CS 559: Machine Learning: Fundamentals and Applications

Assignment 5 Due: 11/20/2024 Wednesday 11:59 p.m.

- The assignment must be individual work and must not be copied or shared. Any tendency to cheat/copy evidence will lead to a 0 mark for the assignment.
- Students must only use Pandas, NumPy, Matplotlib, and Scipy if the problem does not specify libraries/packages.
- All codes will be tested in grading. Any codes with an error will be marked 0. Make sure to restart the kernel and run it all before the submission. Delete any codes that do not want to be graded.
- Results must be displayed.
- All problems must be submitted in a single notebook file. Do not use a text editor to write codes.

1 Unsupervised Learning [65 pts]

Supervised learning techniques cannot be applied when a given data set does not have feature names or unknown targets. Instead, we can apply unsupervised learning techniques such as clustering analysis or dimensionality reduction to learn about the hidden structures of the data set. Sometimes, such methods can be applied in EDA or preprocessing the training data to build a supervised learning model. In this question, we will practice implementing, fitting, and analyzing unsupervised learning algorithms.

1.1 Clustering Analysis [40 pts]

a. Load the provided data - **autompg.csv**. The data contains seven columns; the target is **mpg**. This data set can be used for both classification and regression problems.

b. [5 pts] Relabel **mpg** into three ordinal domains [good (2), average (1), bad (0)] by calculating the quantiles using **numpy.quantitle**. For other continuous columns, convert to four ordinal categorical features as (4) - high, (3) - good, (2) - okay, (1) - poor.

c. [10 pts] Implement a **myKmean**(X, k, iter) algorithm as a **function** (not a class) using only **NumPy**. The parameter X is the data, k is the number of clusters, and iter is the number of iterations. The function returns the final cluster labels, the centroid of clusters, and the total variance of the algorithm.

d. [10 pts] Determine the appropriate k value of the new data set using **myKmean**. Show that the total variance converges as k increases.

f. [15 pts] Using the results of **myKmean** with the best k value from d, build a linear classification model using scikit-learn that predicts the cluster label. Discuss any noticeable characteristics in each cluster. Guess any features that are correlated to each other based on the weight values.

e. Save the data set and the cluster label. It will used in the BN problem,

1.2 Dimensinoality Reduction [25 pts]

a. Load the provided data - **pca_data.csv**.

b. [10 pts] Study the data and hypothesize the appropriate number of dimensions (n) that the data can be reduced to. Use statistics and visualizations to support your answer.

c. [10 pts] Implement a function **myPCA(X,n)**(no class) that takes the data X and n is the number of components found in b. The function will return the principal components.

d. [5 pts] Perform **myPCA** and compare the result with scikit-learn PCA.

2 Bayes Networks [35 pts]

In this problem, we will build a BN model that predicts the cluster labels. We will practice making conditional independence assumptions and building a model based on the assumptions. The table below is to

Variable	mpg	cylinders	displacement	horsepower	weight	acceleration	model-year	cluster label
Denotation	M	C	D	H	W	A	Y	L

a. [10 pts] Assume the following BN model. Draw the layout of the graph and construct the conditional probability tables.

$$P(L, M, W, C) = P(L|M, W, C)P(M|W)P(M|C)P(W|C)P(C)P(Y|C)P(A|Y)P(A|C) \qquad (1)$$

b. [10 pts] Verify the conditional independencies in the graph. Justify your answer.

c. [10 pts] Build a BN model to predict L using pgmpy.