Final exam study guide

Katie Schuler

2024-10-08

Be able to do the lab questions (labs 4 - 10), practice exam questions, and exams 1 and 2 from this semester!!

Studying resources:

- Practice exam 1, solutions
- Practice exam 2, solutions
- Exam 1, solutions
- Exam 2, solutions

1. **Sampling distribution**
    - Explore a dataset with an appropriate figure (histogram, boxplot, scatterplot) and summary statistics appropriate for the distribution.
    - Recognize uniform and Gaussian probability distributions in a plot or equation and use R's functions `d*()`, `p*()`, and `r*()` to work with these distributions
    - Explain the difference between the parameter and the paramter estimate
    - Construct the sampling distribution of a paramater estimate with `infer` and quantify the spread of the distribution with a confidence interval.
    - Understand the difference between constructing a confidence interval the standard error method vs. the percentile method.
2. **Hypothesis testing**
    - Given a set of data, implement the 3-step hypothesis testing framework **nonparametrically**: (1) Pose a null hypothesis, (2) quantify how likely a given pattern of results is under the null, and (3) determine whether to reject the null (conceptually and with the `infer` framework).
    - Given a theoretical distriubiton (e.g. t), implement the 3-step hypothesis testing framework **parametrically**.
    - Given an observed correlation, determine whether a correlation is positive, negative, or no correlation.
3. **Model specification**

- Explain the different types of models and identify appropriate scenarios for selecting each.
- Specify a model conceptually by selecting appropriate terms based on the data.
- Write the model as an equation (functional form), including the appropriate terms.
- Implement the model computationally in R (e.g. y ~ 1 + x).
- Identify the aliases of the linear model equation and understand how they represented a weighted sum of inputs.
- Explain the notions of overfitting and model complexity

4. **Applied model specification**

    - Given a set of data (and a visualization of the these data):
    - Be able to write the functional form of the model as an equation and in R.
    - Read the output of the lm function in R and determine which weight corresponds to each term in the model.
    - Determine which model is being represented by the line plotted in a graph (when given the model as an equation or R specification)
    - Explain the two common approaches to linearlizaing nonlinear models and understand how they make the problem linear.

5. **Model fitting**

    - Compare and contrast the three methods for fitting a linear model: using the lm function, iterative optimization, and the matrix approach, highlighting their differences and advantages. Interpret the output of the lm function or optim function in R, understanding what is returned and how it was acheived. Understand that the goal of model fitting is to identify the best fitting estimates of the free parameters (weights).

6. **Model accuracy**

    - Explain the concept of model accuracy and its important in evaluting model peformance.
    - Identify the components of the coefficient of determination (R^2) and how it quantifies model accuracy.
    - Descibe overfitting and the consequences it has
    - Compare simple and complex models, donsidering their inmpact on model accuracy and interpretability.
    - Apply cross-valiadation techniques to assess model accuracy and prevent overfitting.
    - Given R code from a cross validation with 'tidymodels', understand what model is being validated, how, and what the returned tibble from `collect_metrics()` is showing.

7. **Model reliability**

    - Explain the concept of model reliability and its role in assessing the stability of parameter estimates across different samples.

- Describe how bootstrapping can be used to estimate the reliability of model parameters and construct confidence intervals.
- Identify the differences between model reliability and model accuracy.
- Explain how sample size affects model reliability
- Identify the reliability estimates in the `summary()` of lm().

8. **Classification**

    - Understand the concept of classification and how it is used to predict categorical outcomes
    - Explain the logistic function and it's role in transforming linear output into a probability.
    - Identify the three components required for fitting a model using glm().
    - Fit a glm in R to perform logistic regression and interpret the output (with `glm`), with `infer`, and with `parsnip`.
    - Understand that we can apply all of our model building steps to classification by making the simple change to glm during specification.