Midterm Review Topics
CSCI 452 – Data Mining
Dr. Schwartz

1. Introductory lectures
    a. Definitions of data mining
    b. Potential applications
    c. Types of data mining questions (descriptive, exploratory,…)
    d. Building block problems (association pattern, clustering, outlier, classification)
    e. Experimental design
2. Statistical Learning Basics
    a. General form ($Y = f(X) + e$)
    b. Prediction vs Inference
    c. Types of error (reducible vs irreducible)
    d. Parametric vs. Non-parametric methods
    e. Model flexibility vs interpretability
    f. Training and test sets
    g. Supervised vs unsupervised learning
    h. How to measure performance (confusion matrices, mean squared error)
    i. Overfitting
    j. Bias, variance, and bias-variance tradeoff
3. Statistical Estimation and Prediction
    a. Point estimation
    b. Confidence Intervals
    c. Normal distributions, standard deviation and standard error
    d. t-values and p-values
    e. Margin of error
4. Regression
    a. Simple linear regression (single predictor variable)
    b. Residual sum of squares and least squares
    c. Residual standard error and R^2 statistics
    d. Confidence interval vs prediction interval
    e. Root mean squared error
    f. Multiple linear regression
    g. General formula (what we're estimating)
    h. Correlation matrix
    i. Interaction terms
    j. Extrapolation
    k. Regression with qualitative variables – dummy variables, different encodings
    l. Logistic regression
    m. Pros and cons of regression
5. Preprocessing Data
    a. Handling missing data
    b. How and why to reclassify categorical variables

    c. Binning numeric values
    d. When to remove (or not remove) variables
6. Decision Trees
    a. Uses and error measurement
    b. Types of nodes
    c. How to build a decision tree (Basic algorithm – don't need to know differences between Hunt's algorithm, ID3 family, CART)
    d. Choosing the split condition (nominal, ordinal, continuous)
    e. Choosing the attribute on which to split (gini, entropy, misclassification error)
    f. Information gain/gain
    g. Bias of impurity measures
    h. Overfitting and pruning
    i. Advantages and disadvantages