Practice Exam 1 Questions - Key

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

In supervised learning we have a response variable that acts as a target for what we are modeling. In the unsupervised case we don’t have a response variable and are instead trying to understand patterns in the data.

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

A regression task has a numeric (usually continuous type) response variable. A classification task has a categorical variable (binary or more general) as the response variable.

1. 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?

* This would be classification task.
  1. What is the response variable in this scenario?
* The response variable is whether the person graduated or failed to graduate.
  1. 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.
* For a categorical variable with d categories, we need d-1 dummy variables. Here we have four levels so we need three dummy variables. The naming of these is not unique.
* The case for non-stem major is given when these indicators are all 0.
  1. Considering your answer to c carefully, what are and in this scenario?

1. Suppose we want to use a multiple linear regression model but the relationships between the predictors and response are non-linear. What are strategies for using a multiple linear regression model to account for these non-linear relationships?

We can include polynomial terms. We can transform our predictors or response in other ways (log transforms, square root, etc.). We can also include interaction terms between variables (includin polynomial terms!)

1. Give an example of a non-linear regression model.
2. What is meant by the term ‘parametric model’ vs the term ‘non-parametric model’?

A parametric model is one where we specify a stronger structure for the model. This structure generally has a few parameters that define the relationship. Fitting this model involves estimating these parameters.

Non-parameteric models are generally very flexible models that don’t have parameters that define the relationship (but may have tuning parameters of course). For instance, the KNN model takes a local average of terms. This model doesn’t have parameters to define but we do need to tune to determine an appropriate number of neighbors.

a. What are the advantages of using a parametric model over a non-parametric model? What are the disadvantages?  
  
Parametric models tend to have lower variance in their predictions and do not require as many observations to fit. They are also generally more interpretable than non-parametric models.  
  
However, they sometimes are not flexible enough to capture relationships in the data (or require us to have more of an idea of what the relationships might be in order to model the data well).   
  
b. Give an example of a situation where a non-parametric model would be a preferable choice. The example doesn't need to be in a real-world context but should be detailed enough to ensure the non-parametric model is a good choice.  
  
If we had a lot of observations and a response variable whose relationship with the predictors is highly non-linear, non-parametric models would be appropriate.

1. We discussed the terms overfit and underfit models. Describe these two terms, especially as they relate to bias and variance of the model and training and test errors.

Overfitting is where our model trains too closely to the data it is fit on. This can lead to a low training error but the model will not generalize well on the test set due to being trained to patterns that do not generalize to new data. The predictions in the model have high variance here.

Underfitting is the other extreme where we do not train closely enough to the patterns in the data. In this case our training error may be larger and the test error may be larger. The predictions from this model tend to have lower variance but larger (squared) bias.

1. Suppose we have a large data set where we want to perform a regression task. We want to determine the best overall model between a kNN model and a ridge regression model. We want to use a train test split and compare the best kNN and ridge regression model on the test set. We wish to determine the appropriate tuning parameters on the training set only using the bootstrap. Fully outline the process for splitting the data, tuning, comparing, and fitting a final overall best model.

Here we would first take the full data set and split it into a training and test set (likely 80/20 or 70/30).

Using the training set only we’ll select our tuning parameters for the two models.

For ridge regression, we create a tuning grid the penalty term (). We take a (non-parametric) bootstrap sample from the training data. Let be the number of observations in the training data. We randomly select observations from the training data with replacement. On this bootstrap sample we fit our model with one value of our tuning parameter. As we sampled with replacement, some observations were not used in the fitting of the model. These are called out-of-bag observations and we can use these observations to obtain predictions on unseen data. We find our metric on those prediction. This process is repeated B times for that tuning parameter value. We combine the test errors into one training metric for this tuning parameter value.

We now repeat this process for the next value of the tuning parameter and so on. We choose the tuning parameter for the ridge regression model that has the lowest test MSE found through this process.

For the KNN model, the process is similar. We create a grid of values for k. We then create a bootstrap sample and fit the model to that for a specific value of k. We test on the out-of-bag observations. We repeat many times and combine the test errors into one training metric for this tuning parameter value.

We repeat the process for every value of k. We choose the k that minimizes this test MSE found in this way.

We now fit our RR and KNN models to the training data with the specified tuning parameter values. These models are then tested on the test set and we pick an overall best model.

This best model is then fit to the entire data set!

1. Describe the curse of dimensionality. Where does this cause problems when doing predictive modeling?

The curse of dimensionality can mean a few things but the main way we discussed this was through the idea of having ‘a lot’ of observations for the number of predictors we have. We know we generally need more observations when fitting non-parameteric models in order to have them not overtrain to certain observations.

As our number of predictors increase, we have to increase our number of observations by a huge amount to account for the larger and larger dimension of the predictor space. While 50 observations may have been enough with one predictor, with three predictors we may need 5000 to fit it well.

1. When doing predictive modeling, what is a model metric? What is the most commonly used model metric for a regression task?

A model metric is a function that quantifies how well we do at predicting. The most commonly used metric for a regression task is MSE or RMSE.

1. When using the bootstrap for tuning or training our model, what is an out of bag observation and why are they useful?

When we do a (non-parametric) bootstrap, we randomly sample from the data with replacement. We still end up with observations but some are repeated. This means some observations are not included in the bootstrap resample. These are called out-of-bag observations.

This set of out-of-bag observations gives us a natural test set! We can see how well our model works on data it wasn’t trained on.

1. When doing a regression task and considering prediction as our goal, we discussed the expected squared test error given a particular observation,

* We were able to show this could be decomposed into three pieces. Give those three pieces (non-mathematically is fine) and describe what each term represents.

Irreducible error - this is the inherent variabilty associated with the process. We can’t lower this by having a better model.

Variance - this is the variability in the prediction made by the model.

Squared bias - this represents the amount, on average, our model misses the truth by squared.

1. Consider the multiple linear regression setting. We discussed using as a way to judge the effectiveness of a model solely on the data on which the model was trained. We saw that use of can be misleading. However, we noted that we could use in the screening step of the best subset selection procedure in order to choose the best model of each size. Fully describe this screening step of the algorithm and explain why the use of here is defensible.

Suppose we have predictors. We want to screen to find a possible ‘best’ model with predictors in it. Here we can fit each of the possible models with predictors and pick the best one via its value.

We can use to compare models for a given as these models all have the same number of predictors! This avoids the issues with always increasing, even with nonsense predictors being included.

1. Suppose we are fitting a kNN model for a regression task.
   1. Describe how leave-one-out cross-validation (LOOCV) can be used in this setting to choose .

* We can create a grid of values. For a given , we can train the model on of the observations and test that fitted model on the observation left out.
* We can repeat this process, leaving each observations out once. We can then combine the test errors and get an overall test error for that .
* This process is repeated for each value of interest.
* We compare the test errors. The with the lowest test error is the chosen value.
  1. What are two drawbacks of LOOCV as compared to 5 or 10 fold cross-validation?
* LOOCV is computationally expensive! For each tuning parameter value we have to fit models where as 5 fold CV requires us to fit only five models.
* Another drawback is that the predictions from the fitted models are more highly correlated. This implies we don’t get as much of a reduction in variance of our test error prediction as we get when doing five or ten fold CV.

1. Consider the LASSO procedure for fitting a multiple linear regression model. With this model we minimize the following criterion (recall ):
   1. What are the benefits of fitting a LASSO model as compared to an ordinary least squares model?

* LASSO models shrink coefficients. This results in biased estimates but reduces variability. This can improve our test MSE.
* LASSO models also set some coefficients to 0 for large enough values of . This means it also does variable selection.
  1. What happens to our coefficient estimates for a ‘large’ value of the tuning parameter? What happens for a tuning parmeter value near 0?
* For large values we are greatly penalizing the size of the regression coefficients. This means they must be shrunk more (and some will be set to 0). For a value near 0, we get something close to the OLS fit.

1. Why is it important to standarize our predictor values when doing a ridge regression, LASSO, or elastic net model?

We need to standardize the predictors as the relative size of the estimates will depend on the scale of the corresponding predictor value. By standardizing our values we are able to weight our predictors more evenly.

1. When fitting a regularized or penalized regression model such as the LASSO we discussed using the ‘one-standard error’ method for selecting the tuning parameter. Explain this idea and describe the effect of using this method on the selected model.

We know that the test error is a variable quantity. As such, minimizing it will change from data set to data set. We can often look at the variability of the test error. This comes naturally when we do CV. We have many measurements of the test error and so we can understand its variability.

In an effort to have a more simplistic model, we can choose the tuning parameter for a more simple model that gives a test error within one standard error of the minimum. This means we are still getting a model close to the model with the lowest error but the model will be more penalized so that resulting model should be simpler (say fewer variables in the LASSO or elastic net setting).

1. We perform best subset, forward stepwise, and backward stepwise selection on a single data set with predictors. For each approach we obtain models containing. One model containing 0 predictors, one containing 1 predictor, etc.
   1. Which of the three models with exactly predictors has the smallest training RSS or are we unable to tell?

* This problem is answered in the HW 2 solutions. Those will post on Thursday morning. Please check those out!
  1. Which of the three models with exactly predictors has the smallest test RSS or are we unable to tell?
* This problem is answered in the HW 2 solutions. Those will post on Thursday morning. Please check those out!
  + Indicate whether the following statements are true or false
    1. The predictors in the -variable model identified by **forward** stepwise selection are a subset of the predictors in the -variable model identified by **forward** stepwise selection.
    2. The predictors in the -variable model identified by **backward** stepwise selection are a subset of the predictors in the -variable model identified by **backward** stepwise selection.
    3. The predictors in the -variable model identified by **backward** stepwise selection are a subset of the predictors in the -variable model identified by the **best subset** method.
    4. The predictors in the -variable model identified by **best subset** selection are a subset of the predictors in the -variable model identified by the **best subset** method.
  + This problem is answered in the HW 2 solutions. Those will post on Thursday morning. Please check those out!

1. Suppose we fit a multiple linear regression model to data about how much people earn. Our response variable is the wage (in 1000’s of dollars) and our predictors are marital\_status (married, never\_married, or divorced), age, and year that the data was collected. We include an interaction between marital\_status and age in the model. Output for the model is given below.

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | -2526.362 | 722.229 | -3.498 | 0.000 |
| marital\_statusmarried | 26.527 | 11.997 | 2.211 | 0.027 |
| marital\_statusnever\_married | -13.320 | 12.465 | -1.069 | 0.285 |
| age | 0.489 | 0.236 | 2.078 | 0.038 |
| year | 1.299 | 0.360 | 3.608 | 0.000 |
| marital\_statusmarried:age | -0.205 | 0.250 | -0.817 | 0.414 |
| marital\_statusnever\_married:age | 0.321 | 0.280 | 1.146 | 0.252 |

* 1. Write down the fitted equation for . Define any indicator variables as needed.
* where takes on 1 if the marital status is married and 0 otherwise, takes on 1 if the marital status is never married and 0 otherwise.
  1. One column of the output represents the standard error. What is a standard error generally?
* A standard error is a measure of the variability associated with an estimate. It is the (estimated) standard deviation of the value.
  1. Write down the form of a predicted value for someone that is married, has an age of 30, and had a year of data collection of 2008. No need to simplify.
  2. Write down the form of a predicted value for somone that is divorced, has an age of 30, and had a year of data collection of 2008. No need to simplify.
  3. Consider creating a confidence interval for the mean wage as compared to a prediction interval for a future wage. Which would be wider? Why?
* The prediction interval is wider as we are trying to capture a future response there. The confidence interval is trying to capture mean response. There is more variability in a predicted value so the interval must be wider.
  1. Conceptually, what does including an interaction between marital\_status and age do to our model as compared to a model without that interaction (that still includes main effects for both, and an effect for year)?
* The interaction effect allows the model to fit different slopes for the age variable. Here we have a second numeric variable so really we are getting a different slope in the direction of age depending on the marital status.
  1. Based on the above output, which of the **three** predictors would you deem important for the model? State the hypotheses and p-values you used to make this determination.
* The age, year and marital status variable are all important. We can see this from the tests of year ( that slope is 0 vs not zero, p-value 0.000), age (similar test, p-value 0.038), and the test for the marital status married indicator (null is that ‘slope’ (really an intercept modifier here) is 0 vs it is not zero, p-value 0.027). This test indicates that this level differs from the baseline level, implying this variable is important.
  1. Suppose we wanted to check whether the interaction term between marital\_status and age was important for the response, given the other variables in the model. Describe how we could investigate this (you can’t tell with the output above and I’m not expecting you to give the exact formulas - describe the idea of the inferential method we looked at in class for answering this question).
* You can do the (nested) F-test we looked at in the notes where we compare the sums of squares for the model without either of those terms corresponding to the interaction to the sum of squares for the model that includes them.
  1. What type of plot might we look at to investigate the distributional assumption on the errors?
* A qqplot!
  1. A few observed values and predicted values are given in the table below. What is the residual for the first observation. No need to simplify.

| wage | predicted |
| --- | --- |
| 75.04315 | 81.10693 |
| 70.47602 | 83.36921 |
| 130.98218 | 115.28895 |

* Residual = observed - predicted = 75.04315 - 81.10693