Midterm material

CMSC 320

This document describes what 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.

Preliminaries

Level 1

- Data Analysis Cycle: preparation -> modeling -> communication

Level 2

- Data Analysis Cycle: as presented in slides/Zumen & Mount

R

Level 1

- Variables vs. values
- All the many ways to index vectors/data.frames
- Functions, conditionals, loops
- Lists vs. vectors
- Matrices

Level 2

- vectorization
- the `apply` family

Measurement types

Level 1

- categorical
- ordered categorical (we didn't see this one in class)
- discrete numerical
- continuous numerical

Level 2

- factors/levels in R
- the importance of units

Best practices

Level 1

- the importance of reproducibility
- tools to improve reproducibility

Level 2

- the importance of thinking like an experimentalist

Tidy Data and Data Models

Level 1

- Definition of a tidy data set
- Components of a Data Model
- Basics of the Entity-Relationship Data Model
- The components of an ER diagram

Level 2

- Relationship between tidy data and normal form
- JSON
- Other data models

Data Wrangling

Level 1

- `dplyr` single table verbs
- the SFW SQL query
- different join semantics
- difference between data missing systematically vs. missing at random
- why are database systems helpful and useful?

Level 2

- Keys/Foreign Keys in the Entity-Relationship data model
- How an ER diagram is converted into a set of Relations (data tables)
- Views and integrity constraints in database systems
- Imputing continuous numeric missing data

Exploratory Data Analysis

Level 1

- Plots to show data distribution for one variable/two variables
- Distributional characteristics: range, central tendency, spread
- Statistical summaries: sample mean, sample median, sample standard deviation
- The data/aesthetic mapping/geometric representation scheme for data visualization (ggplot)
- Centering and scaling data transformation (standardization)
- Standard units
- Ways of discretizing continuous numeric data

Level 2

- Rank summary statistics
- Distributional characteristic: skew
- The derivation of the mean as central tendency statistic
- 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

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
- The statement of the central limit theorem
- The Bernoulli, Binomial and Normal distributions
- Joint and conditional distribution for discrete probability distributions
- Conditional expectation for discrete probability distributions

Level 2

- What is the normal distribution for the sample mean given by the CLT
- Using the CLT to get a confidence interval for the mean
- Using the CLT to test a simple hypothesis about the mean