CSCI-GA-2590 Fall 2021 Midterm

Name and section: _____

1 True or False

1. (1 point) CRF is a discriminative model.

   √ **True** ◯ False

2. (1 point) Compared to n-gram language model, feed-forward neural network language model handles out-of-vocabulary words naturally.

   ◯ True √ **False**

3. (1 point) The IDF term in TFIDF upweights words that occur in only a few documents.

   √ **True** ◯ False

4. (1 point) The ambiguity of language is one of the key challenges in NLP.

   √ **True** ◯ False

5. (1 point) Generative models make less assumption about the data generating process than discriminative models.

   ◯ True √ **False**

6. (1 point) MLE for logistic regression has a concave objective.

   ◯ True √ **False**

7. (1 point) Word embeddings can be learned from unlabeled corpus.

   √ **True** ◯ False

8. (1 point) Backpropogation is a dynamic programming algorithm.

   √ **True** ◯ False

9. (1 point) At inference time, the time complexity of the multinomial naive Bayes classifier is linear in the number of words in the input.

   √ **True** ◯ False

10. (1 point) We can obtain both word vectors and document vectors by performing SVD on the word × document matrix.

    √ **True** ◯ False

2 Multiple Choices

Note: There can be more than one correct answers; select all that apply! No partial credits: all correct answers must be checked.

1. (2 points) Which of the following is/are true about multinomial naive Bayes classifiers?

   ◯ Each feature must be a single token.
   √ **It has closed form solution under MLE.**
   √ **The number of parameters is linear in the number of features.**
   ◯ A feature is assumed to be independent of other features.

2. (2 points) Consider the toy sentiment classification corpus:

this is a great movie	(label: +1)
great action movie	(label: +1)
such a boring action movie	(label: -1)

   Using a naive Bayes classifier trained on the corpus, what is the probability of the sentence *action movie* being positive?

   ◯ 1/48
   √ **75/123**
   ◯ 25/57
   ◯ 1/32

3. (2 points) Consider the logistic regression model: $p(y = 1 \mid x) = \frac{1}{1+e^{-wx}}$ where $w, x \in \mathbb{R}, y \in \{1, 0\}$. During MLE using SGD, when would the gradient on example (x, y) be zero? [CORRECTION: assuming the learning rate is 1]

   ◯ $x = 0$
   √ $y = 0$ **and** $wx \to -\infty$
   ◯ $y = 0$
   √ $p(y \mid x) = 1$

4. (2 points) You have trained a logistic regression classifier using n-gram ($n \geq 1$) features and observed overfitting on the validation set. What should you do to mitigate the problem?

   √ **Decrease n (as in n-gram)**
   √ **Add L2 regularization**
   √ **Collect more training data**
   ◯ Collect more validation data

5. (2 points) Which of the following is/are true about recurrent neural networks (RNNs)?

   √ **RNNs can handle sequences of arbitrary length.**
   ◯ Gradient clip can be used to mitigate the vanishing gradient problem.
   √ **RNNs can be trained by truncated backpropagation.**
   ◯ Compared to RNNs, LSTMs are easier to train because they do not have recurrent states.

6. (2 points) Consider add-α smoothing for n-gram language models. What is the effect of increasing α on the probability of unseen n-grams?

   ◯ No effect. The probability of unseen n-grams stays the same.
   ◯ The probability of unseen n-grams will decrease.
   √ **The probability of unseen n-grams will increase.**
   ◯ Not sure. It depends on n.

7. (2 points) Alice and Bob are evaluating a language model. Alice computed the average held-out log likelihood, $\ell = \frac{1}{n}\sum_{i=1}^{n} \log_2 p_i$ (where p_i denotes the probability of the i-th word) and sent the value to Bob. Bob then computed the perplexity by $2^{-\ell}$. However, Alice then found that she had mistakenly used the natural log when computing ℓ. Let a be perplexity that Bob computed using the wrong value sent by Alice. How should he correct it to get the right perplexity?

   ◯ a^e
   √ $a^{\log_2 e}$
   ◯ $a^{\frac{1}{e}}$
   ◯ $a^{\frac{1}{\log_2 e}}$

8. (2 points) Consider the sentence:

   She likes to dress her kid in a blue dress

   Which of the following is/are true?

   √ **This sentence is ambiguous (syntactically).**
   ◯ There are 10 word types in this sentence.
   ◯ There are 5 bigrams in this sentence.
   √ **A dictionary-based POS tagger can get an accuracy of 90% at most on this sentence.**

3 Written Questions

You can either type your answers in the text box or write it on paper and upload an image. If you need to use LaTex to type equations, put it inside $$...$$.

1. (a) (3 points) Consider the labeled training example:

   the/DT brown/JJ fox/NN jumped/VBD over/IN the/DT lazy/JJ dog/NN

   Given the feature template $T(x_i, y_i, y_{i-1})$, write down three features you can extract from the above example. (Assume the start symbol is $*$ and the stop symbol is `STOP`)

   Solution: $\mathbb{I}(x_i = \text{the}, y_i = \text{DT}, y_{i-1} = *)$

(b) (2 points) Suppose the size of the tag set is 50, the size of the vocabulary is 1,000, and the total number of tokens in the training set is 2,000 (including the start/stop symbols). What's your best estimate of the maximum number of features (that will have non-zero weights) using the above feature template $T(x_i, y_i, y_{i-1})$?

Solution: 2,000

(c) (2 points) Give at least one example where it is useful to include features that look at the previous *two* tags (i.e. $T(x_i, y_i, y_{i-1}, y_{i-2})$).

Solution: borrow/VB book/NN from/IN ...
the/DT book/NN is/VB ...
When predicting tags for "from" or "in" in the above examples, it is useful to look at whether it is VB or DT before the previous tag NN.

(d) (3 points) The Viterbi decoding algorithm we learned in class only works with tag bigram features (i.e. feature that depend on the current tag y_i and the previous tag y_{i-1}). Explain in words how you would extend the algorithm to handle tag trigram features that use the previous two tags (HINT: you will need to modify the chart $\pi[\cdot, \cdot]$).

Solution: The second argument in π (which represented the ending tag) now represents a pair of tags.

(e) (3 points) Let t be the tag set size and m be the sequence length. What's the time complexity of your modified Viterbi decoding algorithm (from the previous question)?

Solution: $O(mt^3)$. The lattice now has t^2 (i.e. number of tag pairs) rows, and each maximization is over t transitions (i.e. y_{i-1}, y_i to y_i, y_{i+1}).

2. You are trying to use logistic regression to classify the following dataset:

input	label
beauty	+1
creator	+1
realtor	+1
naughty	-1

Here the positive label means that the input word is a noun and negative label means it is not a noun. You decide to use a single feature $\mathbb{I}(x$ ends with "ty") for the task.

(a) (3 points) Let $x \in \mathbb{R}$ be the feature representation of the input and $y \in \{+1, -1\}$ be the label. Recall that in class we used the following parametrization of the

logistic regression model (where $w \in \mathbb{R}$ is the model parameter)

$$p(y = 1 \mid x; w) = \frac{1}{1 + e^{-wx}} ,$$
$$p(y = -1 \mid x; w) = 1 - p(y = 1 \mid x; w) .$$

Show that we can combine the two equations by writing

$$p(y \mid x; w) = \frac{1}{1 + e^{-wxy}} .$$

Solution:

(b) (2 points) Using the new parametrization in the previous question. Write down the loss function $\ell_i(w)$ (i.e. the negative log likelihood) for a *single* training example (x_i, y_i).

Solution:

$$\ell_i(w) = \log(1 + e^{-wx_iy_i})$$

(c) (3 points) Write down the gradient of $\ell_i(w)$.

Solution:

$$\nabla_w \ell_i(w) = -[1 - p(y_i \mid x_i; w)]x_iy_i$$

(d) (2 points) Let 0 be the initial value of $w_0 = 0$. Using SGD, give the value of w after one update on the example "beauty".

Solution:

$$w = 0.5$$

(e) (3 points) Let 0 be the initial value of $w_0 = 0$. Now using GD (i.e. the gradient is now computed on the whole dataset), give the value of w after one update. Is this the optimal value of w? [CORRECTION: assuming the learning rate is 1]

Solution: $w = 0$. Yes, because the gradient is zero.

(f) (3 points) Will your model achieve zero *training error*? If not, provide one additional feature that will allow you to achieve zero training error on this dataset.

Solution: No. One example feature: $\mathbb{I}$("ea" is a substring of the word).

3. Consider the following RNN:

$$\begin{aligned} h_0 &= 0 \ , \\ h_t &= f(w_1 h_{t-1} + w_2 x_t + b1) \ , \\ y_t &= g(w_3 h_t + b_2) \ , \end{aligned}$$

where the activation function f is

$$f(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 & \text{otherwise} \end{cases} \ ,$$

and g is the identity function

$$g(x) = x \ .$$

(a) (5 points) Consider a binary input sequence (i.e. $x_t \in \{0, 1\}$). We would like the model to output 0 until it receives a 1, upon which it switches to output 1 for *all* subsequent time steps. For example, given an input sequence 001010, the model should output 001111. Give values of w_1, w_2, w_3, b_1, b_2 that produce the desired model behavior. If such a solution does not exist, explain why.

Solution: Let $w_3 = 1$ and $b_2 = 0$ so that $y_t = h_t$.
Let $h_t = 0$ represent the state that there has been no "1" in the sequence so far. Let $h_t = 1$ represent the state that there has been at least one "1" in the sequence so far. We have four cases:

- $h_t = 0, x_t = 0, h_{t+1} = 0$: $w_1 \times 0 + w_2 \times 0 + b_1 < 0$
- $h_t = 0, x_t = 1, h_{t+1} = 1$: $w_1 \times 0 + w_2 \times 1 + b_1 \geq 0$
- $h_t = 1, x_t = 0, h_{t+1} = 1$: $w_1 \times 1 + w_2 \times 0 + b_1 \geq 0$
- $h_t = 1, x_t = 1, h_{t+1} = 1$: $w_1 \times 1 + w_2 \times 1 + b_1 \geq 0$

One solution is $b_1 = -1, w_2 = 1, w_1 = 1$.

(b) (2 points) Give at least one reason why the activation function f defined above is undesirable in practice.

Solution: Not differentiable.

(c) (3 points) Suppose now f is the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}} \ .$$

What would the derivative $\frac{df}{dx}$ be when x is very large? Why this might be a problem for training neural networks?

Solution: Goes to zero. If w is initialized to be very large, then the model gets no gradient before it has learned anything.

(d) (4 points) Let's use the ReLU function for f now, i.e. $f(x) = \max(0, x)$. Give values of w_1, w_2, w_3, b_1, b_2 that produce the desired model behavior. If such a solution does not exist, explain why.

Solution: Let $h_t = 0$ represent the state that there has been no "1" in the sequence so far. Let $h_t = c$ represent the state that there has been at least one "1" in the sequence so far, where c is some positive number (note that $h_t \geq 0$). Following the same argument, we have

- $h_t = 0, x_t = 0, h_{t+1} = 0$: $w_1 \times 0 + w_2 \times 0 + b_1 < 0$
- $h_t = 0, x_t = 1, h_{t+1} = c$: $w_1 \times 0 + w_2 \times 1 + b_1 = c$
- $h_t = c, x_t = 0, h_{t+1} = c$: $w_1 \times c + w_2 \times 0 + b_1 = c$
- $h_t = c, x_t = 1, h_{t+1} = c$: $w_1 \times c + w_2 \times 1 + b_1 = c$

There is no solution satisfying all constraints.