1. Describe how do you do to handle and analyze a new dataset? What are the steps of data cleaning?
2. What is overfitting? How do you prevent overfitting? Give a few examples of how it works and discuss.
3. What are the assumptions of a linear model? What will happen if we have correlated variables in a linear model?
4. What is bias/variance tradeoff? Give a few examples of how it works and discuss.
5. Describe what is a p-value? How would you explain it to people with no statistical knowledge?
6. What are the commonly used error measure for a classification problem? Explain its pros and cons
7. Given a data set and two models, how can you be sure which model is better? Explain what you will do to assess which one is better?
8. Describe the similarity and difference between ridge and lasso regression? Why regularization works?
9. How do you handle highly imbalanced data?
10. What are the commonly used model selection techniques? Pick one method and describe the procedures.
11. Explain what is cross validation and what is bootstrap?
12. How do you identify the most important variables of a model? List some techniques and pick one technique to explain the necessary steps.
13. How would you capture interaction between your features?
14. Why feature engineering is important in model building and describe a few methods used for feature engineering?
15. How to handle missing data?
16. What algorithms can be used for reducing the dimensionality of data? Pick one and describe how it is used.
17. Describe what a principal component analysis does. What are the key steps of performing PCA?
18. If the response variable has different shapes in different regions, what model can we use to accommodate this and why?
19. Explain the pros and cons of spline functions. List a few commonly used spline basis functions. How to choose the number of spline basis functions?
20. What is a generalized additive model (GAM), describe when to use GAM and describe what traits the data set should have such that it is appropriate for using GAM. List R/Python packages that implements GAM.
21. What is a deep neural network? What are the available software packages in R and Python? List some applications you know of and give some comments about them.
22. In the neural network, what’s the intuition behind the bias term in a neuron?
23. What is the role of activation functions in a neural network? Why is the activation function nonlinear?
24. Describe a few activation functions for neural network you have heard of or used and give some comments about them.
25. Explain what a regression tree is and its role in machine learning. Describe the steps of fitting a regression tree?
26. What is a random forest? Why random forest can help reduce variance? What is the pros and cons of random forest? Discuss the R and Python packages of random forest.
27. What is bagging? What are pros and cons of the bagging? Discuss the R and Python packages of bagging.
28. What is boosting? What are pros and cons of the boosting? Discuss the R and Python packages of boosting.
29. What is lasso regression? What are pros and cons of the lasso? Discuss the R and Python packages of lasso.
30. What are the hyper parameters of bagging? Discuss how to choose the hyper parameter.
31. What are the hyper parameters of random forest? Discuss how to choose the hyper parameter.
32. What are the hyper parameters of gradient boosting? Discuss how to choose the hyper parameter.
33. Explain what the K-nearest neighbor algorithm does and discuss its advantages and disadvantages.