CMPE 239 – Spring 2017

Final Study Guide

This guide outlines the topics that will be covered in the final exam. You may notice that several topics are omitted – these will not be included. The focus of the final exam will be on the material that was not covered by the midterm exam. However, there will be theoretical/short answer questions that cover some of the midterm material and "combination" questions that will require knowledge of topics that were covered in either part (no long exercises though).

The midterm will consist of two sections, worth 25 points each. The first section will be multiple-choice, consisting primarily of theoretical and short computation questions. The second section will have a few longer problems you will need to solve. To prepare for the exam, review class materials (slides, activities) and homework assignments. At a minimum, be familiar with the topics mentioned below, which are a subset of those included in the syllabus and slides.

You will need a calculator for the exam. However, if needed, the values of log numbers will be provided to you. **Smart phones cannot be used as calculators and their use is forbidden.** Complex formulas will be provided, but you should know basic ones which were required in homework assignments (e.g., cosine similarity, Euclidean distance) and are mentioned in the list below.

****** pre-midterm material you need to review ******

- Introduction
  - Know the difference between descriptive and predictive methods.
  - Be able to identify the data mining method that should be used to solve a given problem.
  - What are some differences between Web and Data Mining?
  - Understand supervised vs. unsupervised learning.
- Data
  - Understand the difference between objects and attributes and be familiar with other names used for them.
  - Be able to identify the type of an attribute based on their properties (distinctness, order, etc.).
  - Asymmetric attributes
  - Data preprocessing strategies
  - Proximity measures
    - Be able to compute (know formulas) basic proximity functions: cosine, Euclidean, Minkowski, Simple matching, Jaccard coefficient.
    - What is and what are the properties of a metric?
    - Given the formulas (i.e., don't need to remember), be able to compute entropy and mutual information.
    - Combine similarities given multiple representations of the same objects.
    - Understand the concept of density.
- Nearest Neighbors (NN)
  - Understand the difference between exact and approximate NN.
  - Be able to transform text documents into a sparse matrix (see demo).

- Identify algorithms most appropriate for solving NN in different scenarios.
- What are families of locality sensitive functions?
- Understand the general locality sensitive hashing (LSH) procedure (see demo) and how to apply it for hamming distance, l-p norms, Jaccard coefficient, and cosine similarity.
- How do we increase accuracy and/or recall in LSH (slides + demo)?
- What is the Cauchy-Schwarz inequality (know formula) and how can you use it in filtering?

❖ Dimensionality Reduction (DR)

- Understand the difference between feature selection and feature extraction.
- Identify the goal of different dimensionality reduction methods.
- Be familiar with the general procedures for DR using Truncated SVD, PCA, LLE, Random Projections, and LDA (no need to remember formulas).
- How do we choose the number of dimensions k in PCA s.t. at least p% variance is captured by the model (see slides, demo, homework)?

❖ Association Analysis / Pattern Mining

- Understand what are (frequent) itemsets, association rules, sequences.
- Define and find support (count) and confidence for itemsets, association rules, sequences.
- Understand the Apriori principle and how to apply it to prune the frequent itemset search space.
- Compute the number of considered/counted itemsets naïvely and/or using support-based pruning (se homework).
- Understand what is a projected database and how to construct one.
- Compact representation of frequent itemsets: understand what are and how to identify and count closed and maximal frequent itemsets.
- Be able to create a sequence from a sequence database.

****** post-midterm material you need to know ******

Since we spent most of our time on classification and clustering, most questions will come from those areas, though you should be familiar with the other material we discussed as well.

❖ Classification

- Understand the difference between prediction and inference.
- Understand the difference between classification and regression.
- Steps in the general classification process.
- Understand and know the difference between underfitting and overfitting.
- Given a confusion matrix, compute (know the formulas) accuracy, prediction, recall, and F-measure.
- Know how to generate a confusion matrix given two lists, one containing the true and the other the predicted classes for a set of objects.
- Understand the process of model selection.
- Identify different methods of classifier evaluation.
- Apply a decision tree model to classify an object.
- Instance-based learning
    - Know how to apply kNN given a proximity measure.

        - How to choose k
        - Scaling issues
        - How to improve kNN efficiency
    - Naïve Bayes classifier
        - Bayes theorem formulation
        - Know how to identify the hypothesis and evidence in a word problem.
        - Maximum a-posteriori (MAP) hypothesis
        - Why is it naïve? That is the basic assumption? How and why is the Bayes theorem simplified?
        - Be able to apply a Bayes model to make predictions.
        - Issues with Naïve Bayes classifier
    - Bayesian Belief Network
        - Main components
        - Be able to apply a BBN model to predict the class of a sample.
    - Artificial Neural Network
        - What is a perceptron?
        - What type of optimization is used to learn the perceptron weights?
        - Be able to apply the perceptron learning rule for several iterations.
        - What is an activation function?
        - Characteristics of ANN
    - SVM
        - Basic properties (e.g., binary linear classifier, can define soft margin, etc.)
        - How to transform into a nonlinear SVM
        - Advantages and disadvantages
    - Ensemble methods
        - Differences between bagging and boosting
        - Understand how random forest works.
- Regression
    - Be able to apply a linear regression model.
    - Given the intercept and slope formulae, be able to learn a linear regression model for a 2D dataset.
    - Parameter estimation
    - Differences between regularized regression models (ridge, lasso, elastic net)
    - Be able to apply kNN regression to predict the output for a sample
- Clustering
    - What is cluster analysis?
    - K-Means
        - Know the algorithm and be able to apply it on a set of vectors
        - Complexity
        - Limitations of K-means
        - Solutions to the initial-centroid problem
        - Important pre and post-processing steps
        - How to pick the right k
    - Hierarchical clustering

- What is a dendrogram?
- Advantages of hierarchical clustering
- Cluster proximity choices (match name and distance)
- Be able to build dendrogram given a pre-defined cluster proximity.

  - Density-based clustering
    - DBSCAN points definitions. Be able to identify points in each category in a dataset given MinPts and Eps values.
    - How to choose Eps and MinPtrs
  - Cluster validity
    - Measures of cluster validity
    - Formulae for cohesion and separation. Be able to compute WSS and BSS given a set of points and a clustering of those points.
    - General idea about silhouette coefficient
    - External measures of cluster validity
- Recommender Systems
  - User/item representation, utility matrix
  - Key problems in recommender systems
  - When does RS do its job well?
  - Neighborhood-based (user/item-based) collaborative filtering
    - User/Item representation
    - The basic techniques
    - Basic assumptions
    - Similarities, prediction, recommendation
    - Know how to compute a rating given a utility matrix and a target user-item pair.
  - Latent factor models
    - Interpretation of latent factors
    - Basic ides (approximate utility matrix of ratings with two low-rank matrices)
    - Estimation by minimizing the square error
    - Be able to “predict” if given the decomposed matrix
  - Evaluating recommender systems
    - Know formulae and differences between MAE and RMSE
    - Given a list of top-N recommendations and which were hits, be able to compute AP and ARHR.
- Information Retrieval
  - Understand Web search users and their query needs
  - Steps to constructing an inverted index
  - Understand the overall technique of MapReduce
  - Different ways to construct search result summaries
  - Evaluation of a search engine (relevance benchmarks, precision, recall, F1 – how are they different than in the case of classification)
- Link Analysis
  - Motivation
  - Social network analysis
    - Basic definitions & semantics

- Prestige (rank, degree, proximity) – understand the difference, know formulas included in the slides
- PageRank
    - Intuition
    - The PageRank algorithm: understand how it works and what each component represents
    - Advantages & criticisms