Practice Exam 1 Questions

Your actual exam will have questions similar to those below with space below each question to give answers. I’ve tried to emulate that here but I haven’t taken the time to print the questions out and hand write the answers, which means there may be too much or too little space in this document. I’ll make sure an appropriate amount of space is given on the actual exam (which can also help you guage how detailed of an answer to give!)

* Please note that the example questions below are not exhaustive!
* You may notice there are no programming questions below (that is no R or python syntax at all)
* There are some pseudo code questions. Here you are writing out the logic of the process and how you would go about doing it within a programming language without worrying about the syntax or the process.
* There is very little calculation required for these questions. For most answers that involve output and reading/using it, you do not need to simplify calculations.
* If you have any other questions about the content or structure of the exam, please post to the discussion forum!

1. When considering the statistical learning paradigm, we discussed supervised and unsupervised learning. What is the major difference between supervised and unsupervised learning?

   
   
   
 

1. When considering predictive modeling as our goal, what is the difference between a regression and classification task?

   
   
   
   
   
 

1. In the statistical learning paradigm, we discussed three major goals: statistical inference, predictive modeling, and pattern finding. Give a real world example for each of these goals. Specify a possible model or method we discussed in class that would help answer the question from each real world example.

   
   
   
   
   
   
   
   
   
 

1. Consider having models characterized by flexibility. Say going from not very flexible to very flexible.
   1. What type of relationship between flexibilty and squared bias would we expect? Why?
   2. What type of relationship between flexibilty and variance would we expect? Why?
   3. What type of relationship between flexibilty and training error would we expect? Why?
   4. What type of relationship between flexibilty and test error would we expect? Why?
2. Suppose we collect data to understand how the background of graduate school applicants relates to their graduation status (graduated or failed to graduate). We collect the following information on 200 past students:
   * undergraduate major (Mathematics, Statistics, Other STEM field, Non-STEM field)
   * undergraduate GPA (a continuous 0-4 scale)
   * the time since receiving their undergraduate degree in months
   * whether or not they previously conducted academic research (yes or no).
   1. Would this be a regression or classification task?
   2. What is the response variable in this scenario?
   3. Suppose we wanted to turn the undergraduate major predictor into numeric predictors for the purposes of modeling. How many indicator (or dummy) variables are needed to account for this variable? Define these indicator variables.
   4. Considering your answer to c carefully, what are and in this scenario?

* Write down the Multiple linear regression model. Specify the meaning of each term in the model and the common assumptions we would make in order to conduct inference about the model.
  + What kind of plot might we look at to investigate the distributional assumption on the errors?
  + What kind of plot might we look at to investigate the constant variance assumption?
  + What equation do we usually optimize to ‘fit’ the model?
  + What would be the benefit of using a different function to optimize the fit?
  + Residual definition. Calculate/estimate from table or graph?
  + How can we often capture non-linear relationship between the predictors and response using an MLR model?
    - What are ways of accounting for a non-linear relationship in an MLR?
  + What is meant by least squares regression?
  + Output from an MLR model. State things (hypotheses, conclusions). Give fitted equation, equation for a prediction. What does the SE mean?
  + What does an interaction between two predictors mean?
  + WHat is a dummy variable? How do we use them to account for qualitative predictors in an MLR model?
  + Inference
    - What is meant by a standard error?
    - Interpret a CI for beta
    - Conduct a HT for beta
    - Conduct an HT for a subset of betas
    - Interpret prediction vs confidence interval
    - Why is the global F-test needed?
  + Model heredity idea
* Give an example of a nonlinear regression model.
* Regression vs Classification task
* Parametric vs non-parametric models. Advantages/disadvantages. Requirements/best choices for different situations
* overfit vs underfit, bias variance tradeoff
* Tuning or hyperparameter idea
* Curse of dimensionality idea
* Model metric idea. What is the most commonly used model metric for a regression task?
* General process of building a statistical model
  + Why do we need a training/test split?
  + Ideally response variable has the same distribution in train and test set
  + Train/test error relationship, which is likely bigger/smaller in different situations
  + Last step once final model is chosen (fit to entire data set)
  + Explain the resampling procedures (CV, bootstrap) for data splitting and why they appropriately handle the train/test issues discussed.
  + Why might we just use CV or just use bootstrap rather than a train/test split and these?
  + pseudo code for implementing some of these
  + Why do we need to avoid ‘touching’ or using the data set often?
* kNN description/idea. Perhaps finding an estimate and residual from that model
  + Use L1 distance instead
  + Small k implies, large k implies (relate to bias and variance)
  + Training MSE when k = 1?
  + Estimate when k = n?
* When is it ok to use R^2 as a model metric?
* What is LOOCV? What benefits/drawbacks does it have over 5 or 10 fold CV? (heavy computational burden, really a bias variance trade off, 5-fold is biased but LOOCV has highly correlated MSE values so we don’t get a reduction in variance from averaging the test errorrs)
* What is an out of bag observation? Why are they useful?
* What are the sources that contribute to our expected test set error?
* If we have an MLR model and are XTX matrix is not invertible (say due to a dependence in the matrix or very high multi-collinearity. What strategies do we have for dealing with that?
  + Remove highly correlated predictors, use VIF, use dimension reduction techniques, apply shrinkage methods
* What is the idea of Huber loss? Why might we prefer it to using least squares to fit an MLR model?
* Lay out a plan for selecting the best model…
* Why do we not want to use RSE to compare MLR models with differing numbers of predictors? R^2
* Interp R^2
* Adjusted R^2, AIC, BIC, Cp ideas
* BIC tends to pick smaller models
* Best subset selection process
  + Why is it ok to use R^2 or RSE to screen models?
* Forward/backward process
  + Benefits/drawbacks?
* Using holdout set or CV with these treating model size as a tuning parameter
* Regularization methods - idea
* LASSO purpose and benefits
  + Large/small tuning parameter values vs fit
* RR purpose and benefits
  + Large/small tuning parameter values vs fit
* Elastic net purpose and benefits
* Why is it important to standardize our predictors with these methods?
* One SE rule. WHy use it?
* Can you apply RR, LASSO, Elastic Net, subset selection, etc. when p>n?