

CSE4/587 Midterm Exam Review Spring 2018

1. SEMI-SUPERVISED LEARNING: CLASSIFICATION USING K-NN (20 POINTS)

Given a problem, create a model. Based on this model, find the answer for queries values.

2. UNSUPERVISED LEARNING: K-MEANS CLUSTERING (20 POINTS)

Given a data set, derive the clusters for k centroids.

3. HDFS ARCHITECTURE. PROTOCOL (20 POINTS)

HDFS architecture, how is different from you normal file system? Why do you need this? Details of architecture and operations.

4. MAPREDUCE FOUNDATIONS (20 POINTS)

A formal representation of the various elements of the MR workflow: combiners, partitioners. How many reducers? How many reduce calls? Input of mappers and reducers etc. You will derive expression in terms of the variables given. Hadoop architecture.

5. PROBLEM SOLVING USING MR (20 POINTS)

Given a big-data problem provide a MR solution. Pseudo code samples in Ch 2 of Lin and Dryer.

6. DATE, TIME, LOCATION

Location: Knox 20

Date: 3/26/2019

Time: 3.30-4.50PM

Closed book exam.

How to study? Collect the material needed for the topics listed. It is all there in the class notes. Work on problems and discuss it with your friends. There is NO SUBSTITUTE for hard work. Work on mastering these topics, it will provide a solid foundation for data science projects that you may undertake in the future