Machine Learning

Lab 03: Scikit learn

The goal of this lab exercise is to work with Scikit learn and to understand the typical input and output of for the Machine Learning algorithms implemented in this toolbox.

Task 1.

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. Add m outliers uniformly distributed in the unit box and with random labels to the dataset. (Hint: use np.random.rand and np.random.randint)
E. Visualise the data in a scatter plot.

Task 2.

In this task you will train a Support Vector Machine on the data generated in task 1 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)