Technische Universität München WS 2016 Fakultät für Informatik	Prof. Dr. N. Navab, Ph.D. November 23rd, 2016

Exercise Sheet Praktikum Machine Learning

Random Forests

Data: You are provided with three data files following the *.csv* format: TwistData.csv, SpiralData.csv and TuberculosisData.csv. The first two are synthetically generated and the last one consists of first twenty principal components of deeply learnt features[1] extracted from a publicly available tuberculosis Chest X Ray dataset[2]. The last data column in all the files refers to the class information (particularly, in TubercuolosisData *1* - Normal and *2* - Tuberculosis.)

Task 1:

For TwistData and SprialData: Learn random forest classfiers. Spilt the data randomly into two folds. Use the folds interchangeably for training and testing.

a. Refer to: : *http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html*
b. Train RF classifiers varying the number of trees (10,15,...,50) and observe the decision boundaries plotting curves like shown in: *http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py*
c. For fixed number of trees (say, 10) vary the depth of the classifiers from 2,3,..8. Plot similar curves as 1(b).
d. Comment on the classifier behavior for the above cases.

Task 2:

For Tuberculosis data: Learn random forest classifiers. Perform $k = 5$ folded cross-validation, i.e., split data into 5 folds and use 4 folds for training and 1 fold for testing.

a. Classfier 1: Train random forest classifier with 20 trees and max depth of 4.
b. Classifier 2: Train linear SVM classifier with RBF kernel ($\sigma = 1$). (Use codes from previous assignments).
c. Classifier 3: Train logistic regression classifier. (Use codes from previous assignments).
d. Compare the performance of classifiers 1, 2 and 3 by calculating the accuracy, sensitivity and specificity using a one *vs.* all binary confusion matrix.

[1]. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[2].Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X. and Thoma, G., 2014. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, *4*(6), pp.475-477. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256233/