1

Random forest for Regression or Classification.

- For $b = 1$ *to* B :
  - Draw a bootstrap sample Z^* of size N from the training data.
  - Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
    - Select m variables at random from the p variables.
    - Pick the best variable/split-point among the m.
    - Split the node into two daughter nodes.
- Output of ensemble of trees $\{T_b\}_1^B$.

Which of following statements is/are **True** regarding making a prediction at a new point x.

- **A RIGHT**.

$$Regression : \hat{f}_{rf}^{\,B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

- **B**.

$$Regression : \hat{f}_{rf}^{\,B}(x) = majority\ vote\ \{\widehat{C_b}(x)\}_1^B .$$

- **C RIGHT**.

$$Classification : Let\ \widehat{C_b}(x)\ be\ the\ class\ prediction\ of\ the\ the\ b^{th}\ random\ forest\ tree.$$
$$Then\ \widehat{C}_{rf}^{\,B}(x) = majority\ vote\ \{\widehat{C_b}(x)\}_1^B .$$

- **D**.

$$Classification : Let\ \widehat{C_b}(x)\ be\ the\ class\ prediction\ of\ the\ the\ b^{th}\ random\ forest\ tree.$$
$$Then\ \widehat{C}_{rf}^{\,B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x). .$$

2

The average of B identically independent distributed (i.i.d) random variables, each with variance σ^2 has variance $\frac{1}{B}\sigma^2$. If the variables are simply identically distributed, but not necessarily independent, with positive correlation ρ, then the variance is given by

- **A**. $\frac{1}{B}\sigma^2$
- **B**. $\rho\sigma^2$
- **C**. $\frac{1-\rho}{B}\sigma^2$

- **D RIGHT**.

$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$

We have given:

$X_i = \sigma^2 \ \ \forall i = 1, 2, .., B$ ... (equation-1)

$\mathbb{V}\text{ar}[\ \frac{1}{B} \times \sum_{i=1}^{B} x_i\ ] = \frac{1}{B} \times \sigma^2$

We kow that $\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \times \sigma_y}$

$\Rightarrow Cov(x,y) = \rho_{xy} \times \sigma_x \times \sigma_y$

We need to find $\mathbb{V}\text{ar}[\ \frac{1}{B} \times \sum_{i=1}^{B} x_i\ ]$ if they are identcally distributed but not independend.

$$\mathbb{V}\text{ar}[\ \frac{1}{B} \times \sum_{i=1}^{B} x_i\ ]$$

$$= \frac{1}{B^2} \times \mathbb{V}\text{ar}[\sum_{i=1}^{B} x_i\ ]$$

$$= \frac{1}{B^2} \times \sum_{i=1}^{B} \sum_{j=1}^{B} Cov(x_i, x_j)$$

$$= \frac{1}{B^2} \times \sum_{i=1}^{B} \sum_{j=1}^{B} \rho_{x_i x_j} \times \sigma_{x_i} \times \sigma_{x_j}$$

$$= \frac{1}{B^2} \times \sum_{i=1}^{B} \sum_{j=1}^{B} \rho_{x_i x_j} \times \sigma \times \sigma$$

$$= \frac{\sigma^2}{B^2} \times \sum_{i=1}^{B} \sum_{j=1}^{B} \rho_{x_i x_j}$$

$$= \frac{\sigma^2}{B^2} \times (\sum_{k=1}^{B} \rho_{x_k x_k} + \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} \rho_{x_i x_j})$$

$$= \frac{\sigma^2}{B^2} \times (\sum_{k=1}^{B} \frac{Cov(x_k, x_k)}{\sigma_{x_k} \times \sigma_{x_k}} + \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} \rho)$$

$$= \frac{\sigma^2}{B^2} \times (\sum_{k=1}^{B} \frac{\sigma^2}{\sigma \times \sigma} + \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} \rho)$$

$$= \frac{\sigma^2}{B^2} \times (\sum_{k=1}^{B} 1 + \sum_{i=1}^{B} \sum_{j=1, j \neq i}^{B} \rho)$$

$$= \frac{\sigma^2}{B^2} \times (B + (B^2 - B) \times \rho)$$

$$= \frac{\sigma^2}{B} + \rho \times \sigma^2 - \frac{\rho \times \sigma^2}{B}$$

$$= \rho \times \sigma^2 + \frac{1 - \rho}{B} \times \sigma^2$$

- MORE: http://personal.psu.edu/drh20/asymp/lectures/p51to58.pdf

3 K-means is

- **A**. a. A probabilistic algorithm to identify clusters present in data
- **B RIGHT**. b. A non-Probabilistic algorithm to identify clusters present in data
- **C**. None of the above
- **D**. Both a and b

4 Which of the following can act as possible termination conditions in K-Means?

- **A**. Reaching a maximum number of iterations
- **B**. Centroids do not change between successive iterations
- **C**. The squared distance of each data point from its centroid summed over all training data points falls below a threshold
    - I think it is same as variance!
- **D RIGHT**. All of the above
    - Stopping condition of K-means

5 Which of the following are true about Decision trees?

- **A WRONG**. Decision trees can be applied only for classification tasks
    - used for classification and regression
- **B RIGHT**. Decision tree is a non-parametric method
    - Decision trees are non-parametric because they make no distributional assumptions on the data. That's all there is to it. The presence of some numbers that specify certain aspects of the model do not make the model parametric.
- **C WRONG**. Decision tree can handle only categorical variables
    - Able to handle both numerical and categorical data.
- **D RIGHT**. Decision trees tend to overfit data and are high variance classifiers

---

- Decision Trees

- Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning
- Decision Tree - Overfitting

6 You are given 1500 training data points to train a decision tree. The minimum number of observations in each child node should be 300 after a split at the parent node. The minimum number of data points falling into a leaf node should be at least 400. Given these conditions what is the maximum possible depth of the decision tree?

- **A**. 1
- ~~**B RIGHT**~~. 2
- **C**. 3
- **D**. 4

The idea of making a decision tree with maximum height:

1. Have minimum number of childrer for each parent(in decision tree, minimum children is 2)
2. Grow the tree to only one side (say right side). It means children of each parent node except the last one should have one leaf node and the other intermediate node. The last intermediate node will have both children as leaves.

See the tree for this solution.

NB: I didn't understand the question properly. I confused with the statement "The minimum number of observations in each child node should be 300 after a split at the parent node."

If every leaf node have atleast 400 data points, each parent should have at leaat 800 datapoints. That means the above statement is reduntant.

Accepted Answers: 3

I don't know from where that answer came from.

7 Which of the following statements are True with regard to K-Nearest Neighbours?

- **A WRONG**. The decision boundary becomes smoother with decreasing value of K
  - As K increases, our decision surface gets smoother
- **B WRONG**. k-NN requires an explicit training step
  - kNN does not build a model of your data
- **C RIGHT**. The K-Nearest Neighbor algorithm considers the entire training data for each test point classification
  - KNN makes predictions using the training dataset directly.
- **D RIGHT**. Decreasing k increases variance
  - When we increase K, the training error will increase (increase bias), but the test error may decrease at the same time (decrease variance)

- k-NN and some questions about k values and decision boundary
- Would using too many neighbors in the k-nearest neighbor ...
- K-Nearest Neighbors for Machine Learning
- K-Nearest Neighbors and Bias–Variance Tradeoff

8. Download the Old Faithful data from the following link

http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat

Assume, you want to cluster the given data set into 2 clusters, using K-Means clustering algorithm. What will be the cluster centroids, on convergence?

- **A. RIGHT** C1: (2.09, 54.75) , C2: (4.29,80.28)
- **B.** C1: (1.88, 55), C2: (4.15, 88)
- **C.** C1: (2.18, 55), C2: (4.80, 81)
- **D.** None of the above

```
import requests
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

target_url = "http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat"
data_with_linenumbers = np.genfromtxt(target_url, skip_header=26)
data = data_with_linenumbers[:,1:]
print("Shape of the data: ", data.shape)

km = KMeans(n_clusters=2)
colors = km.fit_predict(data)
print("Centeroids: %s" % (km.cluster_centers_.round(3),))

# plot all data
plt.scatter(data[:,1],data[:,0], c=colors)
# plot two center points
plt.scatter(km.cluster_centers_[:,1], km.cluster_centers_[:,0])
plt.axis('equal')
plt.show()
```

9 Which of the following is true with respect to bagging?

- **A RIGHT**. Bagging involves sampling unbiasedly from the data for the purpose of ensemble learning.
    - Each sample have equal chance in model building.
- **B WRONG**. Bagging involves using weights on particular data points while sampling.
- **C RIGHT**. Bagging typically reduces variance in the trained model.
    - How can we explain the fact that “Bagging reduces the variance while retaining the bias” mathematically?
- **D WRONG**. Bagging typically reduces bias in the trained model.

10 Which of the following is true with respect to boosting?

- **A RIGHT**. Boosting can lead to overfitting the data
    - It is a bit subjective statement. boosting is **robust** to overfitting. But if more trees are used, it **can lead** to overfitting.
- **B**. Boosting only reduces variance in the trained model
    - Boosting not only reduces the variance, it also reduces the classification error.
- **C RIGHT**. Boosting tries to bias the overall model by weighting in the favor of good performers
- **D WRONG**. Boosting involves unbiased sampling of the data
    - boosting uses 'weight'. That can be considered as bias.