MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

MACHINE LEARNING.

Lab 02: Pandas and ScikitLearn

BACKGROUND.

The goal of this lab exercise is to work with Pandas to do some preliminary data analysis that is typically required before implementing any machine learning algorithms and to understand the typical input and output of the Machine Learning algorithms implemented in the ScikitLearn toolbox.

Task 1.

In this task you will perform a basic analysis of a bike sharing dataset available at https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. Load the CSV file "day.csv" using Pandas.

A. Compare the average number of casual rentals and registered rentals depending on if it is a holiday or not. What can you observe?
B. The temperature values are already normalised for classification. What is the minimum and maximum temperature in the data set in Celsius?
C. Usually there are more registered than casual renters. On which days in the data set is this not the case?
D. Plot the temperatures against the number of casual and registered rentals. What can you observe?

Task 2.

In this task you will analyse the Titanic passenger dataset, which you can download from Canvas. Load the CSV file "titanic.csv" using Pandas.

A. How many passengers were on the titanic, and what percentage survived?
B. Determine the survival rate for male and female passengers. What can you observe?
C. What is the average fare paid by survivors compared to non-survivors?

Task 3.

In this task you will generate a random dataset with n clusters each containing k data points. The dataset should contain 2 features and a label for each data point. It should be compatible with the classifier input format for Scikit-Learn.

A. Generate the cluster mean, i.e. n vectors of size 2 uniformly distributed in the unit box. (Hint: use np.random.rand)
B. Generate the data points for each cluster by adding Normal distributed noise to the cluster means. (Hint: use np.random.randn)
C. Create the data and target vector formatted as required by Scikit-Learn classifiers.
D. Visualise the data in a scatter plot.

Task 4.

In this task you will train a Support Vector Machine on the data generated in task 3 and visualise the resulting decision function.

A. Train a Support Vector Machine classifier with the data generated in Task 1.
B. Create a mesh grid of all 2d coordinates in the unit box using the following call:
x,y = np.meshgrid(np.arange(0, 1, 0.1), np.arange(0, 1, 0.1))
Then extract a 100x2 matrix of feature vectors covering the whole unit box from x,y.
C. Run the trained Support Vector Machine classifier to predict the label for each point in the unit box.
D. Reshape the predictions into a 10x10 matrix and visualise the labels for the areas in the unit box as an image. (Hint: use matplotlib.pyplot.imshow)