CSE 4/587
Data Intensive Computing

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

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Midterm Review #1

Announcements

- Midterm is Wednesday @ 11:00-11:50 AM
- In Cooke 121 AND Alumni 97
  - If your UB email begins with a-j, you will be in Alumni 97
  - If your UB email begins with k-z you will be in Cooke 121

Midterm Format and Logistics

- [a-j] report to Alumni 97, [k-z] report to Cooke 121
  - Seating will be randomized within each exam room
- Handwritten, Closed Book
  - You must have a pen/pencil, and your UB ID card
  - You may also bring a non-graphing calculator
  - All other bags/electronics/etc will be placed at the front of the room during the exam
- 3-4 short answer questions with a mix of theory and application

Midterm Review

Potential Topics:

1. Data Science Overview
2. Linear Regression
3. Unsupervised Learning: K-Means Clustering
4. Classifiers: General Use
5. Classifiers: K-NN
6. Classifiers: Naive Bayes
7. Classifiers: Logistic Regression

Data Science Overview [Lec 2-5]

1. Understand the overall goals and challenges of DIC
2. Know the four Vs and what they mean
3. Understand the various skills and components that DIC encompasses
4. Know what data cleaning/EDA is and the difference between then two

Linear Regression [Lec 6]

1. Explain the basic components of a Linear Regression model and what they mean/how to interpret them.
2. Understand and discuss evaluation metrics for determining the effectiveness of a given linear regression model.
3. How can you help ensure that your model does not end up overfitting to your particular dataset?

K-Means Clustering [Lec 7, 11-12]

1. Given a set of simple data points, determine centroids and cluster membership for the dataset.
2. Discuss potential interpretations for a given clustering.
3. Understand how K-Means can be used to improve results from other models.
4. Understand and discuss potential issues with K-Means clustering.

Classifiers [8-12]

1. Understand the classification of classifiers
2. Understand the development cycle of a classification problem
3. Understand the basics of the different classifiers we have discussed in class and how to use them
4. Understand the pros/cons of the classifiers discussed in class

Classification with K-NN [Lec 8, 12]

1. Given some simple data points, determine the classification of an unknown point for different values of k.
2. Understand and discuss different evaluation metrics for determining the effectiveness of a given K-NN model.
3. Understand and discuss the potential impact of data scaling and similarity metrics.

Naive Bayes [Lec 9-10]

1. Know the formulation of Bayes Law, and how to apply it to a given problem
2. Know how to take the application of many instances of Bayes Law and aggregate them into a single probability for the Naive Bayes model
3. Understand what Laplace Smoothing is, and what it addresses

Logistic Regression [Lec 11]

1. Know what an odds ratio is
2. Know what the logit function is, and how to apply it to a given odds ratio
3. Know the final formula for logistic regression