===Map-Reduce===

MT0. Which of the following statememts about map-reduce are true? Check all that apply.

(a) If you only have 1 computer with 1 computing core, then map-reduce is unlikely to help

(b) If we run map-reduce using N computers, then we will always get at least an N-Fold speedup compared to using 1 computer

(c) Because of network latency and other overhead associated with map-reduce, if we run map-reduce using N computers, then we will get less than N-Fold speedup compared to using 1 computer

(d) When using map-reduce with gradient descent, we usually use a single machine that accumulates the gradients from each of the map-reduce machines, in order to compute the paramter update for the iterion

Answer: (a),(c), and (d)

===Order inversion===

MT1. Suppose you wish to write a MapReduce job that creates normalized word co-occurrence data form a large input text. To ensure that all (potentially many) reducers receive appropriate normalization factors (denominators) in the correct order in their input streams (so as to minimize memory overhead), the mapper should emit according to which pattern:

(a) emit (*,word) count

(b) There is no need to use order inversion here

(c) emit (word,*) count

(d) None of the above

Answer: (c)

===Apriori principle===

MT2. When searching for frequent itemsets with the Apriori algorithm (using a threshold, N), the Apriori principle allows us to avoid tracking the occurrences of the itemset {A,B,C} provided

(a) all subsets of {A,B,C} occur less than N times.

(b) any pair of {A,B,C} occurs less than N times.

(c) any subset of {A,B,C} occurs less than N times.

(d) All of the above

Answer: (d)

===Bayesian document classification===

MT3. When building a Bayesian document classifier, Laplace smoothing serves what purpose?

(a) It allows you to use your training data as your validation data.

(b) It prevents zero-products in the posterior distribution.

(c) It accounts for words that were missed by regular expressions.

(d) None of the above

Answer: (b)

===Bias-variance tradeoff===

MT4. By increasing the complexity of a model regressed on some samples of data, it is likely that the ensemble will exhibit which of the following?

(a) Increased variance and bias

(b) Increased variance and decreased bias

(c) Decreased variance and bias

(d) Decreased variance and increased bias

Answer: (d)

===Combiners===

MT5. Combiners can be integral to the successful utilization of the Hadoop shuffle. This utility is as a result of

(a) minimization of reducer workload

(b) both (a) and (c)

(c) minimization of network traffic

(d) none of the above

Answer: (b)

===Pairwise similarity using K-L divergence===

In probability theory and information theory, the Kullback–Leibler divergence (also information divergence, information gain, relative entropy, KLIC, or KL divergence) is a non-symmetric measure of the difference between two probability distributions P and Q. Specifically, the Kullback–Leibler divergence of Q from P, denoted DKL(P||Q), is a measure of the information lost when Q is used to approximate P:

For discrete probability distributions P and Q, the Kullback–Leibler divergence of Q from P is defined to be

KLDistance(P, Q) = Sum over i (P(i) log (P(i) / Q(i))

In the extreme cases, the KL Divergence is 1 when P and Q are maximally different and is 0 when the two distributions are exactly the same (follow the same distribution).

For more information on K-L Divergence see:

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence (https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence)

For the next three question we will use an MRjob class for calculating pairwise similarity using K-L Divergence as the similarity measure:

Job 1: create inverted index (assume just two objects) Job 2: calculate/accumulate the similarity of each pair of objects using K-L Divergence

Download the following notebook and then fill in the code for the first reducer to calculate the K-L divergence of objects (letter documents) in line1 and line2, i.e., KLD(Line1||line2).

Here we ignore characters which are not alphabetical. And all alphabetical characters are lower-cased in the first mapper.

http://nbviewer.ipython.org/urls/dl.dropbox.com/s/9onx4c2dujtkgd7/Kullback%E2%80%93Leibler%20diverge
MIDS-Midterm.ipynb
(http://nbviewer.ipython.org/urls/dl.dropbox.com/s/9onx4c2dujtkgd7/Kullback%E2%80%93Leibler%20diverg
MIDS-Midterm.ipynb)
https://www.dropbox.com/s/zr9xfhwakrxz9hc/Kullback%E2%80%93Leibler%20divergence-MIDS-Midterm.ipynb?dl=0
(https://www.dropbox.com/s/zr9xfhwakrxz9hc/Kullback%E2%80%93Leibler%20divergence-MIDS-Midterm.ipynb?dl=0)

MT6. Which number below is the closest to the result you get for KLD(Line1||line2)? (a) 0.7 (b) 0.5 (c) 0.2 (d) 0.1

Answer: (d) 0.1

MT7. Which of the following letters are missing from these character vectors? (a) p and t

(b) k and q

(c) j and q

(d) j and f

Answer: (c)

MT8. The KL divergence on multinomials is defined only when they have nonzero entries. For zero entries, we have to smooth distributions. Suppose we smooth in this way:

(ni+1)/(n+24)

where ni is the count for letter i and n is the total count of all letters.

After smoothing, which number below is the closest to the result you get for KLD(Line1||line2)??

(a) 0.08

(b) 0.71

(c) 0.02

(d) 0.11

In [ ]:

```
Answer: (a)
```

===Gradient descent===

MT9. Which of the following are true statements with respect to gradient descent for machine learning, where alpha is the learning rate. Select all that apply

(a) To make gradient descent converge, we must slowly decrease alpha over time and use a combiner in the context of Hadoop.

(b) Gradient descent is guaranteed to find the global minimum for any function J() regardless of using a combiner or not in the context of Hadoop

(c) Gradient descent can converge even if alpha is kept fixed. (But alpha cannot be too large, or else it may fail to converge.) Combiners will help speed up the process.

(d) For the specific choice of cost function J() used in linear regression, there is no local optima (other than the global optimum).

Answer: (c) and (d)

===Weighted K-means===

Write a MapReduce job in MRJob to do the training at scale of a weighted K-means algorithm.

You can write your own code or you can use most of the code from the following notebook:

http://nbviewer.ipython.org/urls/dl.dropbox.com/s/kjtdyi10nwmk4ko/MrJobKmeans-MIDS-Midterm.ipynb (http://nbviewer.ipython.org/urls/dl.dropbox.com/s/kjtdyi10nwmk4ko/MrJobKmeans-MIDS-Midterm.ipynb) https://www.dropbox.com/s/kjtdyi10nwmk4ko/MrJobKmeans-MIDS-Midterm.ipynb?dl=0 (https://www.dropbox.com/s/kjtdyi10nwmk4ko/MrJobKmeans-MIDS-Midterm.ipynb?dl=0)

Weight each example as follows using the inverse vector length (Euclidean norm):

weight(X)= 1/||X||,

where ||X|| = SQRT(X.X)= SQRT(X1^2 + X2^2)

Here X is vector made up of X1 and X2.

Using the following data answer the following questions:

https://www.dropbox.com/s/ai1uc3q2ucverly/Kmeandata.csv?dl=0 (https://www.dropbox.com/s/ai1uc3q2ucverly/Kmeandata.csv?dl=0)

MT10. Which result below is the closest to the centroids you got after running your weighted K-means code for 10 iterations?

(a) (-4.0,0.0), (4.0,0.0), (6.0,6.0)

(b) (-4.5,0.0), (4.5,0.0), (0.0,4.5)

(c) (-5.5,0.0), (0.0,0.0), (3.0,3.0)

(d) (-4.5,0.0), (-4.0,0.0), (0.0,4.5)

Answer: (c)

MT11. Using the result of the previous question, which number below is the closest to the average weighted distance between each example and its assigned (closest) centroid?

The average weighted distance is defined as sum over i (weighted_distance_i) / sum over i (weight_i)

(a) 2.5 (b) 1.5 (c) 0.5 (d) 4.0

Answer: (d)

MT12. Which of the following statements are true? Select all that apply. a) Since K-Means is an unsupervised learning algorithm, it cannot overfit the data, and thus it is always better to have as large a number of clusters as is computationally feasible. b) The standard way of initializing K-means is setting $\mu1=\cdots=\mu k$ to be equal to a vector of zeros. c) For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide. d) A good way to initialize K-means is to select K (distinct) examples from the training set and set the cluster centroids equal to these selected examples.

Answer: (b), (c), (d)

```
In [ ]:
```