Cairo University
Faculty of Engineering
Computer Engineering Department

CMPN450
Fall 2018

Pattern Recognition and Neural Networks.

Lab 5 – Boosting Algorithms

Objective:

In this lab, we will implement the AdaBoost algorithm as an ensemble learning technique which aims to combine a number of weak classifiers to yield a strong classifier at the end.
The idea of this lab is to identify whether a tumor with given characteristics is malignant or benign. This is a two-class classification problem.

Dataset and Features:

You will be working on the dataset hastie et al, with 10 features representing tumor's area, perimeter, radius, compactness, concavity, texture, symmetry, greyscale level, fractal dimension and coastline approximation. There is one output variable which is diagnosis. It takes one of two values (+1) for malignant and (-1) for benign.

(Q1) Why it is sometimes better to have the two class values (+1) and (-1) instead of (+1) and (0)?
Think about the voting scheme at the end of the boosting algorithm. How can the class values affect this scheme?

Requirement:

You are required to fill the function adaboost_clf(Y_train, X_train, Y_test, X_test, M, clf). This function takes as parameters:

- **Y_train:** The target values for the training set.
- **X_train:** The input features for the training set.
- **Y_test:** The target values for the test set.
- **Y_train:** The input features for the training set.
- **M:** The number of iterations of the AdaBoost Algorithm.
- **clf:** The classifier to be used. (In our case, we are using a decision tree stump as a base classifier). You can use any other classifier.

This function should return two values:

- **The accuracy of the model on the training set.**
- **The accuracy of the model on the test set.**

(Q2) At the end of the python code, there is a snippet code plotting the number of iterations of the AdaBoost Algorithm versus the accuracy of the model. What do you expect will happen as the number of iterations increase?

(Q3) What do you think is more critical in breast cancer detection problem: false positives or false negatives?

(Q4) What are the hyperparameters used in this model? How can you tune them in AdaBoost Algortihm?