IAI 5101: Foundations of ML for Engineers & Scientists

Winter 2023

Assignment 2

Submission Deadline: 24th March 2022 on Brightspace. Submit your Python Notebook (1-2 in a group allowed).

Part A: Supervised Learning (80%)

The compressive strength of concrete (an important material in civil engineering) is the ability to resist failure in form of cracks. The 28-day compressive strength is one of the most widely used metrics to characterize concrete's performance in engineering applications (e.g., holding structural loads). Predicting accurately the concrete strength is important in many construction projects to avoid catastrophic failure of civil infrastructures. Conversely, concrete with overdesigned strength can lead to higher material cost and other environmental issues (e.g., CO_2 emission). Recent development of ML techniques provide an alternative solution to address the prediction problem. While ML can be used to infer the complex and non-linear relationship between concrete mixture proportions and strength, large datasets are needed to train such models. You are hereby provided the *concrete.csv* dataset collected. The available realistic industrial concrete strength data is rather limited, i.e., small dataset problem. Below is a description of the attributes.

- *Cement : measured in kg in a m3 mixture*
- *Blast : measured in kg in a m3 mixture*
- *Fly ash : measured in kg in a m3 mixture*
- *Water : measured in kg in a m3 mixture*
- *Superplasticizer : measured in kg in a m3 mixture*
- *Coarse Aggregate : measured in kg in a m3 mixture*
- *Fine Aggregate : measured in kg in a m3 mixture*
- *Age : day (1~365)*
- *Strength: concrete compressive strength measured in MPa (**Target**)*

I. **EDA (10 marks):**

- Univariate Analysis:
  - Build a histogram to show the distribution and central values of all the variables
  - Use a boxplot to determine if there are outliers in the variables
- Multivariate Analysis:
  - Use a pair plot to determine the relationship and degree of relation between independent variables and between independent variables
  - Use a heatmap to check for correlation between predictor variables

II. **Feature Engineering (10 marks):**

- Ensure data is in the correct format for downstream processes
  - Check for duplicates & missing values and drop, if present
  - Remove possible outliers in the dataset
  - Check for zeros in the dataset and impute with the mean
  - Check for class imbalance and handle, if necessary
  - Scale the data using a standard scaler

III. **Model Development I (30 marks):**

Ensemble Method:

- Split dataset into train (70%) and test (30%) and build predictive models to determine the strength of concrete using the following techniques: K-Nearest Neighbor Regressor, Random Forest, and XGBoost
- Use a voting regressor to predict the values for the ensemble of heterogeneous models above (K-Nearest Neighbor Regressor, Random Forest & XGBoost).

IV. **Model Development II (20 marks):**

Deep Learning:

- Split dataset into train (70%) and test (30%) and train a deep neural network using Keras.
- Try to improve the model by changing the activation function or dropout rate. What effects does any of these have on the result?

V. **Model Comparison, Evaluation (10 marks):**

- Compare the results of the ensemble models with the deep neural network model in terms of the following criteria: RSME, MAE, MSE, and R2 accuracy.
- Identify the model that performed best and worst according to each criterion.

Part B: Unsupervised Learning (20%)

Clustering can be used to identify groups that share common patterns, e.g., geographic areas that share similar weather patterns. You are hereby provided with the *housing* dataset for a state. Cluster the dataset using the median income, longitude and latitude to create economic segments in different regions.

- Perform *k-means* clustering on the selected attributes, specifying k = 6 clusters and plot.
- Apply the elbow method to determine the best k and plot.
- Evaluate the quality of the clusters using the Silhouette Coefficient method.