CSYE 7245–
Big-Data Systems and Intelligence Analytics
Sample Exam and Solutions Two

Q1 (5 Points) Arrange the following functions in increasing order of asymptotic growth:

- log n
- √n
- $5n^5$
- n log n
- $n^{0.1}$

Solution:

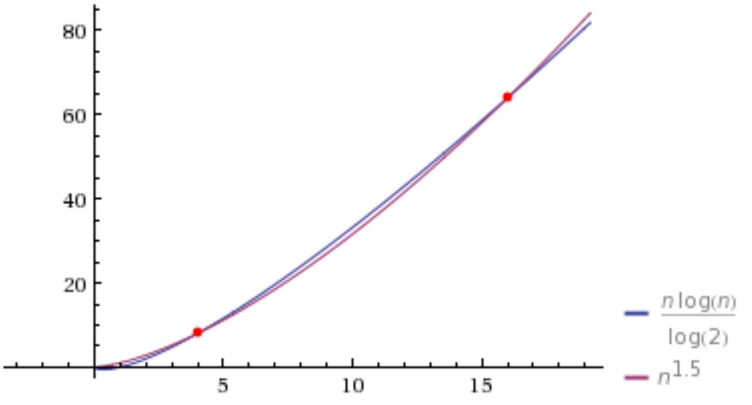
- log n, $n^{0.1}$, √n, n log n, $5n^5$

See http://en.wikipedia.org/wiki/Big_O_notation#Orders_of_common_functions

Q2 (5 Points) Is n log(n) big-O of $n^{1.5}$? Prove your answer. The intersections of the two functions is

$n = 4$ and $n = e^{-2\, W_{-1}\left(-\frac{\log(2)}{2}\right)} \approx 16.0000$

Solution:



No. A function f is O(g) iff there exists a C and N such that $f(x) \leq C\ |g(x)|$ for all $x \geq N$. In the case of proving big-O, you need to find the C and N. We can see after an N=16 and a C=1 the x of $n^{1.5} \geq n \log(n)$ as the n goes to infinity.

Q3 (5 Points) Master Theorem: For the following recurrences, give an expression for the runtime T(n) if the recurrence can be solved with the Master Theorem. Otherwise, indicate that the Master Theorem does not apply.

A. $T(n) = 2T(n/2) + c$
B. $T(n) = 2T(n/2) + n$

Solution:

$T(n) = 2T(n/2) + c$

a=b=2; k=log(2) log(2)=1 f(n) = n which is $\theta(1)$
0<1 so (CASE 1)
In Case 1 $\theta(n^k)$ or $\theta(n^{\log_b a})$ or $\theta(n^1)$

Therefore the runtime T(n) is $\theta(n)$

$T(n) = 2T(n/2) + n$

a=b=2; k=log(2) log(2)=1 f(n) = n 1=1 so (CASE 2)
$f(n) = \Theta(n) = \Theta(n^k \log^p n) = \Theta(n^1 \log^0 n)$ so p=1

$\Theta(n^1 \log^{0+1} n)$ which is $\Theta(n \log n)$ Note: This is the recurrence for MergeSort.

Q4 (5 Points) Is 3-SAT in NP? If so prove it.

Solution:

2 points for yes.
3 points for a polynomial-time certificate.
For 3-CNF with n clauses, given a certificate of an assignment 0s, 1s, is easily verified in poly time by keeping track of assignments given to any literals and check the n clauses for a truth assignment and whether the literals are consistent.

Q5 (5 Points) In information theory, the Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. In another way, it measures the minimum number of substitutions required to change one string into the other. For example, 101 and 111 have a Hamming distance of 1 (1 bit needs to be flipped to make them equal). Is the Hamming distance an admissible heuristic for A* search? Why or why not?

Solution:

Yes. You must mention something about the Hamming distance needing to be a metric space to be an admissible heuristic for A* search. For full credit you need to argue all of the properties of a distance function hold for the Hamming distance.

Distance function d(x,y):
$d(x,y) \geq 0$ (non-negative),
$d(x,y) = 0$ iff $x=y$ (identity of indiscernibles),
$d(x,y) = d(y,x)$ (Symmetric),
$d(x,z) \leq d(x,y) + d(y,z)$ (Triangle inequality)

Q6 (10 Points)

Suppose that you are given a collection of equations and inequalities. An equation or inequality is always between two variables such as $x1 = x2$ or $x2 \neq x3$. We want to find a solution to the collection of equations and inequalities if there is one. Develop an algorithm to solve this problem whose time complexity is better than O(n log n), where n is the number of equations and inequalities. Compute the algorithms time complexity.

Solution:

Algorithm:

- Consider every variable in the given set of equations and inequalities as a vertex of a graph with its parent pointing to self. Also every vertex is assigned a unique value initially.
- For every equation in the given input run Union, so that the variable which have equal value, can be unified into a same subset (or a tree). This step ensures the fact that variable with same values will have a common root.
- For any given inequality, try to find the parents of the two variables that are involved in the inequality. If both the variables have same parent, then no solution is possible as the given inequality contradicts with the equation specified before.
- If in step 3, we don't encounter any contradiction then for all v ε V, find its parent and assign the value of parent to v because of the condition that all the elements in a given tree have same value. Print the values of all the variables giving us the solution for the given set of equations and inequalities

Complexity: Since union and find operations take time O (log*V), the overall running time is O (V + n log*V) where n is the number of equations and inequalities and V is the number of variables. As we know that the value of log*V is small even for large values of n as iterated logarithm grows slower than logarithmic function, the time complexity of this algorithm is better than O (n logn).

Q7 (5 Points) What is the probability of getting exactly 2 heads after flipping three coins?

Solution:

If one is head and 0 tails:
0 0 0
0 0 1
0 1 0
0 1 1
1 0 0
1 0 1
1 1 0
1 1 1

There are three events in getting exactly 2 heads
{0 1 1, 1 0 1, 1 1 0}
So 3/8 is the probability of getting exactly 2 heads after flipping three coins.

We can also the formula for a discrete random variable based on a binomial distribution:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

X = 1, with probability $\binom{3}{2}(\frac{1}{2})^1(\frac{1}{2})^2 = \frac{3}{8}$

Note: 3 choose 2 is 3. The 3/8 comes from 3* 1/2^3

Q8 (5 Points)

What is a cost function in a Neural Network? Give two examples.

Solution:

A cost function is a measure of "how good" a neural network did with respect to it's given training sample and the expected output.

mean_squared_error, and categorical_crossentropy are examples.

Q9 (5 Points)

What is a gradient descent?

Solution:

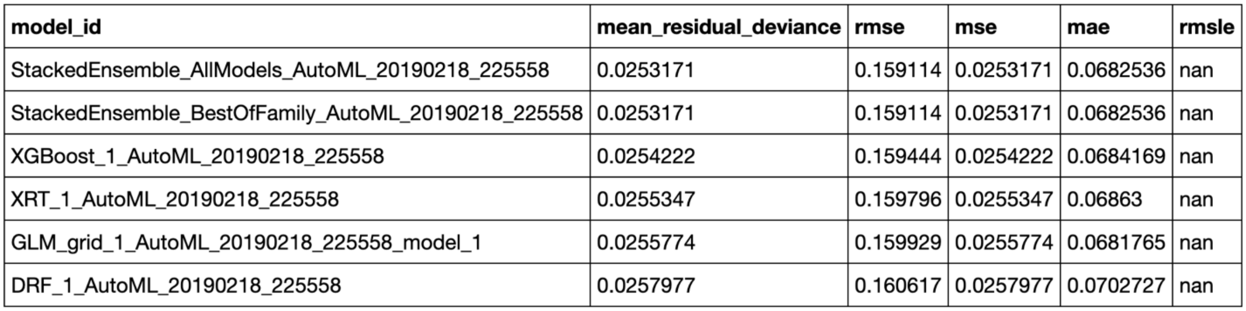
Gradient descent is an optimization algorithm, which is used to learn the value of parameters that minimizes the cost function. It is an iterative algorithm which moves in the direction of steepest descent as defined by the negative of the gradient.

Q10 (5 Points)
What Is the role of an Activation Function?

Solution:

The goal of an activation function is to introduce nonlinearity into the neural network so that it can learn nonlinear functions.

Q11 (5 Points)
The following leaderboard was generated by H2O AI. Was the analysis classification or regression?



model_id	mean_residual_deviance	rmse	mse	mae	rmsle
StackedEnsemble_AllModels_AutoML_20190218_225558	0.0253171	0.159114	0.0253171	0.0682536	nan
StackedEnsemble_BestOfFamily_AutoML_20190218_225558	0.0253171	0.159114	0.0253171	0.0682536	nan
XGBoost_1_AutoML_20190218_225558	0.0254222	0.159444	0.0254222	0.0684169	nan
XRT_1_AutoML_20190218_225558	0.0255347	0.159796	0.0255347	0.06863	nan
GLM_grid_1_AutoML_20190218_225558_model_1	0.0255774	0.159929	0.0255774	0.0681765	nan
DRF_1_AutoML_20190218_225558	0.0257977	0.160617	0.0257977	0.0702727	nan

Solution:

Regression as **mean_residual_deviance** is being used to rank the leaderboard.

Q12 (5 Points)
I want to show variable importance plots from the models in the leaderboard in Q11. Which models should I select?

Solution:

Any of the tree based models. XGBoost, XRT and DRF.

Q13 (5 Points)
We are trying to predict bad_loan an integer with values 0 or 1. How to we tell H2O that we want to perform classification rather than regression on this dependent variable?

Solution:

Convert it to an enum.

```
df["bad_loan"] = df["bad_loan"].asfactor()
```

Q14 (5 Points)

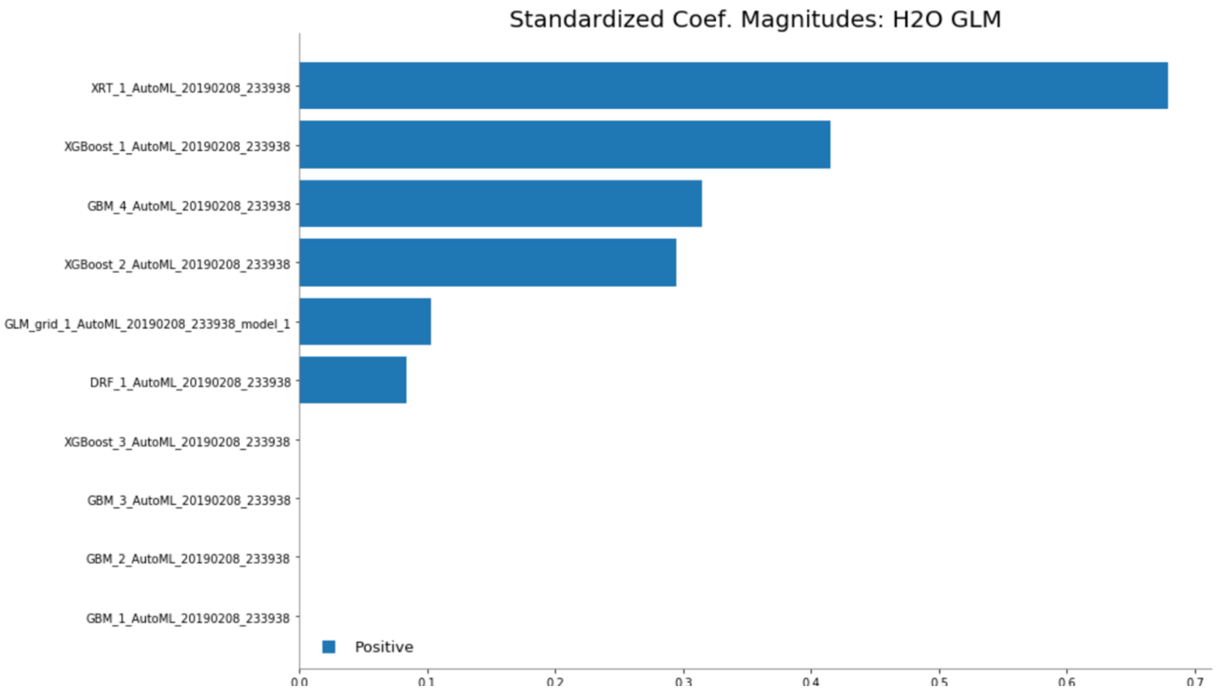
There are several of the same algorithms like XGBoost, GBM, etc resulting in different models. How to models using the same algorithm differ? For example, GBM_1_AutoML_20190208_233938 and GBM_2_AutoML_20190208_233938 so that they give different AUC values?

Solution:

The algorithms differ in the hyperparamters used.

Q15 (5 Points)

Below are the standardized coefficient magnitudes for a stacked ensemble. Are all of the base models used to make the stacked ensembles predictions?



Solution:

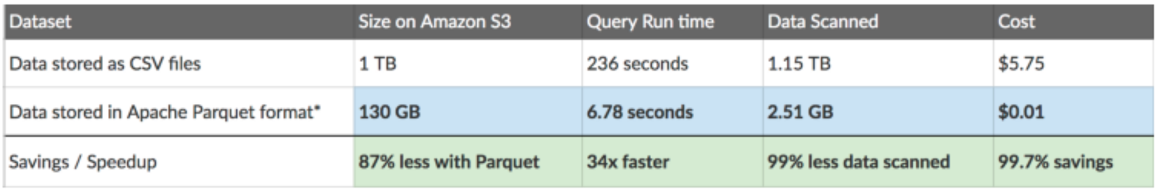
No the models with zero weight are not used.

Q16 (5 Points)

What are the benefits of parquet vs CSV in Apache Spark?

Solution:

Like a **CSV** file, parquet is a file format. The difference is that **Parquet** is designed as a columnar storage format to support complex data processing. ... Apache **Parquet** is column-oriented and designed to bring efficient columnar storage of data compared to row based files like **CSV**
https://dzone.com/articles/how-to-be-a-hero-with-powerful-parquet-google-and



Dataset	Size on Amazon S3	Query Run time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	$5.75
Data stored in Apache Parquet format*	130 GB	6.78 seconds	2.51 GB	$0.01
Savings / Speedup	87% less with Parquet	34x faster	99% less data scanned	99.7% savings

Q17 (5 Points)

You have a dataframe called 'Browns_Lectures' which has 2 columns: 'Date' and 'Professor'. Write a query to calculate how many lectures were taken by Professor Brown in the last 20 days. Assume that today's date is 20th March 2019.

Note: The dataset has data for last 2 years from April 201

Date	Professor
2017-04-12 08:00:00	Nik Bear Brown
2017-04-12 16:00:00	Nik Bear Brown
2017-04-12 20:00:00	Nik Bear Brown
2017-04-13 20:00:00	Nik Bear Brown
2017-04-14 20:00:00	Nik Bear Brown
. . . .	. . . .
2019-03-20 08:00:00	Nik Bear Brown
2017-03-20 16:00:00	Nik Bear Brown
2017-03-20 20:00:00	Nik Bear Brown

Solution:

```
Browns_Lectures.filter(year('Date') == '2019').filter(dayofyear('Date') >= 72).groupBy(dayofyear('Date')).
count().orderBy('dayofyear('Date')').show()
```

Q18 (5 Points)

Name three benefits of Spark over MapReduce?

Solution:

Spark has the following benefits over MapReduce:

Ease of use.

Unlike Hadoop, Spark provides inbuilt libraries to perform multiple tasks from the same core like batch processing, Steaming, Machine learning, Interactive SQL queries. However, Hadoop only supports batch processing.

Hadoop is highly disk-dependent whereas Spark promotes caching and in-memory data storage.
Spark is capable of performing computations multiple times on the same dataset. This is called iterative computation while there is no iterative computing implemented by Hadoop

Q19 (5 Points)
Explain the concept of Resilient Distributed Dataset (RDD)?

Solution:

Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. ... RDDs can be created through deterministic operations on either data on stable storage or other RDDs. RDD is a fault-tolerant collection of elements that can be operated on in parallel.