Machine Learning & Mathematics

1. What is cross-validation? How to do it right?
2. Is it better to design robust or accurate algorithms?
3. How to define/select metrics?
4. Explain what regularization is and why it is useful. What are the benefits and drawbacks of specific methods, such as ridge regression and lasso?
5. Explain what a local optimum is and why it is important in a specific context, such as K-means clustering. What are specific ways of determining if you have a local optimum problem? What can be done to avoid local optima?
6. Assume you need to generate a predictive model using multiple regression. Explain how you intend to validate this model
7. Explain what precision and recall are. How do they relate to the ROC curve?
8. What is latent semantic indexing? What is it used for? What are the specific limitations of the method?
9. Explain what resampling methods are and why they are useful
10. What is principal component analysis? Explain the sort of problems you would use PCA for. Also explain its limitations as a method
11. Explain what a false positive and a false negative are. Why is it important these from each other? Provide examples when false positives are more important than false negatives, false negatives are more important than false positives and when these two types of errors are equally important
12. What is the difference between supervised learning and unsupervised learning? Give concrete examples
13. What does NLP stand for?
14. What are feature vectors?
15. When would you use random forests Vs SVM and why?
16. How do you take millions of users with 100's transactions each, amongst 10k's of products and group the users together in meaningful segments?
17. How do you know if one algorithm is better than other?
18. How do you test whether a new credit risk scoring model works?
19. What is: collaborative filtering, n-grams, cosine distance?
20. What is better: good data or good models? And how do you define "good"? Is there a universal good model? Are there any models that are definitely not so good?
21. Why is naive Bayes so bad? How would you improve a spam detection algorithm that uses naive Bayes?
22. What are the drawbacks of linear model? Are you familiar with alternatives (Lasso, ridge regression, boosted trees)?
23. Do you think 50 small decision trees are better than a large one? Why?

24. Why is mean square error a bad measure of model performance? What would you suggest instead?
25. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything? Are you familiar with A/B testing?
26. What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?
27. Do you know / used data reduction techniques other than PCA? What do you think of step-wise regression? What kind of step-wise techniques are you familiar with?
28. How would you define and measure the predictive power of a metric?
29. Do we always need the intercept term in a regression model?
30. What are the assumptions required for linear regression? What if some of these assumptions are violated?
31. What is collinearity and what to do with it? How to remove multicollinearity?
32. How to check if the regression model fits the data well?
33. What is a decision tree?
34. What impurity measures do you know?
35. What is random forest? Why is it good?
36. How do we train a logistic regression model? How do we interpret its coefficients?
37. What is the maximal margin classifier? How this margin can be achieved and why is it beneficial? How do we train SVM?
38. What is a kernel? Explain the kernel trick
39. Which kernels do you know? How to choose a kernel?
40. Is it beneficial to perform dimensionality reduction before fitting an SVM? Why or why not?
41. (What is an Artificial Neural Network?) What is back propagation?
42. What is curse of dimensionality? How does it affect distance and similarity measures?
43. What is Ax=b? How to solve it?
44. How do we multiply matrices?
45. What is singular value decomposition? What is an eigenvalue? And what is an eigenvector?
46. What's the relationship between PCA and SVD?
47. Can you derive the ordinary least square regression formula?
48. What is the difference between a convex function and non-convex?
49. What is gradient descent method? Will gradient descent methods always converge to the same point?
50. What the Newton's method is?
51. Imagine you have N pieces of rope in a bucket. You reach in and grab one end-piece, then reach in and grab another end-piece, and tie those two together. What is the expected value of the number of loops in the bucket?

Statistics

1. How do you assess the statistical significance of an insight?
2. Explain what a long-tailed distribution is and provide three examples of relevant phenomena that have long tails. Why are they important in classification and regression problems?
3. What is the Central Limit Theorem? Explain it. Why is it important?
4. What is statistical power?
5. Explain selection bias (with regard to a dataset, not variable selection). Why is it important? How can data management procedures such as missing data handling make it worse?
6. Provide a simple example of how an experimental design can help answer a question about behavior. How does experimental data contrast with observational data?
7. Is mean imputation of missing data acceptable practice? Why or why not?
8. What is an outlier? Explain how you might screen for outliers and what would you do if you found them in your dataset. Also, explain what an inlier is and how you might screen for them and what would you do if you found them in your dataset
9. How do you handle missing data? What imputation techniques do you recommend?
10. You have data on the durations of calls to a call center. Generate a plan for how you would code and analyze these data. Explain a plausible scenario for what the distribution of these durations might look like. How could you test, even graphically, whether your expectations are borne out?
11. Explain likely differences between administrative datasets and datasets gathered from experimental studies. What are likely problems encountered with administrative data? How do experimental methods help alleviate these problems? What problem do they bring?
12. You are compiling a report for user content uploaded every month and notice a spike in uploads in October. In particular, a spike in picture uploads. What might you think is the cause of this, and how would you test it?
13. You're about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle?
14. There's one box - has 12 black and 12 red cards, 2nd box has 24 black and 24 red; if you want to draw 2 cards at random from one of the 2 boxes, which box has the higher probability of getting the same color? Can you tell intuitively why the 2nd box has a higher probability
15. What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?
16. Define: quality assurance, six sigma.
17. Give examples of data that does not have a Gaussian distribution, nor log-normal.

18. What is root cause analysis? How to identify a cause vs. a correlation? Give examples
19. Give an example where the median is a better measure than the mean
20. Given two fair dices, what is the probability of getting scores that sum to 4? to 8?
21. What is the Law of Large Numbers?
22. How do you calculate needed sample size?
23. When you sample, what bias are you inflicting?
24. How do you control for biases?
25. What are confounding variables?
26. What is A/B testing?
27. An HIV test has a sensitivity of 99.7% and a specificity of 98.5%. A subject from a population of prevalence 0.1% receives a positive test result. What is the precision of the test (i.e the probability he is HIV positive)?
28. Infection rates at a hospital above a 1 infection per 100 person days at risk are considered high. An hospital had 10 infections over the last 1787 person days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard
29. You roll a biased coin (p(head)=0.8) five times. What's the probability of getting three or more heads?
30. A random variable X is normal with mean 1020 and standard deviation 50. Calculate P(X>1200)
31. Consider the number of people that show up at a bus station is Poisson with mean 2.5/h. What is the probability that at most three people show up in a four hour period?
32. You are running for office and your pollster polled hundred people. Sixty of them claimed they will vote for you. Can you relax?
33. Geiger counter records 100 radioactive decays in 5 minutes. Find an approximate 95% interval for the number of decays per hour.
34. The homicide rate in Scotland fell last year to 99 from 115 the year before. Is this reported change really networthy?
35. Consider influenza epidemics for two parent heterosexual families. Suppose that the probability is 17% that at least one of the parents has contracted the disease. The probability that the father has contracted influenza is 12% while the probability that both the mother and father have contracted the disease is 6%. What is the probability that the mother has contracted influenza?
36. Suppose that diastolic blood pressures (DBPs) for men aged 35-44 are normally distributed with a mean of 80 (mm Hg) and a standard deviation of 10. About what is the probability that a random 35-44 year old has a DBP less than 70?
37. In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?
38. A diet pill is given to 9 subjects over six weeks. The average difference in weight (follow up - baseline) is -2 pounds. What would the standard deviation of the

difference in weight have to be for the upper endpoint of the 95% T confidence interval to touch 0?

39. In a study of emergency room waiting times, investigators consider a new and the standard triage systems. To test the systems, administrators selected 20 nights and randomly assigned the new triage system to be used on 10 nights and the standard system on the remaining 10 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 3 hours with a variance of 0.60 while the average MWT for the old system was 5 hours with a variance of 0.68. Consider the 95% confidence interval estimate for the differences of the mean MWT associated with the new system. Assume a constant variance. What is the interval? Subtract in this order (New System - Old System).
40. To further test the hospital triage system, administrators selected 200 nights and randomly assigned a new triage system to be used on 100 nights and a standard system on the remaining 100 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 4 hours with a standard deviation of 0.5 hours while the average MWT for the old system was 6 hours with a standard deviation of 2 hours. Consider the hypothesis of a decrease in the mean MWT associated with the new treatment. What does the 95% independent group confidence interval with unequal variances suggest vis a vis this hypothesis? (Because there's so many observations per group, just use the Z quantile instead of the T.)

Process & Miscellaneous

1. How to optimize algorithms? (parallel processing and/or faster algorithms). Provide examples for both
2. Examples of NoSQL architecture
3. Provide examples of machine-to-machine communications
4. Compare R and Python
5. Is it better to have 100 small hash tables or one big hash table, in memory, in terms of access speed (assuming both fit within RAM)? What do you think about in-database analytics?
6. What is star schema? Lookup tables?
7. What is the life cycle of a data science project ?
8. How to efficiently scrape web data, or collect tons of tweets?
9. How to clean data?
10. How frequently an algorithm must be updated?
11. What is POC (proof of concept)?
12. Explain Tufte's concept of "chart junk"
13. How would you come up with a solution to identify plagiarism?

14. How to detect individual paid accounts shared by multiple users?
15. Is it better to spend 5 days developing a 90% accurate solution, or 10 days for 100% accuracy? Depends on the context?
16. What is your definition of big data?
17. Explain the difference between “long” and “wide” format data. Why would you use one or the other?
18. Do you know a few “rules of thumb” used in statistical or computer science? Or in business analytics?
19. Name a few famous API’s (for instance GoogleSearch)
20. Give examples of bad and good visualizations