BRAC UNIVERSITY
Department of Computer Science and Engineering

Examination: Midterm Semester: Fall 2024
Duration: 75 minutes Full Marks: 30

CSE 440: Natural Language Processing II

Figures in the right margin indicate marks.

Answer all 3

1. A. You are training a deep neural network, and the training accuracy is 98%, but the validation accuracy is only 70%. What does this indicate, and how can you address it? [4]

   B. What is the impact of increasing the number of features in a dataset on underfitting and overfitting? Explain. [2]

   C. Is overfitting always bad? Discuss one scenario where a slightly overfit model might be acceptable. [4]

2. A. For a dataset with 3 examples, the target values y and predicted probabilities y' are given as follows: y = [1,0,1], y'= [0.9,0.1,0.6]. Calculate the average binary cross-entropy loss for this dataset. [6]

   B. Describe: [4]
      a. Conditional independence and the Naive assumption of the Naive Bayes algorithm.
      b. How Naive Bayes algorithms deal with unseen features (e.g. words that were not seen during training)

3. A. A company is analyzing customer reviews using NLP techniques. Consider the following matrix for three documents and three terms: [8]

Document	Term: "product"	Term: "excellent"	Term: "price"
d1	3	0	1
d2	1	2	0
d3	0	1	4

   Calculate TF-IDF for every term in every document.

   B. Explain how the TF-IDF values would change if the term "price" appeared in every document. [2]