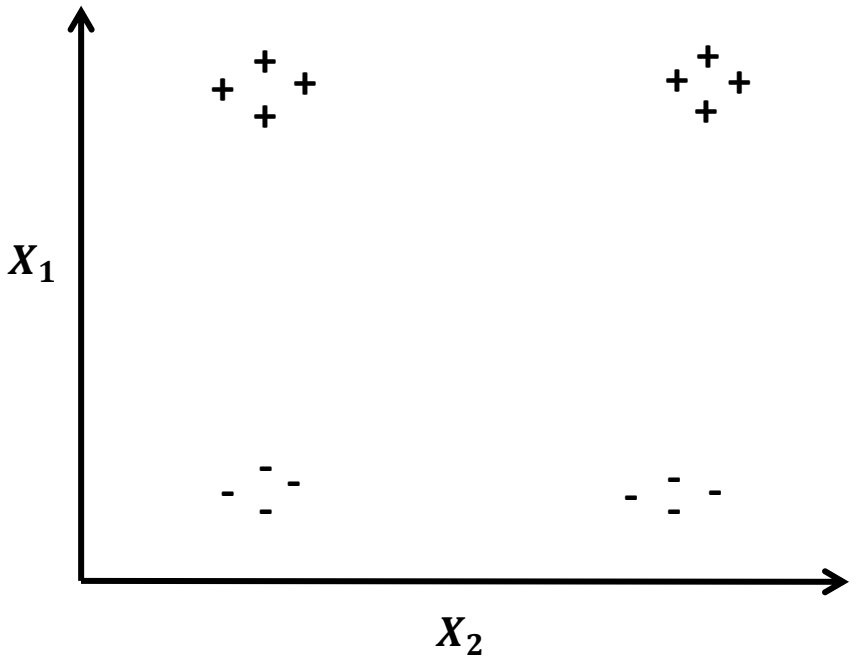
You decide to look at the data and you realize that the training set looks like this:



(a) Which algorithm do you think would perform better on this task: logistic regression or K nearest neighbor (setting K = 1)? Provide a brief justification. **[5 points]**

I think that K nearest neighbors will do better than logistic regression here. This is because logistic regression will output a linear classifier that splits the input into two spaces linearly, where one side is positive and the other is negative. In this case, it is not possible to accurately predict the data in the training set like this. K nearest neighbors, however, can split the input into more than two spaces. Therefore, I think that K nearest neighbors will perform better than logistic regression with the given training set.

(b) Now, suppose the data looks like this:



You plan to train a logistic regression classifier. Internally, logistic regression predicts a binary output (*positive* = 1 or *negative* = 0) using the following equations:

$$output = \frac{1}{1 + e^{-z}}$$

$$z = w_1 x_1 + w_2 x_2 + b$$

After training the logistic regression model, which of the three outcomes below do you think is most likely?

- Outcome 1: $(w_1 > 0)$
- Outcome 2: $(w_1 < 0)$
- Outcome3: $(w_1 \approx 0)$

Briefly justify your choice [**10 points**]

From the equation for z, we know that the output will be 1 if z is positive and 0 if z is negative. From the plot of the data, we know that the output should be 0 if X_1 is small and 1 if X_1 is large. Therefore, only one of these outcomes makes sense: $w_1 > 0$ (with what will likely be a negative b value). This is because we do not want w_1 to invert the value of X_1 so that a big X_1 can lead to a positive z and an output of 1.

5. Naïve Bayes [20 points]

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

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

$$P(+) = \frac{2}{5} = 0.4$$

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

$$P(-) = \frac{3}{5} = 0.6$$

(a) What class (positive or negative) would the model predict for a movie review that just says "terrible!" and what would be the confidence value associated with the predicted class? **Hint:** notice that this test instance would have the following feature values: great=0, fine=0, and terrible=1. **[10 points]**

This model would certainly predict negative for this review, as its feature values perfectly match all three negative instances that it already has. Mathematically this can be shown below, where the model predicts positive if the LHS is greater than the RHS, else the model predicts negative.

Predict positive if:

$$P(POS) * \prod_{i=1}^{n} P(w_i = D_i|POS) \geq P(NEG) * \prod_{i=1}^{n} P(w_i = D_i|NEG)$$

Plugging in, we get:

$$\frac{2}{5} * (0 * 0 * 0) \geq \frac{3}{5} * (1 * 1 * 1)$$

Because this is false, we know for a fact that the model will predict negative.

We can then compute the confidence of the NEG prediction given instance D:

$$P(NEG|D) = \frac{P(D|NEG) * P(NEG)}{P(D)} = \frac{\frac{1}{3} * \frac{3}{5}}{\frac{1}{5}} = \frac{0.2}{0.2} = 1$$

6. 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 Bayes 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, Yawei cares more about precision than recall. In fact, she would like precision to be greater than 90% and recall to be greater than 50%. What value of T would you use for Yawei? Explain your answer in terms of expected level of precision and recall for you chosen value of T. **[10 points]**

T	Precision	Recall
0.9	1	0.5
0.85	1	0.66
0.8	0.8	0.66
0.75	0.66	0.66
0.7	0.66	0.66
0.65	0.66	0.66
0.6	0.57	0.66
0.55	0.63	0.83
0.5	0.63	0.83
0.45	0.63	0.83
0.4	0.66	1

From this chart, we can see that a T value of 0.85 would meet Yawei's criteria. However, because there are no test instances between 0.80 and 0.89, we would need a larger test set to narrow down the exact threshold that would work for Yawei. However, because we know that Yawei cares more about precision than recall, a higher threshold is better here since we want the model to be certain when it predicts an instance as positive. Therefore, the value of T that I would use for Yawei based off this test set is 0.89. That is, P(+|D) must be greater than or equal to 0.89 to be predicted as positive.

(b) With respect the positive class, Carl cares more about recall than precision. In fact, he would like recall to be greater 90% and precision to be greater than 50%. What value of T would you use for Carl. Explain your answer in terms of precision and recall for you chosen value of T. **[10 points]**

As we can see from the chart that I made above, the only way to get recall greater than 90% is to have a threshold of 0.4 or lower. However, because the point in the test set that brought the threshold this low is at 0.41, our threshold must actually be 0.41 or below. However, going lower would only decrease the precision, as our recall is already 1. Therefore, the value of T that I would use for Carl based off this test set is 0.41, as it has a precision of 66% and a recall of 1, meeting both of his criteria.

7. Statistics Significance [15 points]

Suppose we have an experimental system that we want to compare against a baseline system. A reasonable approach is to perform 10-fold cross-validation (using the same folds) and to compare both systems based on their average performance. Suppose we do this and obtain the results below. In this case, the experimental system outperforms the baseline by a margin of 0.135. This result suggests that the experimental system is better. However, before celebrating, we decide to perform a statistical significance test.

Answer the following questions.

Fold	Baseline System	Experimental System
1	0.336	0.748
2	0.241	0.158
3	0.784	0.951
4	0.974	0.046
5	0.504	0.516
6	0.985	0.209
7	0.883	0.555
8	0.474	0.830
9	0.938	0.819
10	0.092	0.026
Average	**0.621**	**0.486**
		0.135

(a) In a statistical significance test, what is the null hypothesis? **[5 points]**

In a significance test, the null hypothesis is the assumption that there is no significant difference between the two given variables, groups, populations, etc. By “no significant difference,” it’s meant that any observed difference is due to random chance or some sort of unaffiliated error. In this case, the null hypothesis would be: “The experimental system is not better than the baseline system.”

(b) Suppose we perform a statistical significance test and the output *p*-value is 0.50. What does this mean? Your answer should start: "A *p*-value of 0.50 means that …. " **[10 points]**

A p-value of 0.50 means that there is a 50% chance of the observed results occurring by chance. A p-value below the confidence level (α) is required to reject the null hypothesis. The standard confidence level is 0.05 and can vary, but our confidence level is definitely not 0.50. Therefore, with a p-value of 0.50, we are unable to reject the null hypothesis that the experimental system is not better than the baseline system. Good thing we didn't celebrate!