Gmail

Andrew Ho <kironide@gmail.com>

Manager Questions & Interview Self-Study

Jonah & Robert Signal Data Science <signaldatascience@gmail.com> Wed, Mar 16, 2016 at 10:32 PM
To: Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, David Bolin <david@bolin.at>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Jonah & Robert Signal Data Science <signaldatascience@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Robert Cordwell <cordwell@gmail.com>, Sam Eisenstat <sam.eisenst@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

Heya!

Some of you have been asking for questions to test yourselves & use for interview prep, so I've compiled a list of what I like to call "Manager Questions," so named because they're the sort of high-level questions that a manger might ask if giving a technical interview. You can get it here. Please add questions from your own interviews!

Note that these questions are a fair bit harder than actual interview questions--I can't answer every single one of these off the top of my head. In fact, we're giving you these questions with the intent that you'll try and answer them, Google the answer if you can't, and fill in some gaps.

Bold questions are somewhat more important than the rest. We strongly recommend that, even if this feels overwhelming, you start by focusing on being able to answer every single bold question comfortably.

- What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

- Describe two ways I can make my model more robust to outliers.

- What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?

- **What is a classifier?** What are some techniques I might use for classification? Explain them and why they're useful.

- What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

- **What's the difference between supervised and unsupervised learning**

- **Explain cross validation, and provide an example of when you might want to use it.**

- What is k-fold cross validation?

- What are various ways to predict a binary response variable? Can you compare two of them and give me a situation where each would be more appropriate?

- What is regularization? Can you describe a situation where it would be useful?

- **Why might it be preferable to include fewer predictors over many?**

- Your linear regression didn’t run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?

- You run your regression on different subsets of your data, and notice that in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

- How and why does ensemble learning work?

- **Can you explain an A/B test to an engineer with no statistics background?**

- Can you explain linear regression to such an engineer?

- Can you explain what "95% confidence" means to the engineer?

- **Tell me about a dataset that you’ve analyzed. What techniques did you find helpful, and which ones didn’t work?**

- **What’s your favorite algorithm? Can you explain what it does and why it works to me?**

Best,
Robert