Some general software notes

Nicholas P. Ross

January 31, 2020

Abstract

This is a simple document that discusses the basis and basics of Data Science.

Contents

1 Algorithms

Heap's algorithm

Heap's algorithm

Steinhaus Johnson Trotter algorithm

Steinhaus Johnson Trotter algorithm

2 Books

Introduction to Algorithms Paperback, T Cormen
Clean Code: A Handbook of Agile Software Craftsmanship (Robert C. Martin)
The Robert C. Martin Clean Code Collection (Collection) (Robert C. Martin Series)
The Pragmatic Programmer Paperback, by Andrew Hunt

3 Top 7 Machine Learning Github Repositories for Data Scientists

Top 7 ML GitHub repos/

4 Why Data Science Matters

https://medium.com/sequoia-capital/why-data-science-matters-ee583f785a55

5 16 Useful Advices for Aspiring Data Scientists

https://towardsdatascience.com/16-useful-advices-for-aspiring-data-scientists-6da9afa8c72c

6 What are the top 30 most essential algorithms you must know for competitive programming?

What are the top 30 most essential algorithms you must know for competitive programming?

1. General techniques: greedy algorithms, dynamic programming - dozens of techniques, divide-and-conquer (binary search and other), sorting algorithms - especially merge sort, heap sort and quick sort, partial sort/top-k elements in [expected] linear time.

2. Data structures: arrays, dynamic arrays, stacks, queues, deques, trees, heaps, hash functions and hash tables including rolling hash/polynomial hash for strings/substrings, binary search trees (treap or splay, including the one with implicit keys), segment trees/Fenwick tree, disjoint set union, sqrt-decomposition.

3. Graphs: exploration - DFS/BFS, shortest paths - **Dijkstra**, Bellman-Ford, Floyd-Warshall, spanning trees - **Prims**, Kruskals, flows - Ford-Fulkerson, Edmonds-Karp, min-cost-max-flow, topological sort and strongly connected components, 2-SAT, LCA, bridges and articulation points, eulerian cycle, biconnected graphs - 2-coloring, matching.

4. Strings: KMP, Z-function, polynomial rolling hash, suffix array, suffix tree or suffix automaton.

5. Algebra/Number Theory: Euclids algorithm, fast exponentiation, sieve of Eratosthenes, inverse modulo prime number, solving ax+by=c and alike Diophantine equations, Eulers function computation, Fast Fourier Transform, Gausss algorithm for matrix inversion/solving system of linear equations, Chinese theorem about remainders.

6. Geometry: intersecting lines, segments, circles, line and polygon, tangents and common tangents, moving points, lines and segments, sorting by angle, computing lengths and areas, convex hull, scanning line algorithms, fast point inside polygon, finding two closest/farthest points, covering circle.

7. Various: number of combinations, Catalans numbers, inclusion-exclusion formula, Burnsides lemma, Stirling numbers, Grundy numbers for games.

(Also::
Dijkstras
Prims
(B/D)FS
Sieve of Eratshtenos
Binary heaps
Mergesort
“and that’s pretty much it”)

7 Glossary (of sorts)

Kernel trick

The “kernel trick”, or “kernel methods” are a class of algorithms for pattern analysis, whose best known member is the support vector machine (SVM). The general task of pattern analysis is to find and study general types of relations (for example clusters, rankings, principal components, correlations, classifications) in datasets. In its simplest form, the kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data.

See also: Understanding the kernel trick.

8 Top firms 100 Data Science interview questions

www.kdnuggets.com/2017/03/top-firms-100-data-science-interview-questions, by Brendan Martin.

A fresh scrape from Glassdoor gives us a good idea about what applicants are asked during a data scientist interview at some of the top companies. Unfortunately for us, almost every company has their interviewees sign NDAs. Since Glassdoor allows anonymity, a few brave souls have given us some fantastic examples of what they were asked during the interview process at top companies like Facebook, Google, and Microsoft.

If you find yourself unable to answer some of the questions below, consider checking out a course or a book on the subject.

If youd like to share your answer(s) to any of the questions, leave a comment and Ill add the top ones to the post. Just make sure to comment with your real name so I can give you credit!

Also, if you dont see a particular question on this list that youve been asked, or you know of one thats asked a lot, comment below. Id love to add it.

8.1 General Questions

Apple::
Suppose youre given millions of users that each have hundreds of transactions and these millions of transactions are for tens of thousands of products. How would you group the users together in meaningful segments?

Microsoft::
Describe a project youve worked on and how it made a difference.
How would you approach a categorical feature with high-cardinality?
What would you do to summarize a Twitter feed?
What are the steps for wrangling and cleaning data before applying machine learning algorithms?
How do you measure distance between data points?
Define variance.
Describe the differences between and use cases for box plots and histograms.

Twitter::
What features would you use to build a recommendation algorithm for users?

Uber::

Pick any product or app that you really like and describe how you would improve it.
How would you find an anomaly in a distribution ?
How would you go about investigating if a certain trend in a distribution is due to an anomaly?
How would you estimate the impact Uber has on traffic and driving conditions?
What metrics would you consider using to track if Ubers paid advertising strategy to acquire new customers actually works? How would you then approach figuring out an ideal customer acquisition cost?

LinkedIn::

Big Data Engineer Can you explain what REST is?

8.2 Machine Learning Questions

Google::

Why do you use feature selection?
What is the effect on the coefficients of logistic regression if two predictors are highly correlated? What are the confidence intervals of the coefficients?
Whats the difference between Gaussian Mixture Model and K-Means?
How do you pick k for K-Means?
How do you know when Gaussian Mixture Model is applicable?
Assuming a clustering models labels are known, how do you evaluate the performance of the model?

Microsoft::

Whats an example of a machine learning project youre proud of?
Choose any machine learning algorithm and describe it.
Describe how Gradient Boosting works.
Data Mining Describe the decision tree model.
Data Mining What is a neural network?
Explain the Bias-Variance Tradeoff
How do you deal with unbalanced binary classification?
Whats the difference between L1 and L2 regularization?

Uber::

What sort features could you give an Uber driver to predict if they will

accept a ride request or not?
What supervised learning algorithm would you use to solve the problem and how would compare the results of the algorithm?

LinkedIn::

Name and describe three different kernel functions and in what situation you would use each.
Describe a method used in machine learning.
How do you deal with sparse data?

IBM::

How do you prevent overfitting?
How do you deal with outliers in your data?
How do you analyze the performance of the predictions generated by regression models versus classification models?
How do you assess logistic regression versus simple linear regression models?
Whats the difference between supervised learning and unsupervised learning?
What is cross-validation and why would you use it?
Whats the name of the matrix used to evaluate predictive models?
What relationships exist between a logistic regressions coefficient and the Odds Ratio?
Whats the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)
If you had a categorical dependent variable and a mixture of categorical and continuous independent variables, what algorithms, methods, or tools would you use for analysis?
Business Analytics Whats the difference between logistic and linear regression? How do you avoid local minima?

Salesforce::

What data and models would would you use to measure attrition/churn? How would you measure the performance of your models?
Explain a machine learning algorithm as if youre talking to a non-technical person.

Capital One::

How would you build a model to predict credit card fraud?
How do you handle missing or bad data?
How would you derive new features from features that already exist?

If youre attempting to predict a customers gender, and you only have 100 data points, what problems could arise?
Suppose you were given two years of transaction history. What features would you use to predict credit risk?
Design an AI program for Tic-tac-toe

Zillow::
Explain overfitting and what steps you can take to prevent it.
Why does SVM need to maximize the margin between support vectors?

8.3 Hadoop

Twitter::
How would you use Map/Reduce to split a very large graph into smaller pieces and parallelize the computation of edges according to the fast/dynamic change of data?
Data Engineer Given a list of followers in the format:123, 345234, 678345, 123Where column one is the ID of the follower and column two is the ID of the followee. Find all mutual following pairs (the pair 123, 345 in the example above). How would you use Map/Reduce to solve the problem when the list does not fit in memory?

Capital One::
Data Engineer What is Hadoop serialization?
Explain a simple Map/Reduce problem.

8.4 Hive

LinkedIn::
Data Engineer Write a Hive UDF that returns a sentiment score. For example, if good = 1, bad = -1, and average = 0, then a review of a restaurant states Good food, bad service, your score might be 1 1 = 0.

8.5 Spark

Capital One::
Data Engineer Explain how RDDs work with Scala in Spark

8.6 Statistics & Probability Questions

Google::

Explain Cross-validation as if youre talking to a non-technical person.
Describe a non-normal probability distribution and how to apply it.

Microsoft::

Data Mining Explain what heteroskedasticity is and how to solve it

Twitter::

Given Twitter user data, how would you measure engagement?

Uber::

What are some different Time Series forecasting techniques?
Explain Principle Component Analysis (PCA) and equations PCA uses.
How do you solve Multicollinearity?
Analyst Write an equation that would optimize the ad spend between Twitter and Facebook.

Facebook::

Whats the probability youll draw two cards of the same suite from a single deck?

IBM::

What are p-values and confidence intervals?

Capital One::

Data Analyst If you have 70 red marbles, and the ratio of green to red marbles is 2 to 7, how many green marbles are there?
What would the distribution of daily commutes in New York City look like?
Given a die, would it be more likely to get a single 6 in six rolls, at least two 6s in twelve rolls, or at least one-hundred 6s in six-hundred rolls?

PayPal::

Whats the Central Limit Theorem, and how do you prove it? What are its applications?

8.7 Programming & Algorithms

Google::
Data Analyst Write a program that can determine the height of an arbitrary binary tree

Microsoft::
Create a function that checks if a word is a palindrome.

Twitter:: Build a power set.
How do you find the median of a very large dataset?

Uber::
Data Engineer Code a function that calculates the square root (2-point precision) of a given number. Follow up: Avoid redundant calculations by now optimizing your function with a caching mechanism.

Facebook::
Suppose youre given two binary strings, write a function adds them together without using any builtin string-to-int conversion or parsing tools. For example, if you give your function binary strings 100 and 111, it should return 1011. Whats the space and time complexity of your solution?
Write a function that accepts two already sorted lists and returns their union in a sorted list.

LinkedIn::
Data Engineer Write some code that will determine if brackets in a string are balanced
How do you find the second largest element in a Binary Search Tree?
Write a function that takes two sorted vectors and returns a single sorted vector.
If you have an incoming stream of numbers, how would you find the most frequent numbers on-the-fly?
Write a function that raises one number to another number, i.e. the pow() function.
Split a large string into valid words and store them in a dictionary. If the string cannot be split, return false. Whats your solutions complexity?

Salesforce::
Whats the computational complexity of finding a documents most frequently

used words?
If youre given 10 TBs of unstructured customer data, how would you go about finding extracting valuable information from it?

Capital One::
Data Engineer How would you disjoin two arrays (like JOIN for SQL, but the opposite)?
Create a function that does addition where the numbers are represented as two linked lists.
Create a function that calculates matrix sums.
How would you use Python to read a very large tab-delimited file of numbers to count the frequency of each number?

PayPal::
Write a function that takes a sentence and prints out the same sentence with each word backwards in $O(n)$ time.
Write a function that takes an array, splits the array into every possible set of two arrays, and prints out the max differences between the two arrays minima in $O(n)$ time.
Write a program that does merge sort.

8.8 SQL

Microsoft::
Data Analyst Define and explain the differences between clustered and non-clustered indexes.
Data Analyst What are the different ways to return the rowcount of a table?

Facebook::
Data Engineer If youre given a raw data table, how would perform ETL (Extract, Transform, Load) with SQL to obtain the data in a desired format?
How would you write a SQL query to compute a frequency table of a certain attribute involving two joins? What changes would you need to make if you want to ORDER BY or GROUP BY some attribute? What would you do to account for NULLS?

LinkedIn::
Data Engineer How would you improve ETL (Extract, Transform, Load)

throughput?

8.9 Brain Teasers & Word Problems

Google::
Suppose you have ten bags of marbles with ten marbles in each bag. If one bag weighs differently than the other bags, and you could only perform a single weighing, how would you figure out which one is different?

Facebook::
You are about to hop on a plane to Seattle and want to know if you should carry an umbrella. You call three friends of yours that live in Seattle and ask each, independently, if its raining. Each of your friends will tell you the truth $\frac{2}{3}$ of the time and mess with you by lying $\frac{1}{3}$ of the time. If all three friends answer "Yes, its raining", what is the probability that is it actually raining in Seattle?

Uber::
Imagine you are working with a hospital. Patients arrive at the hospital in a Poisson Distribution, and the doctors attend to the patients in a Uniform Distribution. Write a function or code block that outputs the patients average wait time and total number of patients that are attended to by doctors on a random day.

Facebook::
Imagine there are three ants in each corner of an equilateral triangle, and each ant randomly picks a direction and starts traversing the edge of the triangle. Whats the probability that none of the ants collide? What about if there are N ants sitting in N corners of an equilateral polygon?
How many trailing zeros are in 100 factorial (i.e. 100!)?

LinkedIn::
Imagine youre climbing a staircase that contains n stairs, and you can take any number k steps. How many distinct ways can you reach the top of the staircase? (This is a modification of the original stair step problem)

9 Planet and Orbital Insight

https://pypi.org/project/planet/
https://orbitalinsight.com/
https://medium.com/analytics-vidhya/satellite-imagery-analysis-with-python-3f8ccf8a7c32

10 References

− 20 articles about core datascience
http://www.kdnuggets.com/2017/03/top-firms-100-data-science-interview-questions.html