1. Describe preprocessing steps that you carry out before using them to train a model and state under which condition they might be applied.
2. What models would you characterize as complex and which one as simple model.
3. What are the relative strength and weakness of choosing a more complex model over simple model?
4. What is dimension reduction? when and Why we want to use it.
5. What is the best method for filling missing value except using regression and Bayesian analysis?
6. Why we use binding method?
7. Why we use data transformation?
8. What is Normalization and its types.
9. Why we use decimal scaling.
10. What are the best method to perform data integration?
11. Why chi-square test is used.
12. How to choose best feature to perform regression and Bayesian analysis.
13. What is cube-aggregation and What is the best method.
14. How to choose class in decision tree?
15. Difference between regression and log-linear model.
16. What is lossy and lossless dimensional reduction.
17. What is wavelet transformation and Why we use it?
18. How to do automated trading.
19. How to build dataset from nifty and other source into one dataset.
20. Why do we use concat, append, joining and merging? And its types
21. Why we use pickle and Why preferred pandas pickle method over python inbuilt method.
22. Why pandas are faster the reading the file.
23. Why we use resampling method.
24. Difference between rolling statics and ohlc
25. What is mapping function and Why we map matplotlib mdates to dataset dates.
26. does the source of data matters?
27. How to represent 5 or more dimension to simple 2-dimension graph
28. How will you judge the quality of data visualization?
29. How will you decide when to use which data visualization method?
30. how to analysis and visualize religion.
31. when drop ignore feature from dataset why we use axis 1 not 0.
32. how to visualize 500 company data` to a single graph for a lay man.
33. Is rotation necessary for PCA? If we don’t rotate if what will happen.
34. when the Ridge regression favorable over Lasso regression.
35. how to convert complex dataset into simple one.
36. how the random forest algorithm different from Gradient boosting algorithm?
37. why we use Euclidean distance over Manhattan distance in K or KNN algorithm.
38. why don’t we use Euclidean distance in SVM algorithm?
39. explain machine learning algorithm to 5-year-old child
40. explain machine learning algorithm to me like a 5-year-old child
41. Consider the long list of ML algorithm, so when to use on which dataset
42. what will you do to gathering 50 GB of data from Internet for Analysis within a minute.
43. why we preferred SMO over SVM and why.
44. what is the use of point plot in Data visualization?
45. In automated car, how will you use genetic algorithm.
46. what if we transplant human brain with humanoid robot?
47. will people alive today have the opportunity to upload consciousness to a new robotic body
48. how will you describe animal working in the form of machine learning?
49. what will be the consequences if we store half memory of a genius body and other half of animal into a robot
50. how to cure disease using ML.?
51. how to analysis brain cognitive function using declarative and procedural knowledge modeling methodology.
52. What is Anthropomorphic action in robotics
53. In Neuroscience what is a strategic pilles of Human Brain and how we can develop robotic brain using that
54. why we use regression and why not category data use for it
55. inner working of regression
56. why we use y=mx+c in regression
57. what is the formula for verify the best fit line?
58. What is the difference between R-squared and Squared error?
59. Why the standard value of Squared error is 2 and what if we increase it by 4 or more
60. Why the standard value of Squared error is always even value
61. What change you do to increase the accuracy in regression analysis.
62. Can we use non-continuous dataset in regression analysis?
63. Which method k-nearest neighbor use to predict class of feature set.
64. Best value of K in k-nearest-neighbor
65. What if we use k=6 or -1 or 0?
66. What is the significance of -99999?
67. Which data structure used in training and testing set.
68. What is the difference between between accuracy and confidence?
69. How will you decide which feature to drop and which to labeled?
70. What is the significance of shuffle the data?
71. Inner working of k nearest neighbors
72. Which will be the best algorithm from scratch and SKlearn? why
73. What if we put n\_jobs=-1 in regression?
74. What is the fundamental difference between regression/KNN/SVM?
75. Why we use vectors and how vectors work in SVM or in SMO
76. What is the logic in dot product of ((vector w and vector u)- bias(b))>=1, dot product of ((vector w and vector u)- bias(b)) <=1, dot product of ((vector w and vector u)- bias(b)) =0?
77. What is the significance of Y sub I (- or +) class?
78. Why to minimize the magnitude of w and maximize the bias b
79. Why we use LaGrange’s equation
80. What is the convex optimization problem?
81. Why is called Support vector machine is an optimization problem
82. Equation of classification in Support vector machine
83. How to decide step size to solve the convex optimization problem in SVM.
84. What is the weakest link in SVM fundamentally? How to fix that link?
85. Which algorithm attempts to fix SVM fundamental problem
86. How to develop SVM from pure scratch
87. How to decide the decision boundary in SVM?
88. How to decide the best optimized value of w and bias b in SVM?
89. How to know which kernel use for which dataset
90. What are genetic algorithm and its equation
91. How to implement genetic algorithm with Machine Learning
92. How to solve Traveling salesman problem with Machine Learning?
93. How to spread virus with machine learning and genetic algorithm?
94. How to use genetic algorithm as a machine learning classifier
95. How to learn that newborn baby will be a boy or a girl in a 1-month pregnancy using machine learning.
96. Difference between univariate, bivariate and multivariate analysis
97. What is power analysis
98. Why is vectorization consideration as a powerful method for optimizing numerical code?
99. How can outlier value can be treated?
100. Explain about the box cox transformation in regression analysis
101. Fundamental difference between Bayesian and maximum like hood estimation
102. What is the curse of dimensionality?
103. What is multicolinearity and how to overcome it
104. How will you explain logistic regression to an economist, physio scientist and biologist?
105. Is naïve Bayes bad? under what circumstances
106. How will you decide value of K in KMeans algorithm?
107. How to do threading in SVM and where to power up it
108. How to visualize sp500 company in one graph
109. How will you classify waste using Artificial intelligence?
110. How to learn Animal natural language Artificial intelligence?
111. Why sklearn SVM has RBF kernel as default
112. How neural net preferred from SVM
113. How decision tree different from Table loop
114. What is Gini’s coefficient and information gain?
115. What is pruning? What are prepruning and postpruning? why choose over the other
116. What makes neural nets versatile for supervised as well as unsupervised learning task
117. Examine the steps in developing artificial neural nets for predicting stock price?
118. What kind of data would be required for a good price predictor system with ANN.?
119. How the Apriori algorithm works
120. What kind of pitfalls should one guard against while doing SNA?
121. What is random sampling? How does it helps
122. Why linear regression and KNN won’t work for spam filtering.
123. What is slack in kernel
124. As said by Dr. Stephen hawkin,Artficial intelligence will end mankind how?
125. Explain what regularization is and why it is useful.
126. How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression.
127. Explain what precision and recall are. How do they relate to the ROC curve?
128. How can you prove that one improvement you’ve brought to an algorithm is really an improvement over not doing anything?
129. What is root cause analysis?
130. Are you familiar with pricing optimization, price elasticity, inventory management, competitive intelligence? Give examples.
131. What is statistical power?
132. Explain what resampling methods are and why they are useful. Also explain their limitations.
133. Is it better to have too many false positives, or too many false negatives? Explain.
134. What is selection bias, why is it important and how can you avoid it?
135. Give an example of how you would use experimental design to answer a question about user behavior.
136. What is the difference between “long” and “wide” format data?
137. What method do you use to determine whether the statistics published in an article (e.g. newspaper) are either wrong or presented to support the author’s point of view, rather than correct, comprehensive factual information on a specific subject?
138. Explain Edward Tufte’s concept of “chart junk.”
139. How would you screen for outliers and what should you do if you find one?
140. How would you use either the extreme value theory, monte carlo simulations or mathematical statistics (or anything else) to correctly estimate the chance of a very rare event?
141. What is a recommendation engine? How does it work?
142. Explain what a false positive and a false negative are. Why is it important to differentiate these from each other?
143. What is polynomial kernel?what if we change the value of n=infinity,p=1000
144. What are RBF kernel and why are they complex.