Artificial Intelligence and Machine Learning Fundamentals

Activity 7: Preparing Credit Data for Classification

In this section, we will discuss how to prepare data for a classifier. We will be using **german.data** from https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/ as an example and will prepare the data for training and testing a classifier. Make sure that all of your labels are numeric, and that the values are prepared for classification. Use 80% of the data points as training data:

1. Save **german.data** from https://archive.ics.uci.edu/ml/machine-learningdatabases/statlog/german/ and open it in a text editor such as Sublime Text or Atom. Add the header row to it.
2. Import the data file using pandas and replace the NA values with an outlier value.
3. Perform label encoding. Transform all of the labels in the data frame into integers.
4. Separate features from labels. We can apply the same method as the one we saw in the theory section.
5. Perform scaling of the training and testing data together. Use **MinMaxScaler** from Scikit's Preprocessing library.
6. The final step is cross-validation. Shuffle our data and use 80% of all data for training and 20% for testing.