Data Application Lab

Decision Tree & Ensemble Methods Quiz Answer

1. **Assume everything else remains same, which of the following is the right statement about the predictions from decision tree in comparison with predictions from Random Forest?**

   A. Lower Variance, Lower Bias
   B. Lower Variance, Higher Bias
   C. Higher Variance, Higher Bias
   D. Lower Bias, Higher Variance

   Solution: D

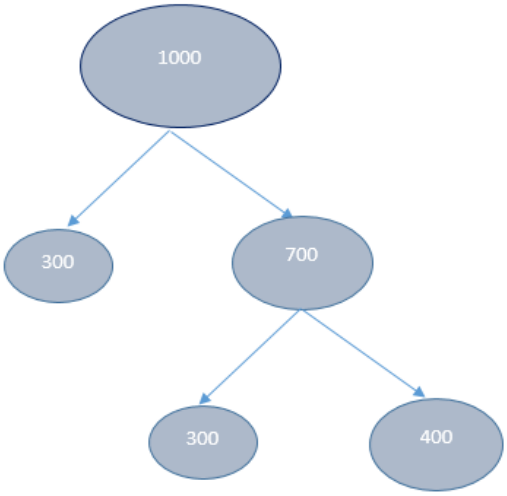
   The predicted values in Decision Trees have low Bias but high Variance when compared to Random Forests. This is because random forest attempts to reduce variance by bootstrap aggregation.

2. **Given 1000 observations, Minimum observation required to split a node equals to 200 and minimum leaf size equals to 300 then what could be the maximum depth of a decision tree?**

   A. 1
   B. 2
   C. 3
   D. 4
   E. 5

   Solution: B

The leaf nodes will be as follows for minimum observation to split is 200 and minimum leaf size is 300:



So only after 2 split, the tree is created. Therefore, depth is 2.

3. **Why do we prefer information gain over accuracy when splitting**?

   A. Decision Tree is prone to overfit and accuracy doesn't help to generalize
   B. Information gain is more stable as compared to accuracy
   C. Information gain chooses more impactful features closer to root
   D. All of these

   Solution: D

4. **In an election, N candidates are competing against each other and people are voting for either of the candidates. Voters don't communicate with each other while casting their votes.**

   Which of the following ensemble method works similar to above-discussed election procedure?

   A. Bagging
   B. Boosting
   C. A or B
   D. None of these

Solution: A

In bagged ensemble, the predictions of the individual models won't depend on each other. So, option A is correct.

5. **Why are ensemble methods superior to individual models?**

   Solution: They average out biases, reduce variance, and are less likely to overfit.

   There's a common line in machine learning which is: "ensemble and get 2%."

   This implies that you can build your models as usual and typically expect a small performance boost from ensemble.

6. **Explain bagging.**

   Solution: Bagging, or Bootstrap Aggregating, is an ensemble method in which the dataset is first divided into multiple subsets through resampling.

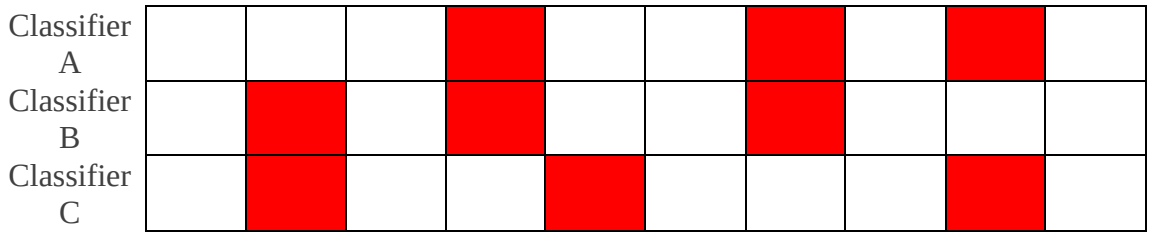
   Then, each subset is used to train a model, and the final predictions are made through voting or averaging the component models.

   Bagging is performed in parallel.

7. **Consider 3 different Classifiers that make predictions on 10 datacases. All 3 classifiers have made exactly 3 mistakes and the mistakes are independent. To generate an ensemble from 3 classifiers we use "majority vote" as the final predictions. Calculate the probability that the ensemble classifier performs worse than each of the individual classifiers (i.e. ensemble classifier makes 4 or more mistakes)**

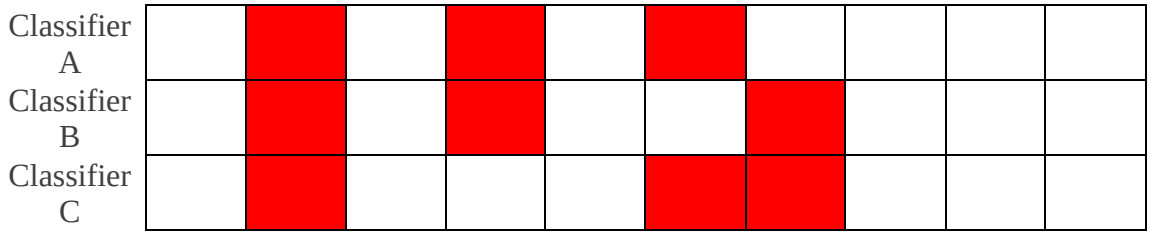
   According to the majority vote, a database will be predicted as mistake by the emsemble method where at least two classifiers have made mistakes in this database.

   After analysis, we will find that in the worst case, the ensemble will make at most 4 mistakes (just as the following figure, we use A, B, C to define the three classifiers). So the target for this problem is converted to let us find the possibility that the ensemble classifier makes 4 mistakes.



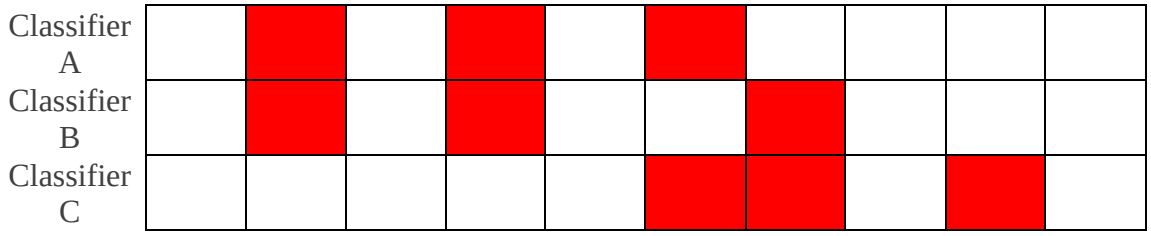
Now we consider several senarios:

1. One mistake is called by all the three classifiers, other three mistakes are made by either two classifiers.



$$C_{10}^{1}C_{9}^{2}C_{2}^{1}C_{7}^{1} = 5040$$

2. All the three mistakes are called only by two classifiers.



$$C_{3}^{1}[C_{10}^{1}C_{9}^{4}(C_{4}^{2}C_{2}^{1})] = 45360$$

$$\therefore \mathrm{P} = \frac{C_{10}^{1}C_{9}^{2}C_{2}^{1}C_{7}^{1} + C_{3}^{1}[C_{10}^{1}C_{9}^{4}(C_{4}^{2}C_{2}^{1})]}{(C_{10}^{3})^{3}} = \frac{7}{240} = 2.92\%$$

8. AdaBoost Clasiffier

When the prediction is right, we have $y_i h_t(x_i) = 1$;
When the prediction is wrong, we have $y_i h_t(x_i) = -1$;

$$z_t = \frac{1}{\sum_i D_{t+1}(i)}$$

$$\sum_{i:y_i \neq h_t(x_i)} D_{t+1}(i) = \sum_{i:y_i \neq h_t(x_i)} \frac{D_t(i)e^{-\alpha_t y_i h_t(x_i)}}{z_t} = \sum_{i:y_i \neq h_t(x_i)} \frac{D_t(i)e^{\alpha_t}}{z_t} = \frac{\sum_{i:y_i \neq h_t(x_i)} D_t(i)e^{\alpha_t}}{\sum_i D_{t+1}(i)}$$

$$= \frac{\sum_{i:y_i \neq h_t(x_i)} D_t(i)e^{\alpha_t}}{\sum_{i:y_i \neq h_t(x_i)} D_t(i)e^{\alpha_t} + \sum_{i:y_i = h_t(x_i)} D_t(i)e^{-\alpha_t}} = \frac{e^{\alpha_t}\varepsilon_t}{e^{\alpha_t}\varepsilon_t + (1-\varepsilon_t)e^{-\alpha_t}} = \frac{1}{1 + \frac{(1-\varepsilon_t)}{\varepsilon_t}e^{-2\alpha_t}}$$

$$= \frac{1}{1 + \frac{(1-\varepsilon_t)}{\varepsilon_t}e^{-\ln(\frac{1-\varepsilon_t}{\varepsilon_t})}} = \frac{1}{1 + \frac{(1-\varepsilon_t)}{\varepsilon_t}(\frac{1-\varepsilon_t}{\varepsilon_t})^{-1}} = \frac{1}{2}$$