Assignment 8 (Sol.)
Introduction to Machine Learning
Prof. B. Ravindran

1. Which of the following is/are true about bagging?

    (a) Bagging reduces variance of the classifier

    (b) Bagging increases the variance of the classifier

    (c) Bagging can help make robust classifiers from unstable classifiers

    (d) Majority is one way of combining outputs from various classifiers which are being bagged

    Sol. a, c, d

    In bagging we combine the outputs of multiple classifiers trained on different samples of the training data. This helps in reducing overall variance. Due to the reduction in variance, normally unstable classifiers can be made robust with the help of bagging.

2. Which among the following prevents overfitting when we perform bagging?

    (a) The use of sampling with replacement as the sampling technique

    (b) The use of weak classifiers

    (c) The use of classification algorithms which are not prone to overfitting

    (d) The practice of validation performed on every classifier trained

    Sol. b

    The presence of over-training (which leads to overfitting) is not generally a problem with weak classifiers. For example, in decision stumps, i.e., decision trees with only one node (the root node), there is no real scope for overfitting. This helps the classifier which combines the outputs of weak classifiers in avoiding overfitting.

3. Consider an alternative way of learning a Random Forest where instead of randomly sampling the attributes at each node, we sample a subset of attributes for each tree and build the tree on these features. Would you prefer this method over the original or not, and why?

    (a) Yes, because it reduces the correlation between the resultant trees

    (b) Yes, because it reduces the time taken to build the trees due to the decrease in the attributes considered

    (c) No, because many of the trees will be bad classifiers due to the absence of critical features considered in the construction of some of the trees

    Sol. c

    The availability of all attributes (at possibly differing levels) allows the original random forest approach to have relatively good classifiers from which to construct the combined classifier. In the proposed approach, many of the constituent classifiers will exhibit very poor performance affecting the performance of the random forest classifier.

4. In case of limited training data, which technique, bagging or stacking, would be preferred, and why?

(a) Bagging, because we can combine as many classifier as we want by training each on a different sample of the training data

(b) Bagging, because we use the same classification algorithms on all samples of the training data

(c) Stacking, because each classifier is trained on all of the available data

(d) Stacking, because we can use different classification algorithms on the training data

Sol. c
When data is at a premium, we would ideally prefer to train all models on all of the available training data.

5. Is AdaBoost sensitive to outliers?

   (a) Yes

   (b) No

Sol. a
See solution to question 7.

6. Considering the AdaBoost algorithm, which among the following statements is true?

   (a) In each stage, we try to train a classifier which makes accurate predictions on any subset of the data points where the subset size is at least half the size of the data set

   (b) In each stage, we try to train a classifier which makes accurate predictions on a subset of the data points where the subset contains more of the data points which were miscalssified in earlier stages

   (c) The weight assigned to an individual classifier depends upon the number of data points correctly classified by the classifier

   (d) The weight assigned to an individual classifier depends upon the weighted sum error of misclassified points for that classifier

Sol. b, d
The classifier chosen at each stage is the one that minimises the weighted error at that stage. The weight of a point is high if it has been misclassified more number of times in the previous iterations. Thus, maximum error minimisation is performed by trying to correctly predict the points which were misclassified in earlier iterations. Also, weights are assigned to the classifiers depending upon their accuracy which again depends upon the weighted error (for that classifier).

7. In AdaBoost, we re-weight points giving points misclassified in previous iterations more weight. Suppose we introduced a limit or cap on the weight that any point can take (for example, say we introduce a restriction that prevents any point's weight from exceeding a value of 10). Which among the following would be an effect of such a modification?

   (a) We may observe the performance of the classifier reduce as the number of stages increase

   (b) It makes the final classifier robust to outliers

   (c) It may result in lower overall performance

Sol. b, c

Outliers tend to get misclassified. As the number of iterations increase, the weight corresponding to outlier points can become very large resulting in subsequent classifier models trying to classify the outlier points correctly. This generally has an adverse effect on the overall classifier. Restricting the weights is one way of mitigating this problem. However, this can also lower the performance of the classifier.

8. Which among the following are some of the differences between bagging and boosting?

   (a) In bagging we use the same classification algorithm for training on each sample of the data, whereas in boosting, we use different classification algorithms on the different training data samples

   (b) Bagging is easy to parallelise whereas boosting is inherently a sequential process

   (c) In bagging we typically use sampling with replacement whereas in boosting, we typically use weighted sampling techniques

   (d) In comparison with the performance of a base classifier on a particular data set, bagging will generally not increase the error whereas as boosting may lead to an increase in the error

   Sol. b, c, d

   With regards to the last option, boosting can result in an increase in error over a base classifier due to over-emphasis on existing noise data points in later iterations.