Data Science Toolbox Question Sheet

02.2 Statistical Testing

Daniel Lawson

Block 2

Short questions

Part A: Linear Algebra for Model Selection

1. Show that:

$$\mathbb{E}\left[(y_i - \hat{y}_i)^2\right] = \text{Var}[y_i] + \text{Var}[\hat{y}_i] - 2\text{Cov}[y_i, \hat{y}_i] + [\mathbb{E}(y_i - \hat{y}_i)]^2$$

   by rearranging the formula for $\text{Var}[y_i - \hat{y}_i]$ (or otherwise).
2. Recall that $\hat{H} = \text{X}((\text{X}^T\text{X})^{-1})\text{X}^T$. Show that $tr(\hat{H}) = p$ when X is an n by p matrix and has full rank.

Part B: Interpreting testing

1. What is the difference between statistical testing and model selection?
2. What is the t-test and under which circumstances would it be appropriate to use?
3. If you intended to use a t-test for a particular problem but found that its assumptions were unsatisfied, describe an alternative test might you use.
4. Define a p-value. Define the power of a test.
5. Make a two-way contingency table of the type of error made when the hypothesis is (true/false) and the test reports (true/false).
6. In a cyber dataset consisting of counts of billions of netflow events over two different years, is it likely that the null hypothesis $\mu_1 = \mu_2$ is worth testing? What about that $\mu_1 < \mu_2$? What about for the chi-squared test?
7. What is the difference between a model-based test and a resampling test? Under which circumstances should they be used?
8. What is a permutation test? What assumption needs to be made in order to construct a valid permutation test?
9. What is a Monte-Carlo test? What advantages and limitations do they have?
10. A paper claims that a larger model will always have better predictions than a smaller model. Discuss.

11. Describe the cross validation procedure, and give a justification for its use in practical settings.
12. What are the advantages of Leave-one-out cross validation and k-fold cross validation?