DS 303 Final Exam Study Guide

Our final exam is scheduled for Tuesday, Dec. 12 from 12 - 2 pm. It will take place in our usual lecture room. It is open note/open internet. All work **must** be your own. You must take the final exam in order to complete the class.

The final exam is comprehensive and any material discussed in lecture or posted on Canvas from the dates of Aug. 21, 2023 - Dec. 3, 2023 is fair game for the exam.

Study Resources

1. Homeworks + homework solutions
2. Midterm 1 and Midterm 2
3. Your lecture notes and `R` scripts
4. In-class activities
5. Textbook

Final Exam Topics (in chronological order)

You will not have time to go through all of these topics perfectly. **Be thoughtful and strategic in your approach to studying.** Many of the methods are constructed from **the same building blocks**. Focus on making sure your understanding of those building blocks are solid.

Since this a timed exam, organization is important. I **strongly** recommend putting together your own `R` script that contain important code snippets. You should practice your coding and implementation. That means **going beyond** just copy/paste of code given to you in class and making sure you understand the nuances of the code.

Topics listed in **red** are my not-so-subtle hint to you to spend extra time on these topics. It does **not** mean other topics are not important or will not appear on the final. But if you are cramming in the last 24 hours to study for the exam - these are the ones I would focus on first.

1. **Bias-variance tradeoff**
    - What is the definition of bias and variance in mathematical terms?
    - What is the definition of bias and variance in conceptual terms (plain language)?
    - For all supervised learning methods, how does bias and variance behave?
    - How do you calculate bias (square) and variance for simulated data?
2. **Multiple Linear Regression**
    - Setup the model. How do we estimate parameters in this model (mathematically and conceptually)?

- When does the method break down?
- When does the method do particularly well?

3. **Inference for Multiple Linear Regression and Multiple Testing Problems**
   - What assumption(s) are needed?
   - Confidence intervals vs. prediction intervals.
4. **Model Selection**
   - Describe model selection techniques available to us (mathematically and conceptually).
   - What are the strengths and limitations of each approach?
   - Be sure to understand implementation details. For example, is AIC computed on the full dataset or training set only?
   - Protect against double-dipping.
5. **MLR other considerations:**
   - Categorical predictors
   - Interaction terms - when do we need them? How do we decide if we want to include them in the model?
   - Polynomial regression
   - Model diagnostics
   - Multicollinearity
6. **Resampling methods: cross-validation and bootstrap**
   - Be sure to understand implementation details very well.
   - What is the purpose of bootstrapping?
   - Strengths and limitations?
7. **Regularization: Ridge and lasso**
   - What is the motivation for shrinkage methods?
   - Setup the models. How do we estimate parameters in regularized models (mathematically and conceptually)?
   - When do the methods break down?
   - When do the methods do particularly well?
   - Distinction between ridge and lasso - why can lasso set regression coefficients to 0?
   - Elastic net: strengths and weaknesses.
8. **Classification**

- What models are available to use to carry out classification?
- Setup the models. If needed, how do we estimate parameters in these models (mathematically and conceptually)?
- When do the methods break down?
- When do the methods do particularly well (i.e. under what scenario would we choose to use LDA over logistic regression?)
- ROC curves

9. **Tree-based Methods**
   - Basic decision trees - how are they fit? How does recursive binary splitting work?
   - Tree-cost complexity pruning: what is it? How is it implemented? How does it tie into bias-variance tradeoff?
   - Classification trees
   - Bagging + random forest: setup the model. Understand the algorithm to ensemble the trees. What are the tuning parameter(s) and how do we tune them? Statistical justification for bagging and random forest?
   - Boosting: setup the model. Understand the algorithm to ensemble the trees. What are the tuning parameter(s) and how do we tune them?
10. **Unsupervised Learning**
   - Clustering: what techniques are available to us? Strengths/limitations of each. What are the practical details that needs to be considered in implementation?
   - PCA: what is it? How does it relate to singular value decomposition (SVD)?
   - Matrix completion and data imputation.