

# Github ml interview

11/10/19 3:35 PM

## DS-ML-Interview-Questions

Last update: 11/16/2018 To be a good DS or MLE, you have to answer most of these questions below quickly and precisely.

### Behavior and Background

1. Introduce yourself.
2. Tell me your past internship or some other projects.

### Machine Learning Concept

#### Linear Regression and Logistic Regression

1. What is linear regression? What is logistic regression? What is the difference?
2. How to find the best parameter for linear regression? / How to optimize the model?
3. Please write down the close form solution for linear regression?
4. What is Stochastic Gradient Descent or Mini-batch Gradient Descent? What is the advantage? and disadvantages?
5. What is mean square error? What is Cross Entropy? What is the difference? Also please write down the formular of these two cost function.
6. What is Softmax? What is relationship between Softmax and Logistic Regression?
7. Explain and derive the SGD for Logistic Regression and Softmax?
8. Does global optimal can be reached by SGD, why?
9. What is the Sigmoid function? What is the characteristic of Sigmoid? What is the advantages and disadvantages of Sigmoid function?

#### Regularization, Overfitting and Model/Feature Selection/Evaluation.

<https://medium.com/analytics-vidhya/regularisation-techniques-in-machine-learning-and-deep-learning-8102312e1ef3>

1. What are L1 and L2? And their difference?
2. From the view of Bayesian, what's the difference between L1 and L2?
3. What is overfitting?
4. How can you know your model is overfitted or not?
5. What is the Bias-Variance Trade-off?
6. How to prevent overfitting?
7. What is cross validation?
8. Let is consider you are training a classifier to classify the cat pictures taken by cell phone. You have 10k cat pictures that taken by cell phone users. How would you split the pictures into training/validation/test set? Now you have got 100k cat pictures from Internet, what dataset would you like to choose to put these 100k cat pictures in?
9. For training/validation/test set, which two sets are most important that you have to keep the distribution of data samples the same?
10. What is data augmentation? Do you know any technique to augment data?
11. What is F1 score? What are recall and precision?
12. What is AUC?
13. How would you handle data imbalance problem?

### Decision Tree

1. What is Decision Tree?
2. What is Information Gain?
3. What is Geni Index?
4. What is the advantages and disadvantages of ID3?
5. What is Random Forrest?
6. What is Bagging?

### Boosting and Ensemble

1. What is AdaBoost and the relation between Adaboost and exponential loss function?
2. What is Gradient Boosting?
3. What is the idea of Bagging and Stacking?
4. Do you know XGBoost? And what is the idea of XGBoost?

### Naive Bayes

1. Write down Naive Bayes equation.
2. Given an example, calculate designate probability by using Bayes Equation.

### Unsupervised Learning

1. What's Clustering?
2. What's K-means? Implement K-means by Python and Numpy. What's the relationship between K-means and EM?
3. What's the pros and cons of K-means?
4. What's the complexity of K-means?
5. What's PCA and SVD? Given the SVD funtion, please implement PCA.
6. Do you know T-sne, simply explain it to me?

### Graph Models

1. What is Hidden Markov Model?
2. What is the assumption that HMM made?
3. There are three matrices in HMM, what are they?
4. What is the three problems of HMM?
5. What is the Viterbi Algorithm? And its complexity?
6. How to optimize the parameters of HMM?
7. In what situation you would likt to use HMM?
8. (Bonus) What are MEMM and CRF?

### Support Vector Machine

1. What is suppor vector?
2. What is the idea of SVM?
3. Explain the idea of kernel method?
4. What is the slack factor in SVM?
5. What is the loss function of SVM?

### EM

1. What is the idea of EM algorithm?
2. In what case we would like to use EM?

### Reinforcement Learning

1. What is the Markov Decision Process?
2. What is the Bellman Equation?
3. What is the Q-function.
4. The difference between Policy Gradient and Q-learning.

### Deep Learning

1. What is the relationship between Logistic Regression and Feedforward Neural Networks?
2. What is Sigmoid, Tanh, Relu? And their pros and cons?
3. What are RNN, LSTM, GRU? And their pros and cons?
4. What are gradient explosion and vanishing?
5. What is CNN, explain the process of CNN and the idea.
6. What is the differencnt between Momentum and Adam optimizer?
7. What is the pros and cons of Tensorflow and PyTorch?
8. What is the compuational graph of deep learning frameworks?
9. Why GPU can accelerate the computation of deep learning models?
10. Why deep learning models are powerful now?
11. What is Batch Normalization and Layer Normalization?
12. What is Dropout?
13. In what case you would like to use transfer learning? How would you fine-tune the model?
14. Do you know any techniques to initialize deep learning models?
15. Why zero initialization is a pitfall for deep learning?
16. Implement a single hidden layer FNN by Python and Numpy.  
<https://towardsdatascience.com/how-to-build-your-own-neural-network-from-scratch-in-python-68998a08e4f6>

### NLP and DL

1. What is tf-IDF?
2. What is word embedding and its idea? What is the difference between sampled softmax and negative sampling? Do you know the relationship between negative sampling and Pointwise Mutual Information(Bouns)
3. What are unigram and bigram?
4. What is the attention machanism?

### Computer Vision and DL

1. What is the difference between simple CNN and Inception Net and ResNet?
2. What are the common techniques to preprocess the image data?
3. What are the kernel and pooling in CNN?
4. What is GAN, the structure and the way we train it.

### Case Study

You are given a specifi case and you need to give a solution for the problem.

### Coding

#### SQL

1. select, group by, left, right, inner outter join...

#### Python coding and concepts

1. What is difference between Tuple and List?
2. What is the list comprehension?
3. Use ML packages to complete certain tasks.
4. Some medium level (refer to Leetcode) coding questions.

#### Pandas

1. Use Pandas to manipulate data.