Midterm 2 material

CMSC 320

This document describes material that will be fair game in the midterm exam. Each section is divided into two levels (level 1 and 2). Mastery of level 1 material is essential to do well in the midterm, level 2 is needed to do great in the midterm.

Exploratory Data Analysis

Level 1

Summary Statistics

- Distributional characteristics: range, central tendency, spread
- Statistical summaries: sample mean, sample median, sample standard deviation

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

- Difference between data missing systematically vs. missing at random
- Centering and scaling data transformation (standardization)
- Standard units
- Ways of discretizing continuous numeric data

Level 2

- The derivation of the mean as an *optimal* central tendency statistic
- Rank summary statistics
- Distributional characteristic: skew
- The five-number summary of data and relationship to boxplot
- Statistical summaries of pairwise relationship between variables: sample covariance and correlation
- The logarithmic transformation for skewed data
- Imputing continuous numeric missing data

Introduction to Statistical Learning

Level 1

- Sources of randomness and stochasticity in data
- 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 Bernoulli, Binomial and Normal distributions
- Joint and conditional distribution for discrete probability distributions
- Bayes Rule
- Conditional expectation for discrete probability distributions

Level 2

- 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

Linear models for regression

Level 1

- The linear regression model
- Estimating linear regression parameters by minimizing residual sum of squares (RSS)
- Fitting a linear regression model in R using the `lm` function
- How the t-statistic and t-test is used in linear regression.
- Diagnostic plots for linear regression
- How to encode categorical predictors in a linear regression model, and how to interpret their coefficient estimates

Level 2

- 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

Level 1

- What is a classification 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.

- Fitting a logistic regression problem using the `glm` function.
- 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?

Level 2

- Understanding classification as a probability estimation problem.
- What are precision and recall?
- How do you construct an Receiver Operator Curve (ROC) using True Positive and False positive rates?

Tree-based methods

Level 1

- What is a regression tree?
- What is a classification (decision) tree?
- Do tree-based methods learn linear or non-linear functions between predictors and outputs?
- How to use recursive partitioning to build a regression tree

Level 2

- What does it mean to "prune" a decision tree, why is that a good idea?
- What is the random forest method? What is it's relationship to regression and decision trees.
- How can we measure "variable importance" using the random forest algorithm.

Model evaluation using resampling

Level 1

- What is the difference between *model assessment* and *model selection*
- Describe how k-fold cross validation is used for model assessment. Describe how k-fold cross validation is used for *model selection*.
- How to compare models using cross-validation estimates of error.

Level 2

- Why is k-fold cross validation preferable over other resampling methods (e.g., single validation set, or resampled validation sets).

Midterm Structure

The midterm will consist of three sections: ~10-15 multiple choice questions, ~5-8 short questions, and 1 or 2 longer questions. Multiple choice will test concept definitions along with problems similar to written exercises in class. Short questions will be similar to written problems done in class or homework, along with concept questions where longer written answers are required. Longer questions are for problem solving (e.g., design a data pipeline to carry out a specific task, prove a property of a summary statistic, etc.)