Study guide for CS484 final exam. This list of topics is not meant to be comprehensive, but outlines the important topics that you should focus on. More emphasis will be on the topics covered after Midterm 2.

Chapter 1. Introduction

- Data mining overview – understand what each task does and when to use them.
  - Classification
  - Clustering
  - Association Analysis
  - Anomaly Detection

Chapter 2. Data

- Common issues with data
- Data preprocessing
- Similarity/distance measures

Chapters 4 & 5 – Classification

- Basic concepts
- Different classifiers (what they are, how they work, strengths/weaknesses, choice of classifier given specific problems/datasets, etc)
- You should know how the following work in depth, i.e. you should know how to build a classifier from scratch and/or classify a test instance using the following.
  - Decision trees
  - Naïve Bayes
  - Nearest Neighbor
- In addition, you should know the following in high-level:
  - Bayesian Networks
  - SVM
  - Ensemble methods (different types, advantages of using ensemble methods, when to use them)
- Common issues/challenges
  - With classification in general
  - With different classifiers
- Evaluation
  - Accuracy, precision, recall, F-measure, ROC curve
- Bias and Variance, Model complexity

Chapter 8 – Clustering

- Basic concepts - what is clustering, how does it differ from supervised learning, when to use
- Different types of clustering algorithms (what they are, how they work, strengths/weaknesses)
  - Partitional – k-means and its variants
  - Hierarchical – agglomerative algorithm with various linkage methods, dendrogram
  - Density-based – dbscan
- Common issues/challenges
  - With clustering in general
  - With different algorithms
- Evaluation/cluster validity

        - SSE, correlation/similarity matrix, cohesion/separation, Silhouette coefficient, cophenetic coefficient (for hierarchical clustering), external measures (what you did for the k-means homework)

Chapter 6 – Association Rules Mining

- Basic concepts
- Apriori-type algorithms/support-confidence framework
    - Candidate generation/pruning (brute-force, $F_{k-1}xF_1$, $F_{k-1}xF_{k-1}$)
    - Rule generation/pruning
    - Efficient candidate counting with hash tree
    - Lattice / frequent itemsets border
- Compact representations
- Pattern evaluation (interestingness measures)
    - support, confidence, lift/interest factor, leverage
- Common issues/challenges

Chapter 10 – Anomaly Detection

- Basic concepts
- Different approaches
- Know how to use supervised & unsupervised algorithms to detect anomalies
- LOF
- Common issues/challenges