Machine Learning and Deep Learning Interview Questions

1. Difference between supervised and unsupervised learning with one example each.
2. What does linear regression do?
3. What will be the the equation for linear regression when we have only one feature x and one predicted value y.
4. If there are **m** training examples what will be the cost function in Q3 above?
5. Which metric is used to measure performance of a linear (or polynomial) regression?
6. Difference between multivariate and linear regression.
7. What does logistic regression do?
8. What is Sigmoid function?
9. What is cross entropy?

10. For an image classification problem involving plant species, with two classes **weed** and **herb** the ground truth and predicted probabilities are given below. What is the cross entropy loss?

	Ground Truth	Predicted Probability
Weed	1	0.79
Herb	0	0.21

11. When is principal component analysis (PCA) used? What does it do?

12. Suppose a training set has **m** examples and **n** features per example. If the examples are stored as the rows of a matrix **X,** what will be the dimensions of the feature covariance matrix that is fed to PCA algorithm?

13. In a training set of heart sounds, consisting of 10,000 recordings each lasting between 10 and 20 seconds, 8847 recordings are labelled as normal while 153 recordings are labelled as abnormal. What metric(s) should be used for measuring the performance of a

classification algorithm? Why is accuracy not a good metric for this dataset?

14. What is cross validation? What will be the size of training and validation sets for a dataset with 8000 examples, if we use 5-fold cross validation.

15. A classification algorithm has 99.8% accuracy on training set but 94% accuracy on validation set. This problem is likely to be caused by overfitting or underfitting?

16. In [15], does the model (algorithm) suffer from high variance or high bias?

17. A classification model has 94% training accuracy and 92% validation accuracy. Does it have high variance or high bias?

18. What is regularization? What do L1 and L2 regularizers do?

19. For a classification problem with just one feature **x0,** bias **w0** and weight **w1**, what will be the expression for **L2** regularization?

20. What is the effect of increasing the number of features on the variance of a model?

21. If a model has high bias and it undefits a dataset, what effect will increasing the dataset size have on the model?

22. One solution for mitigating underfitting (high-bias in model) is to increase the variance of a model by increasing the size (number of parameters/weights) of the model. What issues might arise when using this approach.
23. How does support vector machine (SVM) differ from logistic regression?
24. How does a decision tree classify data?
25. What is impurity in the context of decision tree algorithms?
26. What is a kernel in a SVM classifier?
27. What is the curse of dimensionality?
28. Is K-Nearest Neighbors (KNN) a supervised or unsupervised classification algorithm?
29. Is K-means clustering a supervised or unsupervised method?
30. What are true positives, false positives, true negatives and false negatives.
31. What does precision measure?
32. What does recall measure?
33. How is F1 score useful?

34. In a training set of heart sounds, consisting of 10,000 recordings each lasting between 10 and 20 seconds, 8847 recordings are labelled as normal while 153 recordings are labelled as abnormal. When a classification algorithm is trained on this data, it classifies 100 abnormal sounds correctly while misclassifying 53 abnormal sounds as normal. The algorithm classifies 8800 recordings out of 8847 normal recordings as normal while 47 normal sounds are misclassified as abnormal. What are the accuracy, precision, recall and F1 score for this algorithm?

35. How does Naive Baye's algorithm classify?

36. Given the following data for temperature, humidity (L=low, M=moderate, H=high) and the observed weather conditions (rainy=1, dry=0), apply Naive Baye's classifier to predict the probability of it being a rainy or dry day when temperature is high (H) and so is humidity.

Temperature	Humidity	Rainy(=1)/Dry(=0)
L	L	0
M	L	0
H	L	0
L	M	1

M	M	1
H	M	0
L	H	1
M	H	1
H	H	1

37. Suppose a classifier for herb(=1) vs weed(=0) case is classifying too many weed plants as herbs. One way to fix this is to increase the threshold of the classifier from 0.5 to a larger value. This will reduce the number of false positives, but also, likely, decrease the number of true positives. The threshold can be further varied and we get another pair of true positive and false positive values. By plotting the true positive rate (TPR) against the false positive rate at many different thresholds, one gets a curve. What is the curve called? How is the curve useful in finding the sweet-spot for the threshold of a classifier?
38. What is Area Under Receiver Operating Characteristic (AuC) ?
39. How do sigmoid and softmax activations differ?
40. What is multi-label classification how is it different from multi-class classification?
41. What is the Markov assumption?
42. What is naive about Naive Baye's classifier?

43. What does an activation function do?
44. What is dropout?
45. What does gradient descent do?
46. How is stochastic gradient descent different from gradient descent?
47. What is the purpose of one-hot encoding?
48. Why is training done in mini batches?
49. What is an epoch?
50. What is the formula for ReLU activation?
51. What does backpropagation do (in terms of gradients)?
52. If we have a 28x28 image with RGB data and we apply a 3x3 convolution with stride 1 and number of kernels set to 32 what will be the size of the convolution layer's output?
53. In [52] what will be the number of parameters in the convolution layer?
54. What does pooling do?
55. What is 1x1 convolution used for?
56. What are generative networks?

57. What is a text embedding? Given a dictionary and a text corpus, describe how to build an embedding for the corpus?
58. What is negative sampling in the context of training a Word2Vec model?
59. What is boosting in machine learning?
60. What are some some optimizers (cost optimization algorithms) for neural networks apart from Stochastic Gradient Descent?
61. What is the exploding/vanishing gradients problem?
62. Which kind of neural networks are more powerful: shallow and wide or deep and narrow?
63. In the absence of more data how can one train a high variance model?
64. What is transfer learning?
65. How do you find the hyperparameters of a model?
66. How do GridSearch and Random Search differ work for finding the best hyperparameters?
67. How can you augment audio data, image data?
68. What is the gist of a generative adversarial network?
69. How is the conditional probability P(A|B) different from P(A ∩ B)?

70. Why are convolution model more effective than fully connected models for image classification.

71. What type of classification is best suited for learning to play football?