Machine Learning for Data Science Interviews

Adarsh Chavakula

Columbia Data Science Society

Format:

- Topic
- Easy questions
- Hard questions - Alternate ways to frame easy questions

(Answers provided only to select questions)

Contents

PART 1

Fundamental Concepts

General

Q1 - What is a model?

Q2 - What’s the difference between a model and an algorithm?

Q3 - What’s the difference between supervised and unsupervised learning?

Q4 - What’s the difference between classification, regression, clustering and recommendation?

Concept: Overfitting and Underfitting

EASY:

Q1 - What is overfitting, underfitting, difference between the two

Q2 - How do you prevent overfitting?

Q3 - How do you prevent underfitting?

Concept: Regularization

EASY:

Q1 - What is regularization?

Q2 - What is L1 regularization, L2 regularization?

Concept: Model Evaluation

EASY:

Q1 - What is k-fold cross validation?

Q2 - Why do you need a validation dataset?

Q3 - What is Out-of-Sample and Out-of-Time cross validation?

Concept: Evaluation Metrics

EASY:

Q1 - RMSE, MAE, R^2, Accuracy, AUC-ROC, Gini, Categorical Cross Entropy, Precision/Recall, Confusion Matrix, F1 Score, MAP@K?

Q2 - What's the problem with Accuracy?

Concept: Class Imbalance

EASY:

Q1 - Why is class imbalance a problem?

Q2 - What are some strategies to deal with class imbalance?

Q3 - What evaluation metrics should you use when dealing with imbalanced classes? Which one should you **NOT** use?

Concept: Multi-Class Classification

EASY:

Q1 - Which evaluation metric?

Q2 - What is one-vs-all classification?

Q3 - Can I use RMSE as an evaluation metric?

Concept: Multi-Class Classification

EASY:

Q1 - Which evaluation metric?

Q2 - What is one-vs-all classification?

Q3 - Can I use RMSE as an evaluation metric?

Answer **-** *Sometimes*. Classes can be ordinal.

Concept: Very High Dimensional Data

EASY:

Q1 - How do you deal with it?

Q2 - What is feature selection? Strategies to do it?

Q3 - What is dimensionality reduction? Strategies to do it?

Answer: PCA, t-SNE, Auto-encoders

Topic: Model Parameters

EASY:

Q1 - What is the bias-variance tradeoff? *(Statisticians love asking this)*

Q2 - How do you decide the number of parameters in a model?

Topic: Model Parameters

EASY:

Q1 - What is the bias-variance tradeoff? *(Statisticians love asking this)*

Q2 - How do you decide the number of parameters in a model?

Answer - ***Theoretical***: *Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC).* ***Practical***: *Cross Validation*

Topic: Variable Pre-processing

EASY:

Q1 - What is feature normalization? What are the different methods?

Answer - Standardization (subtract mean and divide by deviation, min-max scaling, log transformation)

Q2 - Why is this required?

Q3 - What is One Hot Encoding?

Topic: Optimization (1)

EASY:

Q1 - What is the mathematical expression for gradient descent for updating weights?

Q2 - How does stochastic gradient descent (SGD) work?

Q3 - Is Gradient descent guaranteed to attain global minima?

Topic: Optimization (2)

EASY:

Q1 - What is learning rate (or step size)?

Q2 - What happens if learning rate is too high (or too low)?

Q3 - What are convex functions?

PART 2

ML Algorithms

Algorithm: Linear Regression

EASY:

Q1 - What happens if the number of features (columns) is greater than the number of samples (records/rows)

Q2 - What is Ridge Regression, Lasso, Elastic Net?

Q3 - What are the limitations of LR? How can they be addressed?

Algorithm: Logistic Regression

EASY:

Q1 - Is there a closed-form solution to Logistic Regression? If yes, derive it. If no, explain how to arrive at the optimal solution.

Q2 - Difference between linear and logistic regression?

Q3 - What are the limitations of LR? How can they be addressed?

Algorithm: Support Vector Machines

EASY:

Q1 - What is the difference between logistic regression and SVM?

Q2 - What is the Kernel Trick.

Q3 - How does Hinge Loss work?

Q4 - What is the dual of an SVM? Why is it useful?

Algorithm: Decision Trees

EASY:

Q1 - How are the “cuts” decided in each node of a tree?

Q2 - What is entropy and gini-impurity?

Q3 - When would a DT overfit? How can it be avoided?

Algorithm: k-Nearest Neighbors

EASY:

Q1 - What is the time complexity of brute-force kNN?

Q2 - Explain how the KD-Tree or Ball-Tree variations of kNN work.

Q3 - How do you choose k?

Algorithm: k-Means Clustering

EASY:

Q1 - What is the time complexity of brute-force k-Means clustering?

Q2 - Difference between k-means clustering and kNN?

Q3 - How do you choose k?

Algorithm: Naïve Bayes

EASY:

Q1 - What's the naive assumption made in Naive Bayes?

Q2 - What are the strengths and weaknesses of NB?

Algorithm: Neural Networks

EASY:

Q1 - What is a Saturated Neuron? (**OR** what is vanishing gradient? **OR** what is the problem with sigmoid/tanh activations)

Q2 - What are dropouts? How do they help?

Q3 - How do you determine the number of nodes and layers?

Algorithm: Neural Networks

EASY:

Q1 - What is a Saturated Neuron? (**OR** what is vanishing gradient? **OR** what is the problem with sigmoid/tanh activations?)

Q2 - What are dropouts? How do they help?

Q3 - How do you determine the number of nodes and layers?

ANSWER: You cannot pre-determine the best architecture. It is often done by trial and error.

Topic: Ensemble Models

EASY:

Q1 - What are the different model ensembling techniques? Explain each of them. Examples of each?

Q2 - Why do ensembling?

Topic: Ensemble Models

EASY:

Q1 - What is the difference between bagging and boosting?

OR What is the difference between Random Forests and Gradient Boosted Decision Trees?

PART 3

Challenging/ Conceptual Questions

Topic: ?

Q1 [Medium] - We have a dataset with a large number of features (several hundred). Based on our business understanding of these features, we know that most of these features may not contain any useful signals. What should be the modeling strategy for a **linear model**?

Answer in next slide

Topic: Regularization / Feature Selection

Answer 1 -

L1 Regularization: Since we know that most features don't have any signal, L1 would allow us to disregard most of them and build a sparse model. *(Why not L2?)*

Feature Selection: We can use a feature selection strategy like Recursive Feature Elimination to retain only useful features (this is time consuming and is best done only once).

We may not want to use a dimensionality reduction technique like PCA *(why?)*

Topic: ?

Q2 [Medium] - A model for detecting cancer has a 98% **test accuracy**. It still fails to detect any positive cases in production. What may be the reason?

Answer in next slide

Topic: Class Imbalance + Evaluation Metrics

Answer 2 -

It is likely that the model was trained on a highly imbalanced dataset (say 98% “not-cancer”). Hence the model can just label everything it sees in the test data as “not-cancer” and still be 98% accurate. A metric like AUC would have been more accurate for this problem. *(what would be the AUC for this model?)*

Topic: ?

Q3 [Medium] I have 2 models - A and B. Model A has a test accuracy of 95% and takes 30 minutes to train. Model B has an accuracy of 96% and takes 4 days to train. Which model should I use?

Answer in next slide

Topic: Evaluation Metrics

Answer 3 - ***It depends***. More information is needed to take a decision.

In some cases, a 1% improvement in accuracy can translate to a lot of $$$ (think algorithmic trading) or could be life critical (think cancer detection). Model B is better for such cases.

Model A is great when the performance is not as critical or if models need to be refreshed often.

Topic: Ensemble Models

Q3 [Difficult] I have 2 models - A and B.

Model A has an accuracy of 60% and

Model B has an accuracy of 80%

Which model is better suited for bagging and which one for boosting?

Topic: Neural Networks

Q4 [Medium] How are neural networks a generalization of Linear/Logistic regression? How can you reduce a NN to Linear/Logistic regression?

Topic: Pre-Processing

Q5 [Difficult] Which ML algorithms require input features to be scaled/normalized and which ones don’t?

Answer in next slide

Topic: Pre-Processing

Answer 5: Tree based methods (DT, Random forest, gradient boosted decision trees) and Naive Bayes are unaffected by any monotonic transformation of the input data.

All others are affected (usually in a positive way) from scaling/normalization

Topic: Pre-Processing / Algorithms

Q6 [Difficult] I accidentally forgot to one-hot encode my categorical variables when using a Random Forest but it still gave results which were comparable to the time when I one-hot encoded them.
Why would this happen? *(Assume the categories are integers and are important indicators for the variable being predicted)*

Answer in next slide

Topic: Pre-Processing / Algorithms

Answer 6: Given sufficient depth, tree based models can eventually slice each variable a large number of times, allowing it to be treated as a true categorical variable.

Topic: ?

Q7 [Difficult] I am building a stock price predictor to forecast price changes in the stock market. I performed k-fold cross validation on my awesome deep learning model and the results were amazing.

When I actually started using it for trading, it was a disaster. What did I do wrong?

Answer in next slide

Topic: Cross Validation!

Q8 [Difficult] When making a model to forecast the future, one should always use out-of-time validation instead of randomly shuffled k-fold cross validation. Data points from the “future” should not be used for training.

Topic: ?

Q9 (difficult): What is the intuition behind X?

Common X substitutes: PCA, Stochastic Gradient Descent, Regularization, Dropouts, Bagging, Boosting, Auto-encoders

The hardest question

Q10: What is your favorite ML algorithm?

(Messing this up would definitely end your interview)

Skills?

Q11: Are you good at using Scikit-Learn?

(There's only one correct answer to this)

Q12. Are you good at using Tensorflow?

Controversial

Q What is the best Machine Learning algorithm?

PART 3

Specific Applications

Topic: Time Series Analysis

Q1: What is Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)?

Q2: What is a stationary time series? How can you “stationarize” a non-stationary series?

Q3: What is heteroscedasticity? Why is it a problem?

Topic: Recommendation Systems

Q1: What is the intuition behind Collaborative Filtering?

Q2: What is the cold start problem?

Q3: How does SVD++ work?

Q4: How do Factor Machines work?

Images - Convolutional Neural Nets (1)

Q1: What is pooling?

Q2: Can a CNN trained on images of a specific size work on images of a different size?

Q3: What is the issue with Fully Connected Feedforward Neural Nets which is addressed by CNNs?

Images - Convolutional Neural Nets (2)

Q1: What is transfer learning?

Q2: What are some frequently used pre-trained networks used in practice for image classification?

Q3: How does image captioning work?

NLP - Recurrent Neural Nets

Q1: What is exploding (or vanishing) gradient?

Q2: What are LSTMs and GRUs?

Q3: How do you address the problem of variable sequence lengths?

Thank You

cdss_execs@columbia.edu

Columbia Data Science Society