DS4400 Project Proposal

For this project, we plan to introduce a new modified approach to the traditional Oblique Decision Tree algorithm, the Deep Oblique Decision Tree (DODT).

We are going to compare the performance of DODT against two traditional classification algorithms.

1. Traditional Decision Tree (TDT)
2. K-Nearest Neighbors (KNN)

In particular, the major difference between our approach and ODT is that ODT uses Perceptron only for binary classification on each node, while DODT utilizes Perceptron for both classification and feature processing.

A deep Oblique Decision Tree uses Perceptron in a way that is similar to Multi-Layer Perceptron, not merely processing the feature vector to be the binary label, but also convert to a new feature vector. In each node, despite the samples being classified as 1, other samples are passed to the next layer in the form of raw feature vectors (before summing up), and theoretically, the Perceptron can also be used as normalizing the feature vectors and removing the set of samples, belong to the same class, with similar potential characteristics before passing to the next layer. Therefore, by leveraging the semi-tree-network structure, we can train the underlying Perceptron model on each node separately and can apply techniques like L2-Regularization and Normalization. Hence, different from a deep learning neural network like MLP, gradient backpropagation is not necessary for tweaking parameters throughout the network.

A traditional Decision Tree is an algorithm that uses a tree-like model of decisions and their possible consequences like event outcomes, resource costs, and utility. Decision Tree is commonly used in operation research and operation management.

K-Nearest Neighbors is a popular classification algorithm. To classify a data point x, KNN compares the distance between every data point to x and finds the k nearest data points. The dominating class among the k data points is the predicted class for x.

Background:
The total rows of the dataset is about 7k, leaving enough freedom for us to separate it into test and training datasets. Meanwhile, we have confidence that there are enough columns for us to model the relationship between wine quality and its chemical metrics.

Significance:
The dataset we are working on is about wine quality with about 7k data. It contains different chemical aspects of the alcohol as columns that we will use as features to predict the quality. We consider it a good dataset given that it has already been processed by tens of users, but is not as well known as the “Iris flower”.