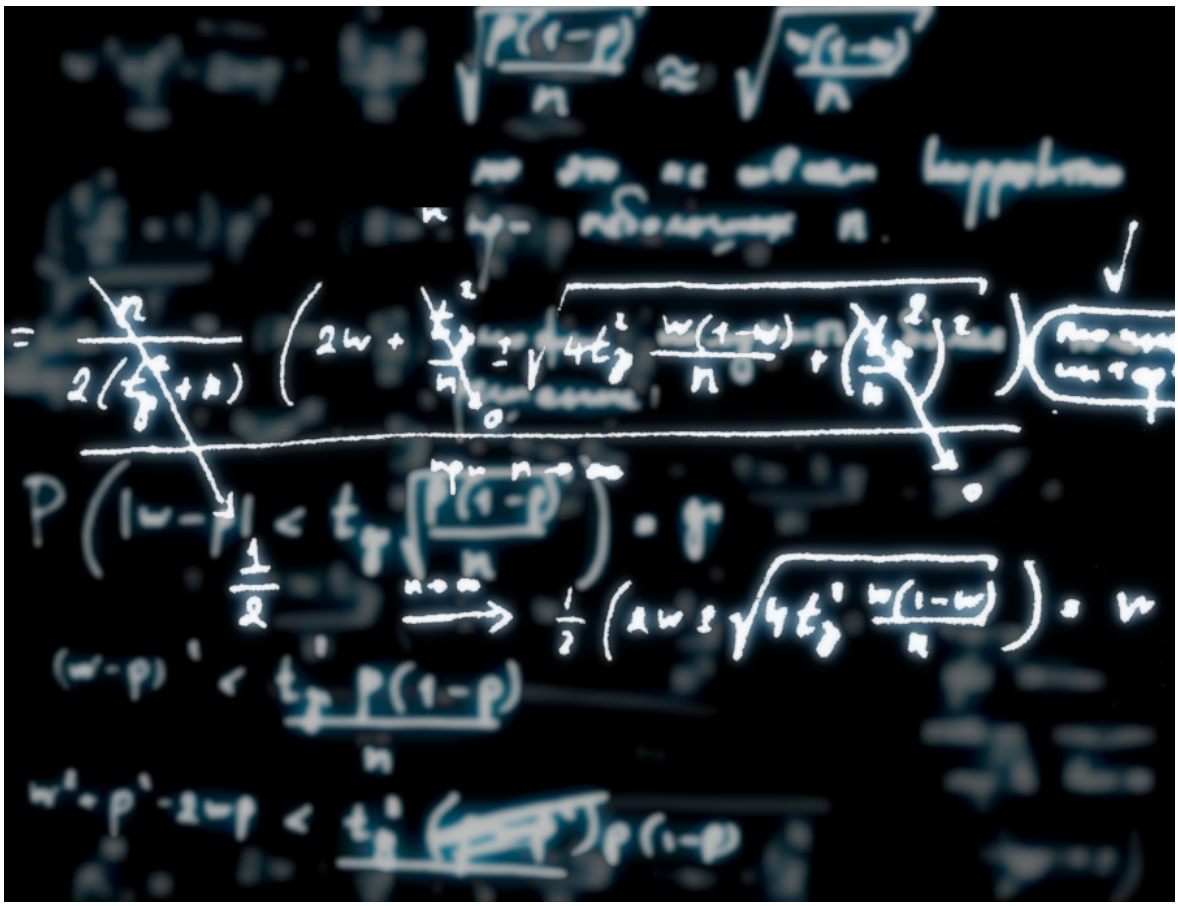
January 9, 2017

41 Essential Machine Learning Interview Questions (with answers)

Roger Huang



Machine learning interview questions are an integral part of the data science interview and the path to becoming a data scientist, machine learning engineer or data engineer. Springboard created a free guide to data science interviews so we know exactly how they can trip candidates up! In order to help resolve that, here is a curated and created a list of key questions that you could see in a machine learning interview. There are some answers to go along with them so you don't get stumped. You'll be able to do well in any job interview with machine learning interview questions after reading through this piece.

Machine Learning Interview Questions: Categories

We've traditionally found that machine learning interview questions pop up in several categories. The first really has to do with the algorithms and theory behind machine learning. You'll have to show an understanding of how algorithms compare with one another and how to measure their performance and accuracy in the right way. The second category has to do with your programming skills and your ability to execute on top of these algorithms and the theory. The third has to do with your general interest in machine learning: you'll be asked about what's going on in the industry and how you keep up with the latest machine learning trends. Finally, there are company or industry-specific questions that test your ability to take your general machine learning knowledge and turn it into actionable points to drive the bottom line for

We've divided this guide to machine learning interview questions into the categories we mentioned above so that you can more easily get to the information you need when it comes to machine learning interview questions.

Machine Learning Interview Questions: Algorithms/Theory

These algorithms questions will test your grasp of the theory behind machine learning.

Q1- What's the trade-off between bias and variance?

More reading: Bias-Variance Tradeoff (Wikipedia)

Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you're using. This can lead to the model underfitting your data, making it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set.

Variance is error due to too much complexity in the learning algorithm you're using. This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead your model to overfit the data. You'll be carrying too much noise from your training data for your model to be very useful for your test data.

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance. You don't want either high bias or high variance in your model.

Q2- What is the difference between supervised and unsupervised machine learning?

More reading: What is the difference between supervised and unsupervised machine learning? (Quora)

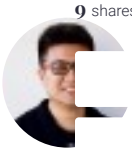
Supervised learning requires training labeled data. For example, in order to do classification (a supervised learning task), you'll need to first label the data you'll use to train the model to classify data into your labeled groups. Unsupervised learning, in contrast, does not require labeling data explicitly.

Q3- How is KNN different from k-means clustering?

More reading: How is the k-nearest neighbor algorithm different from k-means clustering? (Quora)

K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this really means is that in order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled point into (thus the nearest neighbor part). K-means clustering requires only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between different points.

The critical difference here is that KNN needs labeled points and is thus supervised learning, while k-means doesn't — and is thus unsupervised learning.



Roger Huang

Roger has always been inspired to learn more. He has written for Entrepreneur, TechCrunch, The Next Web, VentureBeat, and Techvibes. Previously, he led Content Marketing and Growth efforts at Springboard.

You might also be interested in..

DATA SCIENCE

19 Free Public Data Sets For Your First Data Science Project

Completing your first project is a major milestone on the road to becoming a data scientist. It's also an intimidating process.

READ MORE

DATA SCIENCE

41 Essential Machine Learning Interview Questions (with answers)

Machine learning interview questions are an integral part of the data science interview and the path to becoming a data

READ MORE

DATA SCIENCE

Data Science Career Paths: Different Roles in the Industry

Data Science Career Paths: Introduction We've just come out with the first bootcamp with a data science job guarantee to

READ MORE

RESOURCES

Free learning

E-books and guides

View all resources

COURSES

Introduction to Data Science

Intermediate Data Science: Python

Data Science Career Track

Business Analytics

UX Design

Digital Marketing Career Track

Introduction to Cybersecurity

Cybersecurity Career Track

ABOUT US

About the company

Meet the team

Investors

Jobs

Become a mentor

Hire our students

Corporate training

CONTACT US

Frequently Asked Questions

Send us an e-mail

Like us on Facebook

Tweet us on Twitter

Read our stories on Medium

Springboard