# Tutorial_05_Tasks

March 18, 2019

QBUS6850 - Machine Learning for Business

## 1 Tutorial 5

### 1.1 Task 1 - Spam Email Classification

1. Download the spambase.txt data
2. Load the data and split into training and test sets (75/25)
3. Build a KNN classifier. Use cross validation on the training set to estimate the best $k$
4. Plot the confusion matrix using your final trained model and the test data
5. Predict whether email 647 is spam or not and print the result

**Notes:** - You can find the detailed data description at here http://sci2s.ugr.es/keel/dataset.php?cod=109

### 1.2 Task 2 - Clustering Hand Written Digits

1. Cluster the digits dataset from sklearn using k-means
2. Cluster the digits using another method from the sklearn clustering user guide http://scikit-learn.org/stable/modules/clustering.html.
3. Compare clustering accuracy of k-means and your chosen method using mutual information score and homgeneity score.

**Notes:** - Some methods can be computationally prohibitive but there are solutuions. For example If you choose spectral clustering make sure to set the affinity = 'nearest neighbours'. - Some methods listed are semi-supervised clustering methods. They require you to provide a small number of training or exemplar samples.