Interview Questions

Signal Data Science

A collection of interview questions along with brief notes about how to solve them.

Data science questions

General data analysis

- What is R^2? What are some other metrics that could be better than R^2?
  - R^2 represents the percentage of variation in the target variable explained by variation in the RMSE is a better measure of prediction accuracy than the square of the correlation.
- What is the curse of dimensionality?
  - Algorithms break down in high dimensions. The standard Euclidean distance metric doesn't work as well; distances between pairs of data points all become very similar.
- Is more data always better?
  - Yes from the perspective of pure prediction, but you may not always want to work with all the data you have (at least immediately); it's expensive to store a large amount of data, training models takes longer, the dataset may not fit in memory, etc.
- What are advantages of plotting your data before performing analysis?
  - See Anscombe's quartet – summary statistics don't tell all.
- How can you make sure that you don't analyze something that ends up meaningless?
  - https://www.quora.com/How-can-you-make-sure-that-you-dont-analyze-something-that-ends-up-meaningless
  - Proper exploratory data analysis – graphing, looking at summary statistics, doing sanity checks on a lot of different hypotheses.

- What is the role of trial and error in data analysis? What is the the role of making a hypothesis before diving in?
- How can you deal with missing values in your data?
  - Replace with mean/median/mode or use linear regression for multiple imputation. (Linearly regress each variable against the others).

Predictive modeling

- What is the bias vs. variance tradeoff?
- What is the difference between supervised and unsupervised learning? What is an example of each?
  - Supervised learning has a target variable, unsupervised doesn't. Examples: linear regression vs. K-means clustering.
- How would you explain a linear regression to an engineer with no statistics background?
- What are some ways I can make my model more robust to outliers?
  - Answer on Quora
  - Use a model resistant to outliers (tree models over regression based models), use a more resilient error metric (mean absolute difference instead of mean squared error), cap data at a certain threshold, transform the data, remove outliers manually.
  - Can also project all points onto unit sphere in parameter space.
- 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?
  - Answer on Quora
  - Minimizing squared error finds the "mean" whereas minimizing absolute error finds the "median". The former is easier to compute and the latter is more resistant to outliers.
- What are the differences between the following pairs of terms: Type I error and Type II error, sensitivity and specificity, precision and recall?
- 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?
  - (Multinomial) log loss, which is equivalent to the cross entropy. You can also use sensitivity / specificity for a binary classifier; due to

class imbalance you can't just look at a single loss metric. Finally, one can use area under the ROC curve for a binary classifier.

- What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What's the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)
  - Answer on Quora
  - Logistic regression, especially when regularized, is fairly robust and can be used as a baseline estimate. Its output can also be interpreted as probabilities, unlike some other algorithms. You can also add interaction terms to model nonlinearity.
  - SVMs can take quite a long time to train, even if they work well in practice. Also, the standard formulation of SVMs is only applicable to binary classification (although one can use multiple SVM models for multinomial classification).
  - Ensemble tree-based models are a good "standard" nonlinear technique to apply (random forests and gradient boosted trees) which usually don't overfit too much; they handle high dimensionality, big datasets, and multinomial classification well. Gradient boosted trees perform marginally better than RFs but need more tuning and can be a little more prone to overfitting as a result.
  - Naive Bayes is easy to compute and used for cases where you have many binary features. It makes an assumption of independence – feature A being present provides no information about feature B being present and vice versa.
- What is overfitting and how would you detect if your model suffers from overfitting?
- What is regularization and where might it be helpful? What is an example of using regularization in a model?
  - In the most general sense, regularization is the addition of a term to a model's cost function which measures model complexity. It helps prevent overfitting (learning the background noise instead of the generalizable patterns); it increases model bias but decreases model variance.
  - In the context of linear regression, regularization refers to adding on a norm (usually L^1, L^2, or a mix) of a linear model's coefficients to the cost function.
- What's a hyperparameter? What are the hyperparamters for regularized linear regression?

  - A hyperparameter is a parameter of the model itself. Regularized linear regression has λ (penalization strength). (*Elastic net* regularization also has α, the degree of mixing between L^1 and L^2 norms.)

- Why might it be preferable to include fewer predictors over many? How do you select which predictors to use?
  - With too many predictors, the model will overfit and perform worse on test data. You can use regularization for feature selection. Alternatively, you can regress against a subset of the principal components or look at the correlation matrix for the features and remove those highly correlated with others.

- How can you determine which features are the most important in your model?
  - Look at magnitude of regression coefficients or the variable importance measure of a random forest model. Look also at which ones are the most highly correlated with the response variable.

- You have some variables which are all positively correlated with your target variable. In the linear regression, one of them has a negative coefficient. Does this make sense in light of the positive correlations? What is the interpretation of this result?
  - When accounting for the other variables, the negative coefficient variable has a marginal negative effect.

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

- If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse? Why?
  - You can fit a (cross-validated, properly tuned) regularized linear model with the models' predictions to the response variable. This will perform no worse than the best of the individual models and quite possibly better. It's known as stacking; different models have different strengths and deficiencies, so they compensate for one another.

- What's the difference between a decision tree and a decision forest?
  - A decision forest is an ensemble of many different decision trees, averaged together. In a random forest specifically, at each branch of each tree only a random subset of the predictors is used.

- How would you give different weights to points in a linear regression?
  - In the cost function to be minimized, which is the sum of squared errors, multiply each squared error by the corresponding weight.

Statistical inference

- How would you explain an A/B test to an engineer with no statistics background?
- How would you explain a confidence interval to an engineer with no statistics background? What about the meaning of 95% confidence?
- In an A/B test, how can you check if assignment to the various buckets was truly random?
- What might be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?
  - You can make sure that your A/B testing infrastructure works properly.
- What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?
- What would be some issues if blogs decide to cover one of your experimental groups?
- How would you conduct an A/B test on an opt-in feature?
- How would you run an A/B test for many variants, say 20 or more?
- How would you run an A/B test if the observations are extremely right-skewed?
- I have two different experiments that both change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?
- What is a p-value? What is the difference between type-1 and type-2 error?
- You are AirBnB and you want to test the hypothesis that a greater number of photographs increases the chances that a buyer selects the listing. How would you test this hypothesis?
- How would you design an experiment to determine the impact of latency on user engagement?
- What is maximum likelihood estimation? Could there be any case where it doesn't exist?
  - MLE is estimation of model parameters such that the likelihood of observing the training data is maximized. The maximum likelihood can be infinity. For example, in a Gaussian mixture model, if the center of a Gaussian sits directly on top of a data point, the likelihood can be driven arbitrarily high by making the width of the Gaussian arbitrarily small.

- What's the difference between a MAP, MOM, MLE estimator? In which cases would you want to use each?
- What is a confidence interval and how do you interpret it?
- What is unbiasedness as a property of an estimator? Is this always a desirable property when performing inference? What about in data analysis or predictive modeling?

Product-based questions

- Given training data on tweets and their retweets, how would you predict the number of retweets of a given tweet after 7 days after only observing 2 days worth of data?
- How could you collect and analyze data to use social media to predict the weather?
- How would you construct a feed to show relevant content for a site that involves user interactions with items?
- How would you design the "people you may know" feature on LinkedIn or Facebook?
- How would you predict who someone may want to send a Snapchat or Gmail to?
- How would you suggest to a franchise where to open a new store?
- In a search engine, given partial data on what the user has typed, how would you predict the user's eventual search query?
- Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?
- You're Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?
- You want to run a regression to predict the probability of a flight delay, but there are flights with delays of up to 12 hours that are really messing up your model. How can you address this?
- How could you use GPS data from a car to determine the quality of a driver?
- Given accelerometer, altitude, and fuel usage data from a car, how would you determine the optimum acceleration pattern to drive over hills?
- How would you quantify the influence of a Twitter user?

- You have 100 mathletes and 100 math problems. Each mathlete gets to choose 10 problems to solve. Given data on who got what problem correct, how would you rank the problems in terms of difficulty?
- You have 5000 people that rank 10 sushi plates in terms of saltiness. How would you aggregate this data to estimate the true saltiness rank in each sushi?
- Given data on congressional bills and which congressional representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most Democratic? Most Republican? Most bipartisan?
- How would you come up with an algorithm to detect plagiarism in online content?
- You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters to include?
- Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach this problem?
- What would be good metrics of success for an advertising-driven consumer product? (Buzzfeed, YouTube, Google Search, etc.) A service-driven consumer product? (Uber, Flickr, Venmo, etc.)
- What would be good metrics of success for a productivity tool? (Evernote, Asana, Google Docs, etc.) A MOOC? (edX, Coursera, Udacity, etc.)
- What would be good metrics of success for an e-commerce product? (Etsy, Groupon, Birchbox, etc.) A subscription product? (Netflix, Birchbox, Hulu, etc.) Premium subscriptions? (OKCupid, LinkedIn, Spotify, etc.)
- What would be good metrics of success for a consumer product that relies heavily on engagement and interac- tion? (Snapchat, Pinterest, Facebook, etc.) A messaging product? (GroupMe, Hangouts, Snapchat, etc.)
- What would be good metrics of success for a product that offered in-app purchases? (Zynga, Angry Birds, other gaming apps)
- A certain metric is violating your expectations by going down or up more than you expect. How would you try to identify the cause of the change?
- Growth for total number of tweets sent has been slow this month. What data would you look at to determine the cause of the problem?
- You're a restaurant and are approached by Groupon to run a deal. What data would you ask from them in order to determine whether or not to do the deal?

- You are tasked with improving the efficiency of a subway system. Where would you start?
- Say you are working on Facebook News Feed. What would be some metrics that you think are important? How would you make the news each person gets more relevant?
- How would you measure the impact that sponsored stories on Facebook News Feed have on user engagement? How would you determine the optimum balance between sponsored stories and organic content on a user's News Feed?
- You are on the data science team at Uber and you are asked to start thinking about surge pricing. What would be the objectives of such a product and how would you start looking into this?
- Say that you are Netflix. How would you determine what original series you should invest in and create?
- What kind of services would find churn (metric that tracks how many customers leave the service) helpful? How would you calculate churn?
- Let's say that you're are scheduling content for a content provider on television. How would you determine the best times to schedule content?
- You're a data scientist at Khan Academy. How would you use machine learning to predict a student's success on a subsequent question, based on knowing their performance on past questions they have answered?

Probability

- Bobo the amoeba has a 25%, 25%, and 50% chance of producing 0, 1, or 2 offspring, respectively. Each of Bobo's descendants also have the same probabilities. What is the probability that Bobo's lineage dies out?
  - Form the quadratic $(1/4) + (1/4)p + (1/2)p^2 = p$ and solve.
- In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?
  - $1 - (.8)^4$.
- How can you generate a random number between 1–7 with only a die?
  - Enumerate outcomes for rolling the die twice and categorize them as corresponding to $1, 2, \ldots, 7$. For the remainder outcomes (which aren't divisible by 7), reroll.

- How can you get a fair coin toss if someone hands you a coin that is weighted to come up heads more often than tails?
  - Do two flips. H/T and T/H have same probability; repeat if H/H or T/T.
- You have an 50–50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?
  - Add two normal distributions and look at the second derivative, which is strictly less than zero for $|\mu_1 - \mu_2| < 2\sigma$.
- Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?
  - Look at percentile in the distribution (z-score).
- A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?
  - The possibilities are FF, MF, and FM, so probability is $1/3$.
- You have a group of couples that decide to have children until they have their first girl, after which they stop having children. What is the expected number of children each couple will have? What is the expected gender ratio of the children that are born?
  - The expected number of children is $1(1/2) + 2(1/2)^2 + 3(1/2)^2 + \cdots$ which is an arithmetic-geometric sum. Letting S be the sum and r be the common ratio $(1/2)$, the general idea is that you form expressions for S and rS, subtract the latter from the former, and divide through by $1 - r$. The answer is 2 in this particular case, so the expected gender ratio is 1:1.
- How many ways can you split 12 people into 3 teams of 4?
  - It's $12!/(4!)^3 3!$, where the 3! accounts for team ordering and the 4! accounts for within-group ordering.
- You call 2 Ubers and 3 Lyfts. If the time that each takes to reach you is IID, what is the probability that all the Lyfts arrive first? What is the probability that all the Ubers arrive first?
  - There are 5! different orderings of the cars in total, 3!2! of which have the Lyfts arrive first, so the probability is $3!2!/5!$. Same for the Ubers because of symmetry in the problem (probability of Lyfts all arriving first equal to probability of Lyfts all arriving last).
- On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Alice and Bob randomly pick adjectives, what is the probability that they form a match?

  - Suppose we know which adjectives Alice chose. There are (5 choose 4) ways for Bob to choose 4 of Alice's adjectives and 19 ways for Bob to choose an adjective that Alice did not choose. There's also 1 way for Bob to choose all the same adjectives. Then 19 times (5 choose 4) plus 1 divided by # possibilities is the answer.

- A lazy high school senior types up application and envelopes to n different colleges, but puts the applications randomly into the envelopes. What is the expected number of applications that went to the right college?
  - Formally: "what's the expected number of fixed points in a random permutation for some n?" Use linearity of expectation. Let X_i be an indicator variable corresponding to whether or not i is a fixed point. Then the number of fixed points is $F = \sum_i X_i$ and $E[F] = E[X_1 + \cdots + X_i] = \sum_i E[X_i]$. Each $E[X_i]$ is equal to $1/n$ so the expected number of fixed points is 1 for all n.
- Let's say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?
  - Closer to the mean in general because of regression to the mean. See the conceptual examples on Wikipedia.
- What's the expected number of coin flips until you get two heads in a row? What's the expected number of coin flips until you get two tails in a row?
  - Quora has many solutions.
- Let's say we play a game where I keep flipping a coin until I get heads. If the first time I get heads is on the nth coin, then I pay you $2n - 1$ dollars. How much would you pay me to play this game?
- You have two coins, one of which is fair and comes up heads with a probability $1/2$, and the other which is biased and comes up heads with probability $3/4$. You randomly pick coin and flip it twice, and get heads both times. What is the probability that you picked the fair coin?
- You have a 0.1% chance of picking up a coin with both heads, and a 99.9% chance that you pick up a fair coin. You flip your coin and it comes up heads 10 times. What's the chance that you picked up the fair coin, given the information that you observed?
- What's the expected number of times you need to roll a 6-sided die until each number comes up at least once?
  - Look at expected *time* for the 1st, 2nd, ... number to show up to get $6/6 + 6/5 + \cdots + 6/1$.

Programming questions

Algorithms

- Write a function to calculate all possible assignment vectors of $2n$ users, where n users are assigned to group 0 (control) and n users are assigned to group 1 (treatment).
- Given a list of tweets, determine the top 10 most used hashtags.
  - For each tweet, split the string by spaces, extract substrings that begin with #, and convert them to lowercase.
- You have a stream of data coming in of size n, but you don't know what n is ahead of time. Write an algorithm that will take a random sample of k elements. Can you write one that takes $O(k)$ space?
  - This is reservoir sampling. Populate a list with the first k elements of the list. Then iterate through the rest; at the ith element, generate a random integer j between 1 and i inclusive; if $j \leq k$, then replace element j in the list of items with the ith item. Then the ith element of S is chosen to be included with probability k/i, and at each iteration the jth element is chosen to replaced with probability $(1/k) \cdot (k/i) = 1/i$.
- Write an algorithm to calculate the square root of a number.
  - Repeatedly average the estimate e with n/e.
- When can parallelism make your algorithms run faster? When could it make your algorithms run slower?
  - Answer on Quora
  - Parallelism works when you can subdivide your algorithm into a large number of independent parts. However, if each subdivision of the problem needs to communicate with the others, then parallelism will help much less. Combined with greater overhead (both computationally, from managing the different parallel threads, or on non-CPU hardware, if jumping from place to place in the hard drive slows the process down substantially), parallelism can make your algorithm run slower.
- Write functions to test if a string is a palindrome (the same both forwards and backwards) and to test if a string contains a palindrome as a substring. Do the latter in $O(n)$ time.
  - Existence of a palindrome equals existence of a palindrome of length 2 or 3, so loop through once checking for two consecutive identical

characters and then loop through another time with a "window" of 3 characters checking for palindromes.

- Suppose you have an even number of points in the 2D plane. Write a function to give the parameters of a line which divides these points into two equally sized groups.
  - I haven't done this myself, but maybe you can find the centroid of the points, pick an arbitrary direction radiating out from the centroid to be 0 degrees, and order the points according to the order in which an arc would "sweep out" the points; then look at the 1st and $(n/2+1)$th points and fiddle with the line a bit. This fails when you have collinearity in bad places relative to lines radiating outward from the centroid, so you'd have to pick a different direction for 0 degres.
- Given a list of integers, find 3 different integers in the list which sum to 0.
  - This is the 3SUM problem. Iterate through the list, storing each number in a hash table. Iterate through every pair of integers and check if the negative of their sum is in the hash table. This solution is in $O(n)$.
- Given a list of integers, find the *continuous subsequence* with the largest sum.
  - This is the maximum subarray problem and can be solved in $O(n)$ time. At each position simply compute the m

SQL

- What are the different types of joins? What are the differences between them?
- Why might a join on a subquery be slow? How might you speed it up?
- Describe the difference between primary keys and foreign keys in a SQL database.
  - Microsoft MSDN reference
  - "A table typically has a column or combination of columns that contain values that uniquely identify each row in the table. This column, or columns, is called the primary key (PK) of the table and enforces the entity integrity of the table. Because primary key constraints guarantee unique data, they are frequently defined on an identity column."
  - "A foreign key (FK) is a column or combination of columns that is used to establish and enforce a link between the data in two tables to control the data that can be stored in the foreign key table. In a foreign key reference, a link is created between two tables when the

column or columns that hold the primary key value for one table are referenced by the column or columns in another table. This column becomes a foreign key in the second table."

- Given a `COURSES` table with columns `course_id` and `course_name`, a `FACULTY` table with columns `faculty_id` and `faculty_name`, and a `COURSE_FACULTY` table with columns `faculty_id` and `course_id`, how would you return a list of faculty who teach a course given the name of a course?
  - SELECT faculty_name WHERE course_name = "whatever" FROM COURSES INNER JOIN COURSE_FACULTY ON course_id INNER JOIN FACULTY ON faculty_id
- Given an `IMPRESSIONS` table with `ad_id`, `click` (an indicator that the ad was clicked), and `date`, write a SQL query that will tell me the clickthrough rate of each ad by month.
- Write a query that returns the name of each department and a count of the number of employees in each:

  `EMPLOYEES` containing: `Emp_ID` (Primary key) and `Emp_Name`

  `EMPLOYEE_DEPT` containing: `Emp_ID` (Foreign key) and `Dept_ID` (Foreign key)

  `DEPTS` containing: `Dept_ID` (Primary key) and `Dept_Name`

Fermi estimates

- How many McDonalds are there in the US?
- How many piano tuners are in Seattle?
- How many baseballs could you fit in a football stadium?

Communication

- Tell me about a project that you've worked on with others. What did you add to the group?
- 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 it to me?
- How would you convince a government agency to release their data in a publicly accessible API?

- I'm a local business owner operating a small restaurant. Convince me to switch my advertising budget from print to Internet.