Artificial Intelligence and Expert Systems

Assignment II

Faculty of mechanical engineering

Instructors:

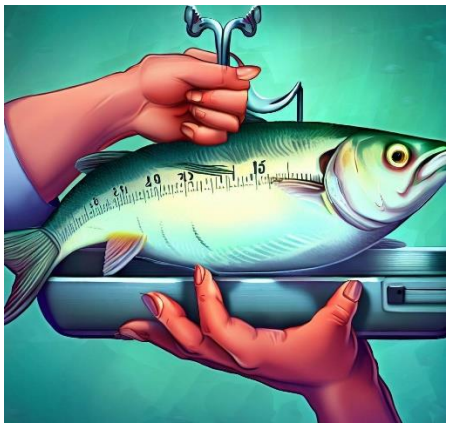
Dr. Esmaeil Najafi

Reza Behbahani Nejad

Due date: 1401/02/10

Problem I)

The Dataset_I contains information on 7 commonly sold fish species in the market. Using this dataset, it is possible to build a predictive model that utilizes machine-readable data to estimate the weight of fish.



Please consider the dataset I and perform the following steps:

a) Determine the number of fish species present in the dataset and analyze their distribution across each class. (Plot a bar chart too!)

b) Find the relationship between the following features: (weight and width) ,(weight and diagonal length), (cross length and vertical length). Use scatter diagrams to visualize the data and provide a detailed explanation of your findings.

c) Develop a simple linear regression model for (cross length vs vertical length). Fit the model and assess its performance using the mean squared error (MSE) and mean absolute error (MAE) metrics. Compare and plot the results.

d) Identify the features that have the most significant impact on fish weight and choose three of these features for further evaluation. Develop a multiple linear regression model to assess the influence of these features on fish weight and evaluate the model using the MSE and MAE metrics. Plot the results to provide a visual representation of your findings.

e) Develop a polynomial regression model for (weight vs width), as well as (weight vs height).

Problem II)

In this exercise, the bias-variance tradeoff in machine learning will be examined. When training a regression model, it is necessary to determine the general formula beforehand. If a simple mathematical model (High bias-low variance) is chosen, it might not be able to capture the underlying relationship present in the dataset. On the other hand, choosing a complicated model (High variance-low bias) might yield good results on training data, but it runs the risk of overfitting and performing poorly on test data. The Dataset II consists of temperature sensor measurements. Based on this data set,

a) Fit a regression model using the following mathematic formulas:

$$f_1(x) = w_1 + w_2 x + w_3 x^2$$

$$f_2(x) = w_1 + w_2 x + w_3 x^2 + \cdots + w_{10} x^9$$

$$f_3(x) = w_1 + w_2 x + \sin(x) + \cos(x)$$

b) Evaluate the models using MSE and MAE. Tabulate the results. Look at which models performed better on this dataset and give your conclusion on the results.

Problem III)

Consider the Dataset_III and do the following steps:

a) Read the text file, how many classes exist in the dataset? Which class has the most amount of data?

b) Plot the histogram of the numerical features. Explain what you understand

c) Use the KNN method in order to fit the appropriate model. For validation, use the F1 score metric.

d) Find the best k.

Dataset Hint:

The Dataset_III contains information on various drinks that generally contain caffeine. It includes instances of drinks that may not fit the typical definition of a drink, such as ground

coffee or tea leaves that would produce a certain volume (in milliliters) if prepared according to the provider's instructions. The dataset does not include calorie information since sugar levels can be controlled. The attributes included in the dataset are the name of the drink, volume quantity, calorie quantity, caffeine quantity (in milligrams), and type of drink (coffee, energy drinks, energy shots, soft drinks, tea, or water).