CIS545 Project Proposal

Group Members: Yihang Zhou, Yinzhou Wang, Hainiu Xu

Data Source

We will be using the United States Drought data from Kaggle (https://www.kaggle.com/cdminix/us-drought-meteorological-data)

Motivation

Droughts in the US are raising concerns in recent years. We want to know what features affect the drought conditions the most (is it possible to only use meteorological data plus soil data) and try to predict the future. It has the potential to further apply our model to other areas in the world and examine its robustness.

Project Plan

1. Area of Study
   In this project, we are looking forward to building a promising classification model that can identify the severity of Drought using various related features. Specifically, this will be a multi-class classification problem with 6 classes.

   Data Wrangling and Exploratory Data Analysis: Since the US Droughts data contain features of two major genres-- weather and soil, which of those are stored in separate csv files, we will start with data wrangling works such as joining the two datasets and conduct rudimentary data cleaning. The data wrangling procedures will be followed by a thorough exploratory data analysis where skills learned in the classes such as correlation analysis and data visualization (static and animated) will be utilized.

   Modeling: The main task of this project is data modeling. We will be exploring models in both linear classification families as well as models proposed based on the notion of neural networks.

2. Objective
   We are going to conduct classification models to predict the different levels of drought in the US based only on meteorological data and potentially generalize our study for other countries in the future. The US drought monitor's drought scale is composed of five different levels, from abnormally drought to exceptionally drought namely D0, D1, D2, D3, and D4. In our datasets, we have an extra 'None' level indicating no drought at the area. All observations are collected at a specific location in a county in a specific time, and it contains other 18 features including wind speed, temperature, humidity, air pressure, precipitation, dew point, etc.

3. Range of Models
   In terms of modeling the US Droughts data, the linear classification models learned in this course will be used to construct a baseline classifier and the ensemble and neural network models will be built to beat the baseline.

   For linear and tree-based ensemble models, we are planning to employ feature selection, dimension reduction, and hyperparameter optimization techniques for further performance boost.

   For neural network models, we will attempt the mainstream models such as MLP, RNNs, and CNN.

Potential Challenges and Obstacles

1. This project is harder in modelling rather than EDA, possible neural networks like RNN and LSTM can be complex enough and beyond the scope of this course.
2. Animated visualization in our plan is relatively hard, but is really cool as well.
3. The original data is massive (~19 million entries). We will need to figure out a way to reasonably reduce the amount of data while roughly keep the distributions of the features.

TA requirement

Yes, we would like to get assistance from our TA staff.
Desired TA: Varun Jana