Midterm 2 material

CMSC 320

This document describes material that will be fair game in the second midterm exam.

Exploratory Data Analysis

Summary Statistics

- Distributional characteristics: range, central tendency, spread
- Statistical summaries: sample mean, sample median, sample standard deviation
- The derivation of the mean as an *optimal* central tendency statistic
- The five-number summary of data and relationship to boxplot
- Statistical summaries of pairwise relationship between variables: sample covariance and correlation

Visualization for EDA

- Plots to show data distribution for one variable/two variables
- The data/aesthetic mapping/geometric representation scheme for data visualization (ggplot)

Data transformations

- Centering and scaling data transformation (standardization)
- Standard units
- Ways of discretizing continuous numeric data
- Relationship between arithmetic and geometric mean
- The logarithmic transformation for skewed data

Introduction to Statistical Learning

- The "inverse problem" way of thinking about data analysis
- Properties of discrete probability distributions
- Expectation for discrete probability distributions
- How the sample mean is an *estimate* of expected value
- The law of large numbers and the central limit theorem
- The Bootstrap procedure
- The Bernoulli, Binomial and Normal distributions
- Using the CLT to get a confidence interval for the mean
- Using the CLT to test a simple hypothesis about the mean

- Application to A/B Testing
- Joint and conditional distribution for discrete probability distributions
- Bayes Rule
- Independence
- Conditional expectation for discrete probability distributions

Linear models for regression

- The linear regression model
- Estimating linear regression parameters by minimizing residual sum of squares (RSS)
- Diagnostic plots for linear regression
- How to encode categorical predictors in a linear regression model, and how to interpret their coefficient estimates
- How to incorporate and interpret predictor interactions in a linear regression model
- Constructing a confidence interval for a parameter estimate in the linear regression model.
- The R^2 measure to assess global fit in a regression model
- What is co-linearity

Linear models for classification

- What is a classification problem?
- Understanding classification as a probability estimation problem.
- Why shouldn't you use linear regression (for continuous outcomes) to predict outcome for a binary categorical variable
- What is log-odds? How do we transform log-odds to probabilities?
- How is the logistic regression problem defined.
- How do we calculate error rate for a classification problem?
- What are False positive and false negative errors?
- What is the False positive rate? True positive rate?
- What are precision and recall?

Gradient descent

- Gradient descent for linear regression
- The Maximum Likelihood principle
- Gradient descent for logistic regression

Midterm Structure

The midterm will consist of two sections: ~5 multiple choice questions, and 2 longer questions. The midterm will be take-home, and you will have 5 days to complete. All resources are at your disposal (lecture notes, recordings, etc.). The code of academic integrity still applies. This is to be done independently.