COSI 231 (Fall 2025): Sample quiz questions

NAME:

1. 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.

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

3. Let $\hat{\boldsymbol{\theta}}$ be the solution to an unregularized logistic regression problem, and let $\boldsymbol{\theta}^*$ be the solution to the same problem, with L_2 regularization. Prove that $||\boldsymbol{\theta}^*||_{\boldsymbol{2}}^{\boldsymbol{2}} \leq ||\hat{\boldsymbol{\theta}}||_{\boldsymbol{2}}^{\boldsymbol{2}}$.

4. 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,

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

5. Explain the difference in how the parameters are updated between a Logistics Regression and a perceptron model.

6. Explain what "error signal" is in a feedforward neural network and how error signals are "cached".

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

8. 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. Why is it a "degenerate" neural network?

9. 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 they work and how they improve the efficiency of the model.

10. 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?

11. Explain how Perplexity is computed as a metric for language models. For a language with a vocabulary of size V, assume the words in the language are uniformly distributed, meaning that the probability for each word in the language is $\frac{1}{V}$, what is the perplexity for a unigram model of this language?

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

13. What are the gates in an LSTM model and how are they defined mathematically?

14. 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 $k \to \infty$.

15. Explain the attention mechanism in an RNN based sequence-to-sequence model for machine translation. What are the variants?