CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Day 14
Midterm Review

Announcements and Feedback

- Project Phase 1 is due tonight @ 11:59PM
  - Submission is via UBLearns – one per team
  - TEST SUBMISSIONS ASAP! It will be easier to address issues in the submission process now, not so much tonight.

Midterm Review

Potential Topics:

1. Linear Regression
2. Supervised Learning: Classification using K-NN
3. Unsupervised Learning: K-Means Clustering
4. HDFS Architecture and Protocol
5. MapReduce
6. PageRank/Graph Processing
7. Word Co-Occurrence

Linear Regression [Lec 4-6]

Explain the basic components of a Linear Regression model and what they mean/how to interpret them.

Understand and discuss evaluation metrics for determining the effectiveness of a given linear regression model.

How can you help ensure that your model does not end up overfitting to your particular dataset?

Classification with K-NN [Lec 4-6]

Given some simple data points, determine the classification of an unknown point for different values of k.

Understand and discuss different evaluation metrics for determining the effectiveness of a given K-NN model.

Understand and discuss the potential impact of data scaling and similarity metrics.

K-Means Clustering [Lec 4-6]

Given a set of simple data points, determine centroids and cluster membership for the dataset.

Discuss potential interpretations for a given clustering.

Understand and discuss potential issues with K-Means clustering.

Hadoop and HDFS Architecture [Lec 7,8]

Understand and discuss the evolution of Hadoop from 1.0 to 2.0.

Understand the basics of the HDFS architecture, the different components involved, and their roles and responsibilities.

Understand and discuss block replication and its importance.

MapReduce [Lec 8-10,12,13]

Understand and discuss the roles of the different types of MapReduce tasks that are part of a MapReduce Job.

Understand and discuss how data flows throughout a MapReduce job and how it is split up over mappers and/or reducers.

Be able to read and understand MapReduce pseudocode.

Be able to write basic MapReduce pseudocode to accomplish a given task.

Understand the basics of the NGS case study: NGS k-mer problem description, basics of implementation, unique characteristics of the k-mer problem, and the basics of spills.

Graph Processing/PageRank [Lec 11,12]

Understand and discuss how graphs can be represented and operated on in MapReduce.

Understand and discuss the basic formulation of PageRank.

Given a simple graph of webpages, be able to compute a few iterations of the PageRank algorithm.

Understand and discuss the dead-end and spider trap issues, how they affect the PageRank computation, and how we can modify the basic PageRank algorithm to deal with them.

Understand and discuss the MapReduce implementation of PageRank, both in the naive formulation, and the modifications needed to deal with dead-ends and spider traps.

Word Co-Occurrence [Lec 13]

Understand and discuss the relevance of word co-occurrence.

Understand and discuss the basic matrix formulation of the problem.

Understand and discuss the two different MapReduce formulations (pairs and stripes), and the pros and cons of each formulation.

Understand and discuss the difference between absolute and relative co-occurrence.

Understand and discuss the modifications needed to compute relative co-occurrence with both the pairs and stripes method.