Practical/Tutorial Week 3 - Supervised Learning (k-NN and Decision Trees)

DATA7703 - Machine Learning for Data Scientists

Aims:

- To apply k-NN and decision trees as examples of supervised learning models for classification and regression problems.
- To get some insight into the trained models, including the influence of a hyperparameter in an ML model.
- To produce some assessable work for this subject.

Note that for this prac, you are not expected to implement the machine learning algorithms/models from first principles. You may use existing machine learning libraries/toolkits and refer to the documentation for those libraries. You can also choose to use tools like ChatGPT to help you. However, if you want to learn something, you will need to be able to analyse the code produced by ChatGPT to verify that it works as you intended and to be able to modify it if necessary to complete the prac questions. Your understanding of this code and the work that you do will be evaluated in the prac demos.

On blackboard you will find two datasets for use in this prac: w3classif.csv and w3regr.csv. Download these datasets.

(Q1) Make scatterplots of each dataset so you can see what they look like.

(Q2) Randomly shuffle the datasets (i.e. the order of the rows) and split them each into 70% (for training) and 30% (for testing).

(Q3) **(a)** Build a k-NN classifier with $k = 3$ for dataset w3classif.csv and find the training and test loss (i.e. misclassification rate).

(b) Plot the decision regions for your classifier together with the training and/or test data points.

(c) Experiment with different k values and see how it affects the loss values and the decision regions.

(Q4) **(a)** Build a k-NN regression model with $k = 3$ for dataset w3regr.csv and find the training and test loss (i.e. sum of squared error).

(b) Plot the training and/or test data together with the predicted “function” of the model.

(c) Experiment with different k values and see how it affects the loss values and the predicted function.

(Q5) **(a)** Build a decision tree classifier for dataset w3classif.csv and find the training and test loss (i.e. misclassification rate).

(b) Plot the decision regions for your classifier together with the training and/or test data points.

(c) Experiment with different maximum depth values and see how it affects the loss values and the decision regions.

(Q6) **(a)** Build a decision tree regression model for dataset w3regr.csv and find the training and test loss (i.e. sum of squared error).

(b) Plot the training and/or test data together with the predicted "function" of the model.

(c) Experiment with different maximum depth values and see how it affects the loss values and the predicted function.