Final Material

CMSC 320

This document describes what will be fair game in the final exam. Each section is divided into two levels (level 1 and 2). Mastery of level 1 material is essential to do well in the final, level 2 is needed to do great in the final. The final covers material from the entire course, but is weighted roughly 2/3 towards material in the second part of the semester (starting with linear models below).

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 (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

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
- How to incorporate and interpret predictor interactions in a linear regression model

Level 2

- The closed form solution for the simple 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
- How the F-test is used to test relationship between outcome and sets of predictors
- 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.
- The LDA (linear discriminant analysis) classification model. How to fit it using group-by/summarize queries.
- The Naive Bayes classification model. How to fit it using group-by/summarize queries.
- 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.

The support vector machine

Level 1

- How should we encode (0/1 or -1/+1) categorical outcome data to fit a support vector machine.
- Why is it called a support vector machine.
- How to fit an svm using the `svm` function in the `e1071` R package.

Level 2

- What is the purpose of the "cost" parameter in an SVM.
- What is a kernel function, why do we use them in SVMs?
- Why is looking at the number of support vectors in a fitted SVM useful?

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).

Unsupervised methods

Level 1

- What is the distinction between unsupervised and supervised methods?
- Why is PCA a “dimensionality reduction” method?
- What is the objective function of the PCA problem?
- The role of scaling and centering transformations in the PCA problem?

Level 2

- What is the relevance of the ‘percent variance explained’ metric for PCA?
- How can we determine predictor correlation from the result of PCA?

Gradient Descent

- What is the update rule for multivariate linear regression
- What is the update rule for logistic regression
- What is the general form of the gradient descent algorithm
- What is the difference between the stochastic and batch versions of gradient descent

Communication

- What are some of the advantages provided by interactivity in the graphical presentation of data.