

Notes for W/4 lecture

# Midterm practice questions

(Last update: Nov 4 before lecture)

UMass CS 490A — Fall 2021

## 1 Midterm information

The midterm will be in-class during the normal time, on Tuesday 11/9. It's closed-book, except you can bring a "cheat sheet," one page of notes (front and back is fine) that you write for yourself. (Typically the act of writing the notes can be a useful study aid!)

## 2 Topics on the midterm

Anything covered in class, readings, homeworks, or exercises can be on the midterm. We're more likely to focus on topics covered in class or referred to in class. This includes the following:

Language concepts

- Regular expressions
  - As noted in the 11/2 lecture, be familiar with the core operators like parens, star, plus, disjunctions, and ranges. In the Python regex library, there are many other convenience features, that are not important to know.

- Text normalization, tokenization

Probability, language modeling, classification

- Probability theory, including Bayes Rule
- Relative frequency estimation and pseudocount smoothing
- N-gram (Markov) language models
- Naive Bayes
- Logistic regression, binary classification

- Logistic regression, multiclass classification
- *Note:* gradient derivations will not be included

#### Evaluation / Annotation

- Classification evaluation metrics: false positives/negatives, precision, recall, F1
- Annotator agreement rates

#### Linguistic structure representations

- Part of speech tags
- Constituency grammars and parses
- Dependency parses

#### Latent embedding representations

- Word embeddings, BoE
- Neural networks: feedforward NNs, deep averaging

### 3 Classification

**Question 3.1.** Consider training and predicting with a naive Bayes classifier for two document classes, and without pseudocounts. The word “booyah” appears once for class 1, and never for class 0. When predicting on new data, if the classifier sees “booyah”, what is the posterior probability of class 1?

**Question 3.2.** For a probabilistic classifier for a binary classification problem, consider the prediction rule to predict class 1 if  $P(y = 1|x) > t$ , and predict class 0 otherwise. This assumes some threshold  $t$  is set. If the threshold  $t$  is increased,

- Does precision tend to increase, decrease, or stay the same?
- Does recall tend to increase, decrease, or stay the same?

#### Classification example

Here’s a naive Bayes model with the following conditional probability table (each row is that class’s unigram language model):

word type	a	b	c
$P(w   y = 1)$	5/10	3/10	2/10
$P(w   y = 0)$	2/10	2/10	6/10

and the following prior probabilities over classes:

By cond. indep.

$$\begin{aligned}
 p(\vec{w} = (a, a, b, c) | y = 1) &= p(a|y=1) p(a|y=1) p(b|y=1) p(c|y=1) \\
 &= \frac{5}{10} \cdot \frac{5}{10} \cdot \frac{3}{10} \cdot \frac{2}{10}
 \end{aligned}$$

$$= \frac{25 \times 3 \times 2}{1000} = \frac{150}{10,000}$$



$P(y = 1)$	$P(y = 0)$
$8/10$	$2/10$

## Naive Bayes

Consider a binary classification problem, for whether a document is about the end of the world (class  $y = 1$ ), or it is not about the end of the world (class  $y = 0$ ).

**Question 3.3.** Consider a document consisting of 2 a's, and 1 c.

*Note:* In this practice and on the midterm, you do not need to convert to decimal or simplify fractions. You may find it easier to not simplify the fractions. On the midterm, we will not penalize simple arithmetic errors. Please show your work.

- (a) What is the probability that it is about the end of the world?
- (b) What is the probability it is not about the end of the world?

**Question 3.4.** Now suppose that we know the document is about the end of the world ( $y = 1$ ).

- (a) True or False, the naive Bayes model is able to tell us the probability of seeing the document  $\vec{w} = (a, a, b, c)$  under the model.
- (b) If True, what is the probability?

$$p(\vec{w} = (a, a, b, c) | y = 1)$$

## Logistic Regression

Consider a logistic regression model for this same problem ( $y = 1$  means the document is about the end of the world), with three features. The model has weights  $\beta = (0.5, 0.25, 1)$ .

*Note:* for this problem you will be exponentiating certain quantities. You do not need to write out your answer as a number, but instead in terms of  $\exp()$  values, e.g.,  $P = 1 + 2\exp(-1)$ .

**Question 3.5.** A given document has feature vector  $x = (1, 4, 0)$ .

- (a) What is the probability that the document is about the end of the world?
- (b) What is the probability that it is not about the end of the world?

$$p(y=1|x)$$

**Question 3.6.** Now suppose that we know the document is about sports ( $y = 1$ ).

- (a) True or False, the logistic regression model is able to tell us the probability of seeing  $x = (1, 1, 2)$  under the model.
- (b) If True, what is the probability? (again, answer in terms of  $\exp()$  values).

$$p(x = (1, 1, 2) | y = 1)$$



(b) If True, what is the probability?

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

### Logistic Regression

Consider a logistic regression model for this same problem ( $y = 1$  means the document is about the end of the world), with three features. The model has weights  $\beta = (0.5, 0.25, 1)$ .

Note: for this problem you will be exponentiating certain quantities. You do not need to write out your answer as a number, but instead in terms of  $\exp()$  values, e.g.,  $P = 1 + 2\exp(-1)$ .

$$p(y=1|x) = \sigma(\beta \cdot x)$$

**Question 3.5.** A given document has feature vector  $x = (1, 4, 0)$ .

- (a) What is the probability that the document is about the end of the world?  
 (b) What is the probability that it is not about the end of the world?

$$z = \beta \cdot x = 0.5 \cdot 1 + 0.25 \cdot 4 + 0 = 1.5$$

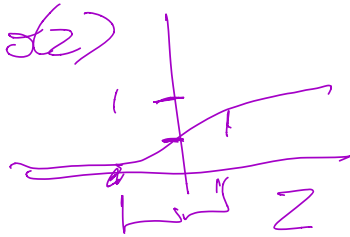
$$p(y=1|x) = \sigma(z) = \frac{1}{1 + e^{-1.5}}$$

$$p(y=0|x) = 1 - \sigma(z) = 1 - \frac{1}{1 + e^{-1.5}}$$

$$= \frac{1 + e^{-1.5}}{1 + e^{-1.5}} - \frac{1}{1 + e^{-1.5}}$$

$$= \frac{e^{-1.5}}{1 + e^{-1.5}} \cdot \frac{e^{1.5}}{e^{1.5}}$$

$$= \frac{1}{e^{1.5} + 1}$$



the end of the world is 0 (or incredibly close).

**Question 3.8.** Show the two standard definitions of the logistic sigmoid function are equivalent.

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad \dots \quad \frac{e^z}{1+e^z}$$

$$\frac{1}{1+e^{-z}} \cdot \frac{e^z}{e^z} = \frac{e^z}{(1+e^{-z})e^z}$$

$$= \frac{e^z}{e^z+1} \quad \checkmark$$

**Question 3.7.** Consider a logistic regression model with weights  $\beta = (\beta_1, \beta_2, \beta_3)$ . A given document has feature vector  $x = (1, -2, -1)$ .

1. What is a value of the vector  $\beta$  such that the probability of the document being about the end of the world is 1 (or incredibly close)?
2. What is a value of the vector  $\beta$  such that the probability of the document being about the end of the world is 0 (or incredibly close)?

**Question 3.8.** Show the two standard definitions of the logistic sigmoid function are equivalent.

**Question 3.9.** In Naive Bayes, if you increase the pseudocount hyperparameter, does your model tend to underfit or overfit more?

**Question 3.10.** In logistic regression, if you increase the L2 norm regularization hyperparameter, does your model tend to underfit or overfit more?

## 4 Distributional and word embedding representations

**Question 4.1.** What is the range of possible values for cosine similarity?

**Question 4.2.** Under what condition is it the case that cosine similarity is equal to one? Show why it's equal to one in this case.

**Question 4.3.** What is one advantage of using averaged word embeddings, instead of bag-of-words, for a document classifier?

**Question 4.4.** What is one disadvantage of using averaged word embeddings, instead of bag-of-words, for a document classifier?

**Question 4.5.** Show that cosine similarity is invariant to rescaling one of its inputs. (Rescaling a vector means, multiplying by a constant scalar.)

## 5 Evaluation

**Question 5.1.** INLP chapter 4, Exercise 2 [Note: too hard to be a test question]

**Question 5.2.** INLP chapter 4, Exercise 3



#### 4 Distributional and word embedding representations

Question 4.1. What is the range of possible values for cosine similarity?

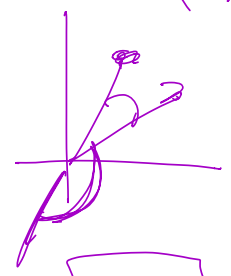
Question 4.2. Under what condition is it the case that cosine similarity is equal to one? Show why it's equal to one in this case.

Question 4.3. What is one advantage of using averaged word embeddings, instead of bag-of-words, for a document classifier?

Question 4.4. What is one disadvantage of using averaged word embeddings, instead of bag-of-words, for a document classifier?

Question 4.5. Show that cosine similarity is invariant to rescaling one of its inputs. (Rescaling a vector means, multiplying by a constant scalar.)

"Bag of embeddings"



$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$$\cos(x, x) = \frac{\sum_i x_i^2}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i x_i^2}}$$

4.3

BoE vs BoW  
 fewer dims (100 vs 100k)  
 ↳ comp.  
 ↳ less sparse  
 Info about other words  
 not in training set  
 but their embeddings are similar

4.4

BoW vs BoE  
 ↳ comp. Embeddings matrix is big



→ what if: specific words matter?

= Embeddings are bad?

↳ Antonyms

↳ Domain shift

## 6 Misc

**Question 6.1.** Please write a regular expression to match any word that has 3 or more instances of the same vowel in a row, like sooooo or haaaaa. (Assume there are 5 vowels: a, e, i, o, u.)

~~\_\_\_\_\_~~

(a|e|i|o|u){3}

~~raae~~

a{3}      aaa

[A-Za-z]\* aaaa\* [A-Za-z]\*

[~~a-z~~]\* (aaa+) | (eee+) | (iii+) | (ooo+) | (uuu+) [~~a-z~~]\*

- non-alphabetic chars? Numbers?

- \w vs \W

## 6 Misc

**Question 6.1.** Please write a regular expression to match any word that has 3 or more instances of the same vowel in a row, like *sooooo* or *haaaaha*. (Assume there are 5 vowels: a, e, i, o, u.)

**Question 6.2.** Consider training a supervised document classifier for sentiment, and compare it to a lexicon counting classifier. If you have a very low number of labeled documents, which model do you expect to be better? If you have a very high number of labeled documents, which model do you expect to be better? Why?

## 7 Text preprocessing (HW2 related)

**Question 7.1.** What is the difference between tokenization and word normalization? (Not vector norms or probability normalization.) Please list a few examples of word normalization.

**Question 7.2.** Why is word normalization used?

**Question 7.3.** (A) What is the difference between lemmatization and stemming? (B) Give a justification why lemmatization may be preferred to stemming. (C) Give a justification why stemming may be preferred to lemmatization.

**Question 7.4.** Why is word normalization used?

**Question 7.5.** What's a pro and a con of using n-gram features, as opposed to bag-of-words features, for a classifier?

## 8 More questions

**Question 8.1.** What's a useful thing you can calculate with NB, without having to calculate  $p(x)$ ?

**Question 8.2.** What's a useful thing you can calculate with NB, but requires you to calculate  $p(x)$ ?

**Question 8.3.** For Naive Bayes with many classes, consider the case where you only care about the ratio between the posterior probabilities for two classes, say, class C and D. Demonstrate (and show your work) that you do not need to calculate the Bayes Rule normalizer  $p(x)$  to calculate this posterior ratio,

$$\frac{p(y = C | x)}{p(y = D | x)}$$

**Question 8.4.** Say you have NB for a binary classification problem. You retrain the model lots of times, and each time you make the pseudocount hyperparameter higher and higher. With each model you do predictions on new data. What happens to Naive Bayes predicted document posteriors as the pseudocount goes higher? [HINT: you can just do this intuitively. It may help to focus on the  $P(w|y)$  terms. A rigorous, if overkill, approach, is to use L'Hospital's rule.]

- (a) They all become either 0 or 1.
- (b) They all become 0.5.
- (c) They all become the class prior.
- (d) There is no stable trend in all situations.

**Question 8.5.** In the typical case, how does the number of parameters compare between BOE (bag of embeddings) and BOW (bag of words)?

- (a) BOW has more parameters than BOE
- (b) BOE has more parameters than BOW
- (c) They are the same

**Question 8.6.** In the typical case for English, how does the number of parameters compare between BOW versus a model where features are counts of character 10-grams?

- (a) BOW has more parameters than character 10-grams
- (b) Character 10-grams has more parameters than BOW
- (c) They are the same

**Question 8.7.** What is an issue if you want to apply BOW to Chinese documents?

**Question 8.8.** Consider an annotation task with 5 items and 2 annotators, for binary classification. Both annotators annotated all items. Draw up a  $5 \times 2$  matrix of their annotations and fill in any values you like, as long as agreement is less than 100%. For all the following questions, show your work. (A) What is the agreement rate? (B) What is the random chance agreement rate? (Use the overall prevalence of classes among all annotations.) (C) Calculate Cohen's kappa.

**Question 8.9.** What is the range of possible values of Cohen's kappa?

**Question 8.10.** Give an example of a task where Cohen's kappa might be high, and one where it might be low. Why the difference?

**Question 8.11.** Constituency and dependency trees focus on different aspects of a sentence's syntactic structure. What does a constituency tree focus on? What does a dependency tree focus on?

dep tree: relationships between words

constit. tree: hierarchical phrases of words

~~2~~

$$p(y=1|x) \equiv$$

$$= C \cdot p(y=1) \prod_i p(w_i/y=1)$$

density  
change

similar

$$p(y=0|x) = C$$

$$p(y=0) \prod_i p(w_i/y=0)$$



**Question 8.12.** Draw a lexicalized constituency tree for an example sentence. (For example, use Figure 12.11 in SLP3.) Draw the unlabeled dependency tree it corresponds to.

**Question 8.13.** Give a simple CFG that can parse the following POS-tagged sentence, with an analysis conforming to the standard type of grammar used in your readings (broadly similar to the Penn Treebank style). You can exclude unary expansions from POS tags to the lexicon; assume POS tags are given to the parser as input. Draw the corresponding parse tree.

(PRP I) (VB run) (ADJ fast)

**Question 8.14.** Amend your CFG so it can also parse this sentence. Draw the corresponding parse tree.

(PRP I) (VB run) (ADJ fast) (IN on) (NNPS Mondays)

