MTH 325, Quiz #8, Spring 2023 Name _____

Instructions: Answer each question as thoroughly as possible. Round answers to 4 decimal places as needed. Exact answers are best when possible. Be sure to answer all parts of each question.

1. Describe the difference between supervised learning and reinforcement learning.

Supervised learning is a type of machine learning where the algorithm learns from labeled data, which means the data has input features (independent variables) as well as corresponding output labels (dependent variables). Reinforcement learning, on the other hand, is a type of machine learning where the algorithm learns by interacting with an environment. The algorithm receives rewards or penalties for the actions it takes in the environment, and its goal is to learn the optimal sequence of actions that maximizes the cumulative reward. Reinforcement learning is often used in decision-making problems where there is no labeled data available, and the agent learns by trial and error.

2. Explain how factor analysis is related to dimensionality reduction.

Factor analysis is a statistical technique that aims to identify latent variables (also called factors) that explain the correlations among a set of observed variables. In essence, factor analysis attempts to identify the underlying structure of a set of variables by grouping them into a smaller set of factors that explain the majority of the variability in the original data. Dimensionality reduction is a broader concept that refers to the process of reducing the number of variables in a dataset while retaining as much of the original information as possible. This can be achieved through various techniques, such as principal component analysis (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF). Factor analysis is a type of dimensionality reduction technique that specifically aims to identify latent variables that can explain the correlations among the observed variables. By reducing the number of observed variables to a smaller set of factors, factor analysis can simplify the data analysis process and make it easier to interpret the underlying structure of the data.

3. Describe two advantages and at least one disadvantage of ensemble methods.

Answers may vary

Advantages: Improved accuracy and Robustness.
Disadvantage: Complexity.

4. Describe k-fold cross validation and why we use it to validate models.

K-fold cross-validation is a technique used to validate the performance of a predictive model. In k-fold cross-validation, the dataset is divided into k equal-sized subsets or folds. One fold is used as the validation set, and the remaining k-1 folds are used as the training set to fit the model. This process is repeated k times, with each fold serving as the validation set once. K-fold cross-validation is used because it provides a more reliable estimate of the model's performance than a single train-test split. By using multiple folds, we can assess the model's generalization ability, ensuring that it does not overfit or underfit the data. Additionally, it allows us to make more efficient use of the available data by using each observation both for training and validation.