University of Glasgow

Friday 26 April 2019
2.00 pm – 3.30 pm
(Duration: 1 hour 30 minutes)

DEGREES OF MSc, MSci, MEng, BEng, BSc,MA and MA (Social Sciences)

Text as Data H

(Answer All Questions)

This examination paper is worth a total of 60 marks

The use of a calculator is not permitted in this examination

INSTRUCTIONS TO INVIGILATORS
Please collect all exam question papers and exam answer scripts and retain for school to collect. Candidates must not remove exam question papers.

4 questions, 60 marks

1. (a) Consider two documents: 1. *"This is the first document"*. 2. *"This is the second document"*.

(i) *Show* the output of SkLearn's CountVectorizer on applying `fit_transform()` to the first document.

[1]

(ii) *Show* and *Explain* the output of the same vectorizer when applying `transform()` to the second document.

[2]

(iii) Explain what TFIDFVectorizer does that makes it take longer to perform `fit_transform()` than CountVectorizer.

[2]

(b) Consider two documents, with vocabulary of size 5. The documents have been vectorized and have the following word counts:

D1 = [3, 2, 1, 0, 0] D2 = [1, 2, 3, 2, 4]

View the documents as a one-hot encoding of terms. Provide the formula for the Dice similarity. Calculate the Dice similarity between D1 and D2. Show your workings.

[2]

(c) You have been tasked with clustering documents using K-means. You are currently using a TF (Count) vector representation. Your manager has asked you to use TF.IDF vectors instead. Explain how would you expect the obtained clustering to change? Use an illustrative example of words or clusters to explain your answer.

[4]

(d) This question is about Word2Vec (W2V) embeddings.

(i) What type of normalization is used for the W2V model to create a probability? Describe the reason for this normalization and how it relates to the type of classification performed.

[2]

(ii) What are the two main model variants for Word2Vec embeddings? Name and describe each model, including what the model predicts.

[2]

2. This question is about supervised and unsupervised learning.

(a) For a Logistic Regression classifier in a binary classification problem, consider a prediction rule to predict class 1 if $P(y = 1|x) > t$, the decision boundary. If the $P(y = 1|x)$ is greater than t, it will predict class 1, otherwise it will predict class 2.

(i) What is the typical assumed value for t?

[1]

(ii) If the threshold t is increased, what happens to the precision and recall values? Does each tend to increase, decrease, or stay the same? Explain why.

[1]

(b) You are building a web document classifier to identify pages about a rare topic, ACDC (the very rare disease, not the band). There are a small number of positive examples (0.0001%) and a large number of negative examples (99.9999%). This task has a common problem in classification. Name and describe this problem. How can this problem be addressed? Name and describe two methods to solve this issue -- include the potential drawbacks of using each method.

[3]

(c) Binary-classification is the most common type of classification. Name and define one other type of classification. Give an example of the type task.

[2]

(d) Describe logistic regression and kNN classification models and parameters. Compare and contrast why you would prefer one over the other for the task of categorizing news articles by their high-level topic.

[4]

(e) We want to classify a document as being about politics or not using Naive Bayes. Each document is associated with a pair $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is a feature vector of word counts of the document and $\boldsymbol{y}$ is the label for whether it is about politics (y = 1 if yes, y = 0 if false). There are 20 instances in the training data, with 5 positive examples and 15 negative examples. The vocabulary size on the training data is 4, with the values illustrated below.

The following conditional probability table has been computed for each feature in the training data, (each row is the class's unigram probability):

word feature *w*	**theresa**	**may**	**europe**	**pizza**
P(w | y = 1)	8/20	8/20	4/20	0/20
P(w | y = 0)	10/20	8/20	2/20	1/20

Consider a test document containing the following terms:

x = {theresa, europe, spain}.

For this document, give the definition of the Naive Bayes with the maximum a posterior (MAP) estimate, e.g. $P_{max}(x|y)$.

(i) Which class has the highest probability ($P_{max}(x|y)$? Show your workings.

[2]

(ii) Consider a test document **x** = {italy, pasta}. What class is most likely? Why?

[1]

(iii) What is *naive* about Naive Bayes? Describe this problem and give an illustration.

[2]

3. This question is about language modeling.

(a) You are in a loud Glasgow bar studying for your exam during a football match. A friend tries to get your attention using two words ("hoy" and "aye"). They yelled a sentence, but you only got fragments of it.

w1=hoy w2=aye, w3=??? w4=hoy

Based on your experience with your friend, you have learned the following bigram language model.

p(hoy | aye) = 7/10
p(aye | aye) = 3/10
p(hoy | hoy) = 4/10
p(aye | hoy) = 6/10

Calculate the sequence probabilities of the noisy sentence using a bigram model for both possibilities of w3. What is the most likely value for w3? Show your workings. (Note: There is no start or end padding.)

[3]

(b) Smoothing is needed in language modeling for documents to avoid zero probabilities. Define Dirichlet smoothing by giving its formula. What is the effect when the value of μ (mu) is increased?

[3]

(c) This question concerns document similarity using KL-Divergence.

(i) Given the table of collection counts below, provide the definition of the unigram probability, P(w). Complete the table for with the probability for each term in the collection below.

[2]

t	count	P(w)
the	100	
mouse	30	
elephant	20	
haggis	20	
cheese	10	
wears	5	
likes	5	
pajamas	5	
said	5	

(ii) Define KL Divergence with its formula. What does KL measure and how can this be interpreted?

[2]

(ii) What is the value for KL(D1|| D2) without smoothing? Give a specific reason.

D1

t	Pr(t|d1)
elephant	0.2
pajamas	0.05

D2

t	Pr(t|d1)
mouse	0.2
likes	0.1
pajamas	0.08

[2]

(iii) What is the value for KL(D1||D2) using Jelenik-Mercer smoothing with lambda = 0.7 for the calculation. Use the collection probabilities from (i) for the background and the document probabilities from (ii). Show your workings (you do not need evaluate your arithmetic).

[3]

4. This question concerns NLP, word embeddings & dialog systems.

Below is an example conversation with an agent.

> **User**: I'd like to book a flight to the Bay Area.
> System: There are three airports there: San Jose, San Francisco, and Oakland.
> **User**: How far is SF from Google?
> System: It has multiple offices: Google Mountain View and Google San Francisco. Which one are you referring to?
> **User**: MTV.
> System: The office is 34 miles away and the journey takes approximately one hour by car.
> **User**: Book me a hotel close to where Greg lives.
> System: A hotel near Gregor is the Marriott in Burlingame.
> **User**: Not that one. Maybe somewhere closer to where he lives.
> System: How about the Holiday Inn in San Mateo?

(a) This question is about coreference resolution in the dialogue system.

(i) Define the problem of coreference resolution. Given an example of it from the interactions above.

[2]

(ii) What type of classification problem is a pairwise coreference resolution model? What machine learning model would you use to solve it? Justify your decision.

[2]

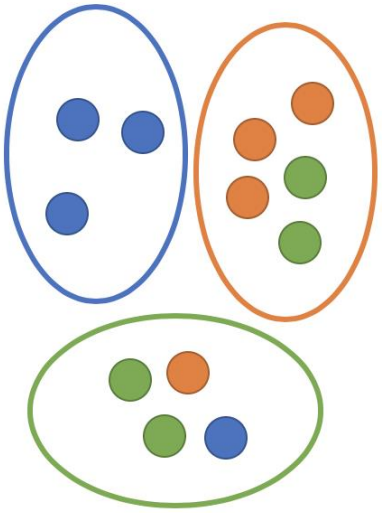
(iii) Define two binary features to use for a coreference resolution pairwise model. For each feature provide three things: an informative name, a description of the feature value, and a positive example for that feature from the dialogue above.

[4]

(iv) Below is an example of chains that form three clusters. The colors of the balls are the true clusters.

Name and describe an appropriate evaluation measure that incorporates both precision and recall. Specify its formula in terms of precision and recall.

[2]



(b) This question concerns dialog systems.

(i) The dialogue agent above is a task-oriented travel agent. What task-oriented dialogue architecture would you use to implement a dialogue system of this type? Describe the architecture and provide a compelling justification for this design decision.

[2]

(ii) A text classifier is created to classify an utterance into set of predefined *intents* in a dialog system. How could word embeddings be used as features for this task? What language problems do word embeddings address in this problem?

[2]