COSI 134 (Fall 2020): Quiz questions

NAME:
DATE: Dec 3, 2020

Instruction: You only need to answer 10 of the following questions. If you need extra space, you can use a blank page, but make sure you clearly indicate which question you are answering.

1. **10pts** Explain the limitation of the conditional independence assumptions of Naïve Bayes classifiers in terms of using more features for the model.

2. **10pts** A logistic regression model defines a posterior distribution as $p(y|\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_i^N \theta_i f_i(x, y)\right)$ where Z is the partition function. Write an expression for Z.

3. **10pts** Prove that the softmax and sigmoid functions are equivalent when the number of possible labels is two. Specifically, for any $\boldsymbol{\Theta}^{(z \to y)}$ (omitting the offset $\boldsymbol{b}$ for simplicity), show how to construct a vector of weights $\boldsymbol{\theta}$ such that

$$\text{SoftMax}(\boldsymbol{\Theta}^{(z \to y} \boldsymbol{z})[0] = \sigma(\boldsymbol{\theta} \cdot \boldsymbol{z})$$

4. **10pts** Write an expression for a multi-level feedforward neural network with two hidden layers. Specify the dimensions of the input, the weight matrices, as well as the biases.

5. **10pts** When training a CBOW model, the output requires a softmax over all words in the vocabulary of a language, which is computationally expensive. Name two ways that make a CBOW model more efficient, and explain how it works.

6. **10pts** What is a filter in a Convolutional Network? Why does a pooling layer need to be applied to the convolution layer before its output can be used for classification?

7. **10pts** The pseudocode below illustrates the perceptron learning algorithm for sequence labeling. Use an example to explain how the paremeters Θ are updated for each training instance. Your example needs to include the correct tag sequence for a sentence, and the predicted output where there is at least one error.

The averaged perceptron algorithm

procedure AVE_PERCEPTRON($\boldsymbol{x}^{1:N}, y^{1:N}$)
  $t \leftarrow 0, \boldsymbol{\theta}^{(0)} \leftarrow \mathbf{0}, \boldsymbol{m} \leftarrow \mathbf{0}$
  repeat
    $t \leftarrow t + 1$
    Select a sequence i ▷ Online training
    $\hat{\boldsymbol{y}} \leftarrow \arg\max_{\boldsymbol{y}} \boldsymbol{\theta}^{(t-1)} \cdot \boldsymbol{f}(\boldsymbol{w}^{(i)}, \boldsymbol{y})$ ▷ Decoding by Viterbi
    if $\hat{\boldsymbol{y}} \neq \boldsymbol{y}^{(i)}$ **then**
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} + \boldsymbol{f}^{(global)}(\boldsymbol{w}^{(i)}, \boldsymbol{y}^{(i)}) - \boldsymbol{f}^{(global)}(\boldsymbol{w}^{(i)}, \hat{\boldsymbol{y}})$
    else ▷ Feature count for entire sequence
      $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$
    end if
    $\boldsymbol{m} \leftarrow \boldsymbol{m} + \boldsymbol{\theta}^{(t)}$
  until tired
  $\bar{\boldsymbol{\theta}} = \frac{1}{t}\boldsymbol{m}$
  return $\bar{\boldsymbol{\theta}}$
end procedure

8. **10pts** Consider the garden path sentence, *The old man the boat.* Given word-tag and tag-tag features, what inequality must in the weights must hold for the correct tag sequence to outscore the garden path tag sequence for this example?

9. **10pts** Let $\alpha(\cdot)$ and $\beta(\cdot)$ indicate the forward and backward variables in the forward-backward algorithm. Show that $\alpha_{M+1}(\blacklozenge) = \beta_0(\lozenge) = \sum_y \alpha(y)\beta_m(y), \forall m \in \{1, 2, \cdots, M\}$

10. **10pts** To handle VP coordination, a grammar includes the production VP → VP CC VP. To handle adverbs, it also includes the production VP → VP ADV. Assume all verbs are generated from a sequence of unary productions, e.g., VP → V → *eat*.

    - Show how the binarize the production VP → VP CC VP.
    - Use your binarized grammar to parse the sentence *They eat and drink together*, treating *together* as an adverb.
    - Prove that a weighted CFG cannot distinguish the two possible derivations of this sentence. Your explanation should focus on the productions in the non-binary grammar.
    - Explain what condition must hold for a parent-annotated WCFG to prefer the derivation in which *together* modifies the coordination *eat and drink*.

11. **10pts** Assuming the following grammar:

S	→	NP VP
VP	→	V NP
NP	→	JJ NP
NP	→	*fish* (the animal)
V	→	*fish* (the action of fishing)
JJ	→	*fish* (a modifier, as in *fish sauce* or *fish stew*)

    Show how the sentence "Fish fish fish fish" can be derived with a series of shift-reduce actions.

12. **10pts**

    Define the actions in a "arc-standard" transition-based dependency parsing system. What constraints need to be applied to ensure the resulting dependency tree is well-formed?