Interview questions every Data Scientist should know!

Interview questions asked for Data Science role in 2020. I have put together around 100 questions which are the basic questions required to pursue a career in Data Science. These questions were asked by interviewers of top MNCs. But these are also the basic things to know for anyone working as a data scientist or aspiring to work as a data scientist. I have also added additional questions which are essential for working in real time and there is no harm in learning right?

1. What is Bias Variance tradeoff?
2. Why is Bias required? What role does bias play in ML?
3. How do you calculate weights and Bias?
4. Why to standardize a distribution and how?
5. Is it always required to normalize a distribution for our data analysis? If yes, how will I normalize the data without any loss of information? If not, how to proceed in such a case?
6. Calculate mean, median, mode, standard deviation, variance in a given dataset
7. What is the standard error?
8. What are the assumptions of any Regression algorithm for any problem statement?
9. What is homoscedasticity?
10. Difference between Numpy and Pandas
11. Difference between Numpy array and Python list
12. What is PCA and its working
13. Explain the use and working of SVM and why is it no longer used?
14. In what cases, regression is used and in what cases classification is used?
15. What do you infer from the output of Covariance, Correlation and Confidence interval?
16. Explain Rejection region
17. What do you mean by hypothesis?
18. Why that always null hypothesis needs to be rejected?
19. Explain Type I and Type II errors with example.
20. When is z-test used and interpret the results of a z-test in real time.
21. When is p-test used and interpret the results of a p-test in real time.
22. What is significance level
23. Calculation of statistical errors
24. How do you interpret selection bias? What are the different types of bias available?
25. Name sampling methods used
26. KNN – explain working
27. K means how do you determine the value of K chosen
28. Difference between Kmeans and KNN
29. Ensemble methods. What are they?
30. Time series – explain with POC
31. What is regularization? What are the methods and why is it used?
32. Explain precision and recall
33. What information will I get in Confusion matrix?
34. Regression steps
35. What is an Activation function?

36. Why is non linearity introduced in a network? How will activation function solve this problem?
37. Explain CNN
38. Why is backpropagation required?

39. What is Gradient descent?
40. What are the optimization techniques in CNN?
41. Explain vanishing gradient
42. Explain exploding gradient problem how to solve and when will it occur?
43. What is the difference between covariance, variance and correlation
44. What are the types of distribution?
45. What is Central Limit Theorem and what is Poisson distribution?
46. When do we use t score and what are its uses?
47. What is Naïve Bayes algorithm and how does it work?
48. Explain bagging and boosting. In which scenario are these used?
49. What is XG Boosting?
50. What do you mean by ARIMA?
51. Explain error metrics and regression metrics?
52. What is performance metrics?
53. What are the normalization techniques in ML?
54. ROC curve what information does this give?
55. What is multicollinearity?
56. Difference between numpy array and array? Why do we use numpy array when we can compute results using arrays and lists?
57. LSTM working
58. GAN working
59. Types of activation functions. When to use ReLU?
60. What are tensors?
61. Types of pooling
62. Calculate zero padding, output size, when input size is given
63. What is transfer learning?
64. Explain object detection and how it works in NN?
65. What is the difference between Numpy array and python list?
66. How does Numpy reduce the memory and increase the computational speed?
67. I have a pandas dataframe. I need to filter the dataframe based on some Boolean condition. How do you do it?
68. Explain loc and iloc
69. What is covariance in a data
70. How do you measure Covariance between two categorical variables? Which test do you use?
71. Tests for categorical variables
72. If in my data, all my missing values are replaced by mean, how will it affect my variance and why?
73. How will you measure the goodness of linear regression?
74. Evaluate a classification model
75. What are the other techniques to evaluate a classification model?
76. What is F1 score
77. Which models are used for Unsupervised learning? Explain one POC for the same
78. If I give you a clustering problem, which clustering will you be using and why?

79. In Kmeans, for numerical variables you can find Euclidian distance. But for categorical variables, how will you find distance? What is your approach in handling such cases?
80. How will you decide how many clusters you want in your model?
81. Specific method to find the value of K in ML

82. Difference between Transfer Learning model and building a model from scratch
83. What is Overfitting
84. Have you faced any overfitting problem in your model? How did you overcome it?
85. Millions of data points are there in training. Still the model is overfitted. How will you identify and overcome it?
86. What is dropout and why is it used?
87. I don't want to use dropout layer in my model. I have 5 hidden layers. There is overfitting. I can increase/decrease the number of layers. What will I do?
88. What is Random Forest?
89. Why do we use Random Forest?
90. How can you control over fitting in decision tree?
91. Difference between Gradient descent and Stochastic GD
92. Hyperparameters of RandomforestClassifier
93. What is pruning?
94. How do you select features? Steps
95. pipeline for every deep learning project in python
96. ML pipeline
97. Performance metrics explain
98. Eigen values and Eigen vectors where is it used?

These are the basic questions every data scientist should be able to answer. I have widely touched the topics in ML, DL.

Will be sharing the topics to learn step by step for those who are looking to make a career in data science.