DSC3006 Assignment 2: Ensemble Learning

Goal

The goal of assignment 2 is to practice ensemble learning.

Task

Load the breast cancer wisconsin dataset from sklearn.

```
from sklearn.datasets import load_breast_cancer
data = load_breast_cancer()
```

1. Please report the average accuracy and standard deviation for 5-fold cross validation using i) decision tree; ii) SVM; iii) logistic regression.

2. Perform bagging on the data set for 10, 20, 50, 100, 200, 500 number of estimators with i) decision tree ii) SVM iii) logistic regression as the base classifier, conduct the 5-fold cross validation and calculate the average classification accuracy and standard deviation of those bagged classifiers. Plot the accuracy curve with the horizontal axis being the number of bagging estimators, and vertical axis being the average accuracy, please also plot the standard deviation of the accuracy as the error bars.

3. Perform Adaboost on the data set for 10, 20, 50, 100, 200, 500 number of estimators with i) decision tree ii) SVM iii) logistic regression as the base classifier, conduct the 5-fold cross validation and calculate the average classification accuracy and standard deviation of those bagged classifiers. Plot the accuracy curve with the horizontal axis being the number of bagging estimators, and vertical axis being the average accuracy, please also plot the standard deviation of the accuracy as the error bars.

4. Compare the results of 2 and 3 and answer the following questions

1) Will bagging or boosting effectively improve the classification performance for each classifier?
2) Will bagging or boosting effectively improve the classification stability in terms of the standard deviation of the 5-fold cross validation classification accuracy?
3) With the increasing of the number of bagging/boosting estimators, will the performance converge (convergence means the cross-validation performance will not increase with more estimators)?
4) Do you observe overfitting with the increase on the number of estimators for bagging/boosting?

Deliverable

dsc3006_assignment_2_yourname.ipynb