1. You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)[Reference: https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/]
2. Is rotation necessary in PCA? If yes, Why? What will happen if you don't rotate the components?[Reference :https://www.quora.com/Is-rotation-necessary-in-PCA-If-yes-why-What-will-happen-if-you-don%E2%80%99t-rotate-the-components]
3. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?[Reference :https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/]
4. You are given a data set on cancer detection. You've build a classification model and achieved an accuracy of 96%. Why shouldn't you be happy with your model performance? What can you do about it?[Reference:https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/]
5. You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?[https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/]
6. You are assigned a new project which involves helping a food delivery company save more money. The problem is, company's delivery team aren't able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?
7. You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?
8. You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?
9. After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled .
10. Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?
11. You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?
12. You've got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?
13. You have built a multiple regression model. Your model R^2 isn't as good as you wanted. For improvement, your remove the intercept term, your model R^2 becomes 0.8 from 0.3. Is it possible? How?
14. After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he's true? Without losing any information, can you still build a better model?
15. You are given a data set consisting of variables having more than 30% missing values? Let's say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?
16. 'People who bought this, also bought…' recommendations seen on amazon is a result of which algorithm?
17. Which data visualisation libraries do you use? What are your thoughts on the best data visualisation tools?
18. How would you implement a recommendation system for our company's users?
19. How can we use your machine learning skills to generate revenue?
20. What are the last machine learning papers you've read?
21. Do you have research experience in machine learning?
22. What are your favorite use cases of machine learning models?
23. How would you approach the "Netflix Prize" competition?
24. Where do you usually source datasets?
25. How do you think Google is training data for self-driving cars?
26. What do you understand by Type I vs Type II error ?
27. You are working on a classification problem. For validation purposes, you've randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy i
28. State the universal approximation theorem? What is the technique used to prove that?
29. Given the universal approximation theorem, why can't a MLP still reach a arbitrarily small positive error?
30. What is the mathematical motivation of Deep Learning as opposed to standard Machine Learning techniques?
31. In standard Machine Learning vs. Deep Learning, how is the order of number of samples related to the order of regions that can be recognized in the function space?
32. What are the reasons for choosing a deep model as opposed to shallow model? (1. Number of regions O(2^k) vs O(k) where k is the number of training examples 2. # linear regions carved out in the function space depends exponentially on the depth.)
33. How Deep Learning tackles the curse of dimensionality?(Other sources(https://www.analyticsvidhya.com/blog/2017/01/must-know-questions-deep-learning)
34. How will you implement dropout during forward and backward pass?
35. What do you do if Neural network training loss/testing loss stays constant? (ask if there could be an error in your code, going deeper, going simpler…)
36. Why do RNNs have a tendency to suffer from exploding/vanishing gradient? How to prevent this? (Talk about LSTM cell which helps the gradient from vanishing, but make sure you know why it does so. Talk about gradient clipping, and discuss whether to clip the gradient element wise, or clip the norm of the gradient.)

37. Do you know GAN, VAE, and memory augmented neural network? Can you talk about it?
38. Does using full batch means that the convergence is always better given unlimited power? (Beautiful explanation by Alex Seewald: https://www.quora.com/Is-full-batch-gradient-descent-with-unlimited-computer-power-always-better-than-mini-batch-gradient-descent)
39. What is the problem with sigmoid during backpropagation? (Very small, between 0.25 and zero.)
40. Given a black box machine learning algorithm that you can't modify, how could you improve its error? (you can transform the input for example.)
41. How to find the best hyper parameters? (Random search, grid search, Bayesian search (and what it is?))
42. What is transfer learning?
43. Compare and contrast L1-loss vs. L2-loss and L1-regularization vs. L2-regularization.
44. Can you state Tom Mitchell's definition of learning and discuss T, P and E?
45. What can be different types of tasks encountered in Machine Learning?
46. What are supervised, unsupervised, semi-supervised, self-supervised, multi-instance learning, and reinforcement learning?
47. Loosely how can supervised learning be converted into unsupervised learning and vice-versa?
48. Consider linear regression. What are T, P and E?
49. Derive the normal equation for linear regression.
50. What do you mean by affine transformation? Discuss affine vs. linear transformation.
51. Discuss training error, test error, generalization error, overfitting, and underfitting.
52. Compare representational capacity vs. effective capacity of a model.
53. Discuss VC dimension.
54. What are nonparametric models? What is nonparametric learning?
55. What is an ideal model? What is Bayes error? What is/are the source(s) of Bayes error occur?
56. What is the no free lunch theorem in connection to Machine Learning?
57. What is regularization? Intuitively, what does regularization do during the optimization procedure? (expresses preferences to certain solutions, implicitly and explicitly)
58. What is weight decay? What is it added?
59. What is a hyperparameter? How do you choose which settings are going to be hyperparameters and which are going to be learnt? (either difficult to optimize or not appropriate to learn – learning model capacity by learning the degree of a polynomial or coefficient of the weight decay term always results in choosing the largest capacity until it overfits on the training set)
60. Why is a validation set necessary?
61. What are the different types of cross-validation? When do you use which one?
62. What are point estimation and function estimation in the context of Machine Learning? What is the relation between them?
63. What is the maximal likelihood of a parameter vector $theta$? Where does the log come from?
64. Prove that for linear regression MSE can be derived from maximal likelihood by proper assumptions.
65. Why is maximal likelihood the preferred estimator in ML? (consistency and efficiency)
66. Under what conditions do the maximal likelihood estimator guarantee consistency?
67. What is cross-entropy of loss? (trick question)
68. What is the difference between an optimization problem and a Machine Learning problem?
69. How can a learning problem be converted into an optimization problem?
70. What is empirical risk minimization? Why the term empirical? Why do we rarely use it in the context of deep learning?
71. Name some typical loss functions used for regression. Compare and contrast. (L2-loss, L1-loss, and Huber loss)
72. What is the 0-1 loss function? Why can't the 0-1 loss function or classification error be used as a loss function for optimizing a deep neural network? (Non-convex, gradient is either 0 or undefined.
73. 1.What's the difference between a generative and discriminative model?
74. When should you use classification over regression?
75. What evaluation approaches would you work to gauge the effectiveness of a machine learning model?
76. models are known to return high accuracy, but you are unfortunate. Where did you miss?
77. When is Ridge regression favorable over Lasso regression?
78. While working on a data set, how do you select important variables? Explain your methods.
79. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn't. How ?
80. Explain machine learning to me like a 5 year old.
81. Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?
82. Do you suggest that treating a categorical variable as continuous variable would result in a better predictive model?
83. When does regularization becomes necessary in Machine Learning?
84. What are parametric models? Give an example?
85. What are 3 data preprocessing techniques to handle outliers?
86. What are 3 ways of reducing dimensionality?
87. How much data should you allocate for your training, validation, and test sets?
88. If you split your data into train/test splits, is it still possible to overfit your model?
89. How can you choose a classifier based on training set size?
90. Explain Latent Dirichlet Allocation (LDA)
91. What are some key business metrics for (S-a-a-S startup | Retail bank | e-Commerce site)?
92. How can you help our marketing team be more efficient?
93. Differentiate between Data Science , Machine Learning and AI.((https://www.dezyre.com/article/100-data-science-interview-questions-and-answers-general-for-2018/184))
94. Python or R – Which one would you prefer for text analytics?
95. Which technique is used to predict categorical responses?
96. What is Interpolation and Extrapolation?

97. What is power analysis?
98. What is the difference between Supervised Learning and Unsupervised Learning?
99. Explain the use of Combinatorics in data science.
100. Why is vectorization considered a powerful method for optimizing numerical code?
101. What is the goal of A/B Testing?
102. What are various steps involved in an analytics project?
103. Can you use machine learning for time series analysis?
104. What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)?
105. What is multicollinearity and how you can overcome it?
106. What is the difference between squared error and absolute error?
107. Differentiate between wide and tall data formats?
108. How would you develop a model to identify plagiarism?
109. You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?
110. What do you understand by long and wide data formats?
111. What is the importance of having a selection bias?
112. What do you understand by Fuzzy merging ? Which language will you use to handle it?
113. How can you deal with different types of seasonality in time series modelling?
114. What makes a dataset gold standard?
115. Can you write the formula to calculate R-square?
116. Difference between Generative and Discriminative models.
117. How will you assess the statistical significance of an insight whether it is a real insight or just by chance?
118. How would you create a taxonomy to identify key customer trends in unstructured data?
119. What do you understand by feature vectors?
120. How do data management procedures like missing data handling make selection bias worse?
121. How's EM done?
122. How can you plot ROC curves for multiple classes. – There is something called as amacro-averaging of weights where PRE = (PRE1 + PRE2 + — + PREk)/K49.Text methods (latent, etc), he asked if I knew anything about these.
123. What is the difference between inductive machine learning and deductive machine learning?
124. How will you know which machine learning algorithm to choose for your classification problem?
125. What are Bayesian Networks (BN) ?
126. What is algorithm independent machine learning?
127. What is classifier in machine learning?
128. In what areas Pattern Recognition is used?
129. What is Genetic Programming?
130. What is Inductive Logic Programming in Machine Learning?
131. What is inductive machine learning?
132. What are the five popular algorithms of Machine Learning?
133. What are the different Algorithm techniques in Machine Learning?
134. List down various approaches for machine learning?
135. What are the different methods for Sequential Supervised Learning?
136. What is batch statistical learning?
137. What is PAC Learning?
138. What is sequence learning?
139. What are two techniques of Machine Learning ?
140. How to use labeled and unlabeled data?
141. What if you don't have any labeled data?
142. What if your data set is skewed (e.g. 99.99 % positive and 0.01% negative labels)?
143. How to make training faster?
144. How to make predictions faster?
145. Write the equation describing a dynamical system. Can you unfold it? Now, can you use this to describe a RNN? (include hidden, input, output, etc.)
146. What determines the size of an unfolded graph?
147. What are the advantages of an unfolded graph? (arbitrary sequence length, parameter sharing, and illustrate information flow during forward and backward pass)
148. What does the output of the hidden layer of a RNN at any arbitrary time t represent?
149. Are the output of hidden layers of RNNs lossless? If not, why?
150. RNNs are used for various tasks. From a RNNs point of view, what tasks are more demanding than others?
151. Discuss some examples of important design patterns of classical RNNs.
152. Write the equations for a classical RNN where hidden layer has recurrence. How would you define the loss in this case? What problems you might face while training it? (Discuss runtime)
153. What is backpropagation through time? (BPTT)
154. Consider a RNN that has only output to hidden layer recurrence. What are its advantages or disadvantages compared to a RNNhaving only hidden to hidden recurrence?
155. What is Teacher forcing? Compare and contrast with BPTT.
156. What is the disadvantage of using a strict teacher forcing technique? How to solve this?
157.

158. Explain the vanishing/exploding gradient phenomenon for recurrent neural networks. (use scalar and vector input scenarios)
159. Why don't we see the vanishing/exploding gradient phenomenon in feedforward networks? (weights are different in different layers – Random block intialization paper)
160. What is the key difference in architecture of LSTMs/GRUs compared to traditional RNNs? (Additive update instead of multiplicative)
161. What is the difference between LSTM and GRU?
162. Explain Gradient Clipping.
163. Adam and RMSProp adjust the size of gradients based on previously seen gradients. Do they inherently perform gradient clipping? If no, why?
164. Discuss RNNs in the context of Bayesian Machine Learning.
165. Can we do Batch Normalization in RNNs? If not, what is the alternative? (BNorm would need future data; Layer Norm)
166. What is an Autoencoder? What does it "auto-encode"?
167. What were Autoencoders traditionally used for? Why there has been a resurgence of Autoencoders for generative modeling?
168. What is recirculation?
169. What loss functions are used for Autoencoders?
170. What is a linear autoencoder? Can it be optimal (lowest training reconstruction error)? If yes, under what conditions?
171. What is the difference between Autoencoders and PCA (can also be used for reconstruction – https://stats.stackexchange.com/questions/229092/how-to-reverse-pca-and-reconstruct-original-variables-from-several-principal-com).
172. What is the impact of the size of the hidden layer in Autoencoders?
173. What is an undercomplete Autoencoder? Why is it typically used for?
174. What is a linear Autoencoder? Discuss it's equivalence with PCA. (only valid for undercomplete) Which one is better in reconstruction?
175. What problems might a nonlinear undercomplete Autoencoder face?
176. What are overcomplete Autoencoders? What problems might they face? Does the scenario change for linear overcomplete autoencoders? (identity function)
177. Discuss the importance of regularization in the context of Autoencoders.
178. Why does generative autoencoders not require regularization?
179. What are sparse autoencoders?
180. What is a denoising autoencoder? What are its advantages? How does it solve the overcomplete problem?
181. What is score matching? Discuss it's connections to DAEs.
182. Are there any connections between Autoencoders and RBMs?
183. What is manifold learning? How are denoising and contractive autoencoders equipped to do manifold learning?
184. What is a contractive autoencoder? Discuss its advantages. How does it solve the overcomplete problem?
185. Why is a contractive autoencoder named so? (intuitive and mathematical)
186. What are the practical issues with CAEs? How to tackle them?
187. What is a stacked autoencoder? What is a deep autoencoder? Compare and contrast.
188. Compare the reconstruction quality of a deep autoencoder vs. PCA.
189. What is predictive sparse decomposition?
190. Discuss some applications of Autoencoders.
191. What is representation learning? Why is it useful? (for a particular architecture, for other tasks, etc.)
192. What is the relation between Representation Learning and Deep Learning?
193. What is one-shot and zero-shot learning (Google's NMT)? Give examples.
194. What trade offs does representation learning have to consider?
195. What is greedy layer-wise unsupervised pretraining (GLUP)? Why greedy? Why layer-wise? Why unsupervised? Why pretraining?
196. What were/are the purposes of the above technique? (deep learning problem and initialization)
197. Why does unsupervised pretraining work?
198. When does unsupervised training work? Under which circumstances?
199. Why might unsupervised pretraining act as a regularizer?
200. What is the disadvantage of unsupervised pretraining compared to other forms of unsupervised learning?
201. How do you control the regularizing effect of unsupervised pre-training?
202. How to select the hyperparameters of each stage of GLUP?
203. What cross-validation technique would you use on a time series dataset?(Time series data)
204. How would you handle an imbalanced dataset?(Classification Algo in various situations)
205. Name an example where ensemble techniques might be useful.(Ensemble models)
206. What's the "kernel trick" and how is it useful?(SVM)((https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/)
207. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?
208. What is convex hull ? (svm)
209. What are the advantages and disadvantages of decision trees?(DT)
210. What are the advantages and disadvantages of neural networks?(Deep Learning)
211. Why are ensemble methods superior to individual models?(Ensemble Models)
212. Explain bagging.(NLP)
213. What are Recommender Systems?(Recommendation system)
214. Why data cleaning plays a vital role in analysis?(EDA)
215. Differentiate between univariate, bivariate and multivariate analysis.(EDA)
216. What is Linear Regression?
217. What is Collaborative filtering?(recommendation systems)
218. Are expected value and mean value different?
219. What are categorical variables?(classification algo)
220. How can you iterate over a list and also retrieve element indices at the same time?(Python)

221. During analysis, how do you treat missing values?(EDA)
222. Write a function that takes in two sorted lists and outputs a sorted list that is their union. (Python)
223. How are confidence intervals constructed and how will you interpret them?(Probability)
224. How will you explain logistic regression to an economist, physican scientist and biologist?(Logistic regression)
225. Is it better to have too many false negatives or too many false positives?(Performance measurement models)
226. What do you understand by statistical power of sensitivity and how do you calculate it?(Probability)
227. Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.(SVM)
228. Write a program in Python which takes input as the diameter of a coin and weight of the coin and produces output as the money value of the coin.(Programming)
229. What are the basic assumptions to be made for linear regression?(Linear regression)
230. Difference between convex and non-convex cost function; what does it mean when a cost function is non-convex? (SVM)
231. Stochastic Gradient Descent: if it is faster, why don't we always use it?(Linear regression)
232. Difference between SVM and Log R – Easy(SVM)
233. Does SVM give any probabilistic output – I said no it doesn't and it was wrong! He gave me hints but I couldn't figure it out!(SVM)
234. What are the support vectors in SVM
235. Mention the difference between Data Mining and Machine learning?(General)
236. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?(EDA)
237. Why is Naïve Bayes machine learning algorithm naïve?(Naive Bayes)
238. Explain prior probability, likelihood and marginal likelihood in context of naïve Bayes algorithm?
239. What are the three stages to build the hypotheses or model in machine learning?
240. What is the standard approach to supervised learning?
241. What is 'Training set' and 'Test set'?
242. List down various approaches for machine learning?
243. How to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?
244. Name some feature extraction techniques used for dimensionality reduction.
245. List some use cases where classification machine learning algorithms can be used.
246. What kind of problems does regularization solve?
247. How much data will you allocate for your training, validation and test sets?
248. Which one would you prefer to choose – model accuracy or model performance?
249. Describe some popular machine learning methods.
250. What is not Machine Learning?
251. Explain what is the function of 'Unsupervised Learning'?
252. How will you differentiate between supervised and unsupervised learning? Give few examples of algorithms for supervised learning?
253. What is linear regression? Why is it called linear?
254. How does the variance of the error term change with the number of predictors, in OLS?
255. Do we always need the intercept term? When do we need it and when do we not?
256. How interpretable is the given machine learning model?
257. What will you do if training results in very low accuracy?
258. Does the developed machine learning model have convergence problems?
259. Which tools and environments have you used to train and assess machine learning models?
260. How will you apply machine learning to images?
261. What is collinearity and what to do with it?
262. How to remove multicollinearity?
263. What is overfitting a regression model? What are ways to avoid it?
264. What is loss function in a Neural Network?
265. Explain the difference between MLE and MAP inference.
266. What is boosting?
267. If the gradient descent does not converge, what could be the problem?
268. How will you check for a valid binary search tree?
269. How to check if the regression model fits the data well?
270. What are parametric models?()
271. What's the trade-off between bias and variance?
272. Explain how a ROC curve works.(Performance Measurement Models)
273. What is the Box-Cox transformation used for?(Probability)
274. Define precision and recall.(Performance measurement models?
275. what is the function of 'Unsupervised Learning'?(Unsuperwised learning)
276. What is Perceptron in Machine Learning?(Deep Learning)
277. What is ensemble learning?(Ensemble Models)
278. What are the two paradigms of ensemble methods?(Ensemble Models)
279. What is PCA, KPCA and ICA used for?
280. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?(https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/)
281. You are given a data set consisting of variables having more than 30% missing values? Let's say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?(https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/)

282. Compare "Frequentist probability" vs. "Bayesian probability"?
283. What is a random variable?
284. What is a joint probability distribution?
285. What are the conditions for a function to be a probability mass function?
286. What are the conditions for a function to be a probability density function
287. What is a marginal probability? Given the joint probability function, how will you calculate it?
288. What is conditional probability? Given the joint probability function, how will you calculate it?
289. State the Chain rule of conditional probabilities.
290. What are the conditions for independence and conditional independence of two random variables?
291. What are expectation, variance and covariance?
292. Compare covariance and independence.
293. What is the covariance for a vector of random variables?
294. What is a Bernoulli distribution? Calculate the expectation and variance of a random variable that follows Bernoulli distribution?
295. What is a multinoulli distribution?
296. What is a normal distribution?
297. Why is the normal distribution a default choice for a prior over a set of real numbers?
298. What is the central limit theorem?
299. What are exponential and Laplace distribution?
300. What are Dirac distribution and Empirical distribution?
301. What is mixture of distributions?
302. Name two common examples of mixture of distributions? (Empirical and Gaussian Mixture)
303. Is Gaussian mixture model a universal approximator of densities?
304. Write the formula for logistic and softplus function.
305. Write the formula for Bayes rule.
306. What do you mean by measure zero and almost everywhere?
307. If two random variables are related in a deterministic way, how are the PDFs related?
308. Define self-information. What are its units?
309. What are Shannon entropy and differential entropy?
310. What is Kullback-Leibler (KL) divergence?
311. Can KL divergence be used as a distance measure?
312. Define cross-entropy.
313. What are structured probabilistic models or graphical models?
314. In the context of structured probabilistic models, what are directed and undirected models? How are they represented? What are cliques in undirected structured probabilistic models?
315. What is Bayes' Theorem? How is it useful in a machine learning context?
316. Why is "Naive" Bayes naive?
317. What's a Fourier transform?
318. What's the difference between probability and likelihood?
319. Explain prior probability, likelihood and marginal likelihood in context of naive Bayes algorithm?
320. What is the difference between covariance and correlation?
321. Is it possible capture the correlation between continuous and categorical variable? If yes, how?
322. What is the Box-Cox transformation used for?
323. What do you understand by the term Normal Distribution?
324. What does P-value signify about the statistical data?
325. A test has a true positive rate of 100% and false positive rate of 5%. There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?
326. How you can make data normal using Box-Cox transformation?
327. Explain about the box cox transformation in regression models.
328. What is the difference between skewed and uniform distribution?
329. What do you understand by Hypothesis in the content of Machine Learning?
330. How will you find the correlation between a categorical variable and a continuous variable ?
331. What does LogR give ? I said Posterior probability (P(y|x=0 or x=1))
332. Evaluation of LogR –
333. How are the params updated – I was able to answer with formulae!
334. When doing an EM for GMM, how do you find the mixture weights ? I replied that for 2 Gaussians, the prior or the mixture weight can be assumed to be a Bernoulli distribution.
335. If x ~ N(0,1), what does 2x follow
336. How would you sample for a GMM
337. How to sample from a Normal Distribution with known mean and variance.
338. In experimental design, is it necessary to do randomization? If yes, why
339. How do you handle missing or corrupted data in a dataset?
340. Do you have experience with Spark or big data tools for machine learning?
341. In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbours. Why not manhattan distance ? (https://www.analyticsvidhya.com/blog/2017/09/30-questions-test-k-nearest-neighbors-algorithm/)
342. How to test and know whether or not we have overfitting problem?(https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/how-to-determine-overfitting-and-underfitting/)
343. How is kNN different from k-means clustering?(https://stats.stackexchange.com/questions/56500/what-are-the-main-differences-between-k-

means-and-k-nearest-neighbours)

344. Can you explain the difference between a Test Set and a Validation Set?(https://stackoverflow.com/questions/2976452/whats-is-the-difference-between-train-validation-and-test-set-in-neural-netwo)
345. How can you avoid overfitting in KNN?(https://www.appliedaicourse.com/course/applied-ai-course-online/lessons/how-to-determine-overfitting-and-underfitting/)
346. Which is more important to you– model accuracy, or model performance?
347. Can you cite some examples where a false positive is important than a false negative?
348. Can you cite some examples where a false negative important than a false positive?
349. Can you cite some examples where both false positive and false negatives are equally important?
350. What is the most frequent metric to assess model accuracy for classification problems?
351. Why is Area Under ROC Curve (AUROC) better than raw accuracy as an out-of- sample evaluation metric?
352. Define Similarity or Distance matrix.?
353. Time complexity of Naive Bayes algo Best and worst cases?
354. What are the differences between "Bayesian" and "Frequentist" approach for Machine Learning?
355. Compare and contrast maximum likelihood and maximum a posteriori estimation.
356. How does Bayesian methods do automatic feature selection?
357. What do you mean by Bayesian regularization?
358. When will you use Bayesian methods instead of Frequentist methods? (Small dataset, large feature set)
359. After analysing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he's true? Without losing any information, can you still build a better model?(https://google-interview-hacks.blogspot.in/2017/04/after-analyzing-model-your-manager-has.html)
360. What are the basic assumptions to be made for linear regression?(Refer:https://www.statisticssolutions.com/assumptions-of-linear-regression/)
361. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)? (https://stats.stackexchange.com/questions/317675/gradient-descent-gd-vs-stochastic-gradient-descent-sgd)
362. When would you use GD over SDG, and vice-versa?(https://elitedatascience.com/machine-learning-interview-questions-answers)
363. How do you decide whether your linear regression model fits the data? (https://www.researchgate.net/post/What_statistical_test_is_required_to_assess_goodness_of_fit_of_a_linear_or_nonlinear_regression_equation)
364. Is it possible to perform logistic regression with Microsoft Excel?(https://www.youtube.com/watch?v=EKRjDurXau0)
365. When will you use classification over regression?(https://www.quora.com/When-will-you-use-classification-over-regression)
366. Why isn't Logistic Regression called Logistic Classification?(Refer :https://stats.stackexchange.com/questions/127042/why-isnt-logistic-regression-called-logistic-classification/127044)
367. Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa. (https://datascience.stackexchange.com/questions/6838/when-to-use-random-forest-over-svm-and-vice-versa)
368. What is convex hull ?(https://en.wikipedia.org/wiki/Convex_hull)
369. What is a large margin classifier?
370. Why SVM is an example of a large margin classifier?
371. SVM being a large margin classifier, is it influenced by outliers? (Yes, if C is large, otherwise not)
372. What is the role of C in SVM?
373. In SVM, what is the angle between the decision boundary and theta?
374. What is the mathematical intuition of a large margin classifier?
375. What is a kernel in SVM? Why do we use kernels in SVM?
376. What is a similarity function in SVM? Why it is named so?
377. How are the landmarks initially chosen in an SVM? How many and where?
378. Can we apply the kernel trick to logistic regression? Why is it not used in practice then?
379. What is the difference between logistic regression and SVM without a kernel? (Only in implementation – one is much more efficient and has good optimization packages)
380. How does the SVM parameter C affect the bias/variance trade off? (Remember C = 1/lambda; lambda increases means variance decreases)
381. How does the SVM kernel parameter sigma^2 affect the bias/variance trade off?
382. Can any similarity function be used for SVM? (No, have to satisfy Mercer's theorem)
383. Logistic regression vs. SVMs: When to use which one? (Let's say n and m are the number of features and training samples respectively. If n is large relative to m use log. Reg. or SVM with linear kernel, If n is small and m is intermediate, SVM with Gaussian kernel, If n is small and m is massive, Create or add more features then use log. Reg. or SVM without a kernel)
384. What is the difference between supervised and unsupervised machine learning?
385. You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?(Refer :https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/)
386. Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?(Refer:https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/)
387. You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?(Refer : https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/)
388. How would you implement a recommendation system for our company's users?(https://www.infoworld.com/article/3241852/machine-learning/how-to-implement-a-recommender-system.html)
389. How would you approach the "Netflix Prize" competition?(Refer http://blog.echen.me/2011/10/24/winning-the-netflix-prize-a-summary/)
390. 'People who bought this, also bought…' recommendations seen on amazon is a result of which algorithm?(Please refer Apparel

recommendation system case study,Refer:https://measuringu.com/affinity-analysis/)

391. Pick an algorithm. Write the psuedo-code for a parallel implementation.
392. What are some differences between a linked list and an array?(Programming)
393. Describe a hash table.
394. What is sampled softmax?
395. Why is it difficult to train a RNN with SGD?
396. How do you tackle the problem of exploding gradients? (By gradient clipping)
397. What is the problem of vanishing gradients? (RNN doesn't tend to remember much things from the past)
398. How do you tackle the problem of vanishing gradients? (By using LSTM)
399. Explain the memory cell of a LSTM. (LSTM allows forgetting of data and using long memory when appropriate.)
400. What type of regularization do one use in LSTM?
401. What is Beam Search?
402. How to automatically caption an image? (CNN + LSTM)
403. What is the mathematical motivation of Deep Learning as opposed to standard Machine Learning techniques?
404. In standard Machine Learning vs. Deep Learning, how is the order of number of samples related to the order of regions that can be recognized in the function space?
405. What are the reasons for choosing a deep model as opposed to shallow model? (1. Number of regions O(2^k) vs O(k) where k is the number of training examples 2. # linear regions carved out in the function space depends exponentially on the depth.)
406. How Deep Learning tackles the curse of dimensionality?
407. Why do RNNs have a tendency to suffer from exploding/vanishing gradient? How to prevent this? (Talk about LSTM cell which helps the gradient from vanishing, but make sure you know why it does so. Talk about gradient clipping, and discuss whether to clip the gradient element wise, or clip the norm of the gradient.)
408. What is the problem with sigmoid during backpropagation? (Very small, between 0.25 and zero.)
409. What is transfer learning?
410. Write the equation describing a dynamical system. Can you unfold it? Now, can you use this to describe a RNN? (include hidden, input, output, etc.)
411. What determines the size of an unfolded graph?
412. What are the advantages of an unfolded graph? (arbitrary sequence length, parameter sharing, and illustrate information flow during forward and backward pass)
413. What does the output of the hidden layer of a RNN at any arbitrary time *t* represent?
414. Are the output of hidden layers of RNNs lossless? If not, why?
415. RNNs are used for various tasks. From a RNNs point of view, what tasks are more demanding than others?
416. Discuss some examples of important design patterns of classical RNNs.
417. Write the equations for a classical RNN where hidden layer has recurrence. How would you define the loss in this case? What problems you might face while training it? (Discuss runtime)
418. What is backpropagation through time? (BPTT)
419. Consider a RNN that has only output to hidden layer recurrence. What are its advantages or disadvantages compared to a RNNhaving only hidden to hidden recurrence?
420. What is Teacher forcing? Compare and contrast with BPTT.
421. What is the disadvantage of using a strict teacher forcing technique? How to solve this?
422. Explain the vanishing/exploding gradient phenomenon for recurrent neural networks. (use scalar and vector input scenarios)
423. Why don't we see the vanishing/exploding gradient phenomenon in feedforward networks? (weights are different in different layers – Random block intialization paper)
424. What is the key difference in architecture of LSTMs/GRUs compared to traditional RNNs? (Additive update instead of multiplicative)
425. What is the difference between LSTM and GRU?
426. Explain Gradient Clipping.
427. Adam and RMSProp adjust the size of gradients based on previously seen gradients. Do they inherently perform gradient clipping? If no, why?
428. Discuss RNNs in the context of Bayesian Machine Learning.
429. Can we do Batch Normalization in RNNs? If not, what is the alternative? (BNorm would need future data; Layer Norm)
430. What is representation learning? Why is it useful? (for a particular architecture, for other tasks, etc.)
431. What is the relation between Representation Learning and Deep Learning?
432. What is one-shot and zero-shot learning (Google's NMT)? Give examples.
433. What trade offs does representation learning have to consider?
434. What is greedy layer-wise unsupervised pretraining (GLUP)? Why greedy? Why layer-wise? Why unsupervised? Why pretraining?
435. What were/are the purposes of the above technique? (deep learning problem and initialization)
436. Why does unsupervised pretraining work?
437. When does unsupervised training work? Under which circumstances?
438. Why might unsupervised pretraining act as a regularizer?
439. What is the disadvantage of unsupervised pretraining compared to other forms of unsupervised learning?
440. How do you control the regularizing effect of unsupervised pre-training?
441. How to select the hyperparameters of each stage of GLUP?