COSI 134 (Fall 2020): Sample quiz questions

NAME:

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

2. 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. What is L2 regularization and how is it different from L1 regularization?

4. Write an expression for a multilayer feedforward neural network with two hidden layers. Specify the dimensions of the input, the weight matrices, as well as the biases.

5. Write down the mathematical expression for the "momentum" optimization and explain how and why it improves the gradient descent algorithm.

6. Explain what is the input, the hidden layer, and the output for a CBOW Word2Vec model. Specify the dimensions of the weight matrices of the model.

7. 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

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

8. 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?

9. A Recurrent Neural Network is a flexible model that is capable of addressing many NLP tasks. What is an appropriate RNN for POS tagging? What is an appropriate model for Machine Translation? Write down the mathematical expressions for each model, and explain the dimensionality of each weight matrix, bias, input layer, hidden layer, and output layer where appropriate.

10. Consider a recurrent neural network with a single hidden unit and a sigmoid activation, $h_m = \sigma(\theta h_{m-1} + x_m)$. Prove that the gradient $\frac{\partial h_m}{\partial h_{m-k}}$ goes to zero as $k \to \infty$.

11. The problem of sequence labeling typically involves finding the tag sequence that has the highest score given an observation sequence (say a sequence of words). In HMM-based sequence labeling, given a matrix of transition probabilities between two tags $P(t_i|t_{i-1})$ and a matrix of emission probabilities $P(w_i|t_i)$, where i is the time step, w_i is the observed word token at i, and t_i is the tag for w_i, can you find the tag sequence for the sentence with the highest score by doing greedy search, that is, finding the tag t_i with the highest score at each time step? Why or why not? Explain with an example.

$$\hat{t}_i = \arg\max_t P(t_i|t_{i-1})P(w_i|t_i)$$

12. 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?

13. Show how to compute the marginal probability $Pr(y_{m-2} = k, y_m = k'|\boldsymbol{w}_{1:M})$ in terms of the forward and backward variables, and the potentials (local scores) $s_n(y_n, y_{n-1})$.

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

15. Name and briefly describe *two* independence assumptions associated with PCFGs

16. 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.
    - Provide 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 conditions must hold for a properly annotated WCFG to prefer the derivation in which *together* modifies the coordination *eat and drink*.

17. 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.

18. Attention is an important concept in neural network based models. Given the encoder-decoder example in Figure 1, write down the expressions used to compute the context vector for $\mathbf{h}_2^{tgt}$:

19. Can transition-based constituent parsing be paired with the CKY decoder? If not, which decoder should be used? Explain how the decoder works

20. 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?

21. Provide the UD-style unlabeled dependency parse for the sentence *Xi-Lan eats shoots and leaves*, assuming *shoots* is a noun and *leaves* is a verb. Provide arc-standard and arc-eager derivations for this dependency parse.

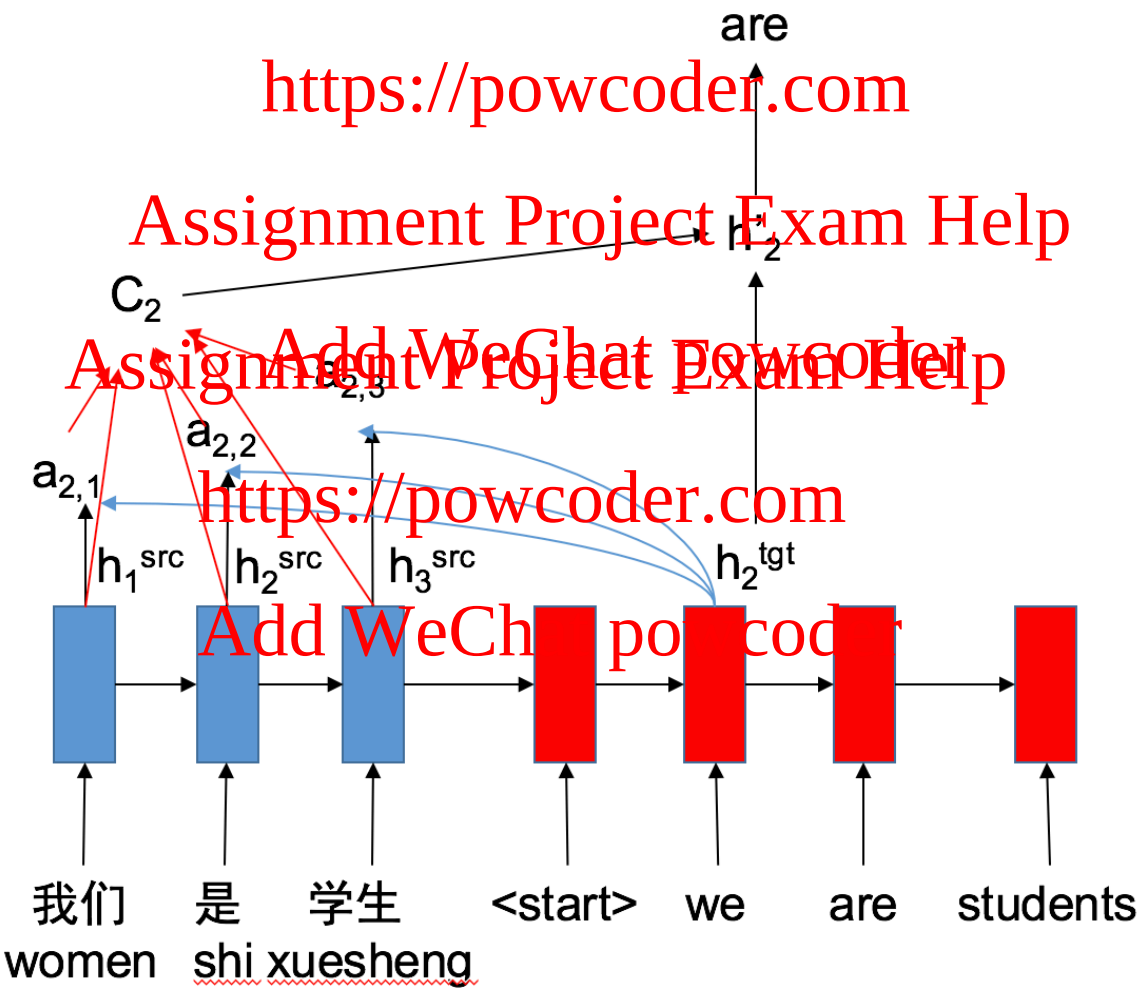

Figure 1: Encoder-decoder with attention