Indian Institute of Technology, Guwahati
Department of Computer Science and Engineering

Mid Semester Examination
Semester I, 2021-22
CS 565: Intelligent Systems and Interfaces

Total Marks: 40 Time: 1 Hours

Instructions: No doubts will be entertained. If you feel question is incorrect, or incomplete or required assumptions, please write your justifications clearly.

1. Zipf's law states that if the word types in a corpus are sorted by frequency, then the frequency of the word at rank r is proportional to r^s, where s is a free parameter, usually around 1. Solve for s using the counts of the first and second most frequent words, c_1 and c_2. [5]

2. For a trigram language model, show the maximum likelihood estimate of conditional probability $\hat{p}_{mle}(w_i|w_j, w_k) = \frac{c(w_j,w_k,w_i)}{c(w_j,w_k)}$, where c indicates count. Your derivation must include all required assumption. [10]

3. Consider a unigram language model over a vocabulary of size V. Assume that a word appears m times in a corpus with M tokens in total. With additive smoothing of α, for what values of m, is the smoothed probability greater than the unsmoothed probability? [5]

4. Discuss intuition and advantage of backoff smoothing discussed in the class over additive smoothing techniques. [5]

5. (a) Let $\mathbf{x}$ and $\mathbf{w}$ are two vectors in $\mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{m \times d}$ be a matrix. Prove the following [5]

$$\mathbf{w} \cdot (\mathbf{A}\mathbf{x}) = \mathbf{w}' \cdot \mathbf{x}$$

where $\mathbf{w}' = \mathbf{A}^T w$ and $\mathbf{A}^T$ is the transpose of the matrix.

(b) Consider a single layer neural network defined as [5]

$$\phi(\mathbf{x}; \theta) = g(\mathbf{W}\,\mathbf{x} + \mathbf{b})$$

where $x \in \mathbb{R}^d, W \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m$ and g is defined as

$$g(\mathbf{z}) = \alpha \times \mathbf{z} + \mathbf{c}$$

where $\alpha \in \mathbb{R}$ is a constant, and $\mathbf{c} \in \mathbb{R}^m$ is a vector. For the output layer, we define conditional probability as

$$p(y|\mathbf{x}; \theta, \mathbf{v}) = \frac{\exp\{\mathbf{v}(\mathbf{y}) \cdot \phi(\mathbf{x}; \theta) + \gamma_y\}}{\sum_{y'} \exp\{\mathbf{v}(\mathbf{y'}) \cdot \phi(\mathbf{x}; \theta) + \gamma_{y'}\}}$$

Show that for any parameter value $\mathbf{v}(\mathbf{y}) \in \mathbb{R}^d$ and γ_y for $y \in \mathcal{Y}$, there are parameter values $\mathbf{v}'(\mathbf{y})$ and γ'_y such that for all $\mathbf{x}, y$,

$$p(y|\mathbf{x}; \theta, \mathbf{v}) = \frac{\exp\{\mathbf{v}'(\mathbf{y}) \cdot \mathbf{x} + \gamma'_y\}}{\sum_{y'} \exp\{\mathbf{v}'(\mathbf{y'}) \cdot \mathbf{x} + \gamma'_{y'}\}}$$

(c) Consider the function $L : \mathbb{R}^m \to \mathbb{R}^m$ defined component-wise as [5]

$$L_{y_i}(\mathbf{l}) = l_{y_i} - \log \sum_{y_j \in \mathcal{Y}} \exp\{l_{y_j}\}$$

(a) Find out the value for $\frac{\partial L_{y_i}(l)}{\partial l_{y_i}}$.

(b) Find out the value for $\frac{\partial L_{y_i}(l)}{\partial l_{y_j}}$ for $j \neq i$.