Midterm 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.

Preliminaries

Level 1

- Data Analysis Cycle: acquisition -> 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 (ordinal)
- 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
- data science ethics and responsible conduct of research

Level 2

- the importance of thinking like an experimentalist

Data Wrangling

Level 1

- `dplyr` single table verbs
- the Select-From-Where SQL query
- different join semantics
- 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)

Tidy Data and Data Models

Level 1

- Components of a Data Model
- Basics of the Entity-Relationship and Relational Data Models
- The components of an ER diagram
- The relationship between tidy data, the ER and the Relational models

Level 2

- JSON
- Other data models

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)
- Imputing continuous numeric missing data
- 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

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 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

- Using the CLT to get a confidence interval for the mean
- Using the CLT to test a simple hypothesis about the mean

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.)