Introduction to AI Coursework

Runze Yuan 2217498

May 19, 2023

1 Introduction

1.1 Question

For a Combined Cycle Power Plant, use its hourly average ambient variables to predict the net hourly electrical energy output.

To be specific, the task is to use four input value to predict one output value.

1.2 Which kind of algorithms to use and why

Use **Regression** for the task.

Reasons:

- The aim of regression algorithms is to capture the relationship between inputs and outputs, which is what the task asks for (to predict energy output with four ambient value).
- Regression algorithms are capable of predicting future or unseen data.

2 Methods

2.1 Algorithms

In this report I will show results with **K Neighbors Regressor** and **Decision Tree Regressor**.

- KNR: Find K nearest neighbors in the feature space for the input vector and use the output value of the neighbors to calculate the predicted output.
- DTR: DTR builds a decision tree by recursively splitting points into groups based on the result of separation.

2.2 Metrics

The Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) are selected as performance metrics for the algorithm. These are common metrics for regression algorithms.

- MSE and MAE are metrics that reflects the difference between the predicted result and ground-truth, the lower the better.
- R^2 reflects the goodness of fitting of the algorithm, and a value close to 1 means a good fit.

2.3 Baseline

For the baseline model, use the dummy model in the scikit learn.

Set all hyperparameters of the dummy to the default value, which means set "parameters" to mean, and both "constants" "quantile" to None.

With these parameters, the dummy model is not an actual regressor and would always output the mean value of the given training output data.

2.4 Hyperparameters

Hyperparameters are parameters that are set prior for models and not changed in the learning process. These parameters control the characteristic of the model and allow adjustments for the user by tuning the hyperparameters.

- **KNR**: In this report I will try to find the optimal **n_neighbours**, **weights**, and **p**.
  - n_neighbors: how many nearest neighbors the algorithm use to predict the output value.
  - weights: changes the weights used in the output generating.
    - uniform: all nearest neighbors have the same weight.
    - distance: use distance as weights for the neighbors, distant neighbors have lower weights.
  - p: changes the type of distance used in the algorithm
    - p=1: apply Manhattan distance for distance calculation.
    - p=2: apply Euclidean distance for distance calculation.
- **DTR**: In this report I will try to find the optimal **max_depth**, **min_samples_leaf**, and **splitter**
  - max_depth: the maximum depth of the tree.
  - min_samples_leaf: the minimum sample amount for a node to be a leaf node.
  - splitter: changes the strategy of splitting samples into different groups in nodes.
    - best: evaluate all possible splits and apply the best one which have the best performance.
    - random: apply the best split among a random generated strategy for randomly selected subsets of the features.

2.5

3 Result and Analysis