CS771

The Neural Nuts

May 2023

Problem 1

Solution : We have used **Linear Regression** to get the best performing linear model(in terms of MAE on training data). It gives the best fit line that describes the linear relationship between the dependent variable and the independent variables.

Steps taken in Linear Regression

Data preparation: This involves collecting and cleaning the data, identifying the dependent and independent variables, and splitting the data into training and testing sets.

Model selection: Here we choose the appropriate type of linear regression (simple linear regression or multiple linear regression) and selecting the variables to include in the model.

Model fitting: Estimate the values of the intercept and slope coefficients using a method such as least squares regression. The coefficients are chosen to minimize the sum of the squared errors between the predicted values of the dependent variable and the actual values.

Model evaluation: The performance of the model is evaluated by computing metrics such as mean absolute error (MAE), mean squared error (MSE), or R-squared, and the evaluation is carried out on the testing data to confirm the model's ability to generalize to new, unseen data.

Model refinement: Adjustments to the model are made such as adding or removing variables or applying regularization techniques to improve its performance.

We iterate through these steps to get the best-performing linear model.

The Mean Absolute Error (MAE) values on training set are :

Mean absolute error NO2	Mean absolute error OZONE
5.754484278500234	6.473547759023771

Problem 2

Solution : We have used **K-Nearest Neighbouring** algorithm. The idea behind KNN is to predict the class of a new data point based on the class of its closest neighbours in the feature space. The algorithm determines the k closest neighbors to the new data point, where k is a hyper-parameter that needs to be set by the user. KNN is a non-parametric algorithm, which means it does not make any assumptions about the underlying data distribution.

Feature Selection

We have used the four voltage parameters used in Q1, in addition we have also used humidity and temperature as features. We have modified the time feature, removing the date as the dates are not changing much and hence doesn't have a large effect on the result and introducing a new feature as:-

$$\theta = \frac{2 * \pi * hours}{24} + \frac{2 * \pi * minutes}{24 * 60}$$

The θ is defined to take into account the periodicity of the time-cycle. Here we have defined the cosine and sine of θ as our two features which is analogous to being x and y coordinates in a plane.

Hyper-parameter tuning

Here the only hyper-parameter we have is of K, which we checked for different values and the optimum result came for K=6.

Training-Strategy

We divided the training set into training and testing data with a split ratio of 70-30%.

Loss-Function

There is no loss function for KNN, at the time of training only the data is memorised at the feature space is created for the data.