Answer the questions below. When finished, export this notebook as a pdf or html file, and then submit that file to the appropriate assignment in the Google classroom.

1. What are the three sources of error for a given model?

σ^2: Irreducible error.

$Var[\hat{f}]$: Variance of the model

$Bias^2[\hat{f}]$: The squared bias of the model.

2. Can these error sources be measured or observed directly? If they can, how does one measure them? If they cannot, how do we ensure that all sources of error are minimized?

Irreducible error can not be reduced, as the name implies.

The bias and the variance can not be observed directly. To ensure that all sources of error are minimized, we can use cross validation to tune the hyperparameters and check the overall error using the diferent combinations of hyperparameters we use in the cross validation process.

3. In your own words, explain each source of error.

Irreducible error comes from the data itself, especially from the actual process for collecting the data.

The variance comes from making the model too complex, which can make the model especially sensitive to the parameters of the model. The increased sensitivity of the model means that any slight changes to the parameters will have a larger effect on the model, which will increase the differences between model output, which means the model results will have a wider distribution of results.

Bias in the model comes from over-simplifying aspects of the model. If key aspects of the data are not represented in the model, then the model will systematically get incorrect results by a consistent amount. The consistent amount that the over simplified model is due to how the missing key features of data would affect the ideal model.

4. Which source of error is associted with "overfitting" a model, and which is associated with "underfitting?"

Overfitting is associated with variance; it makes the model overly complex and sensitive.

Underfitting is associated with too much bias; the model is overly simple.

5. What is a hyperparameter?

A hyperparameter is a parameter that is used in fitting the model, but is not an actual weight solved for in the model. Examples of hyperparameters are the learning rate, the number of epochs, the number of features, etc...

6. Describe the k-folds cross validation framework.

When performing the k-folds cross validation, you randomize the observations, then split them up into 'k' as close to equal sized data sub-sets. Going through each k subset of data, you train with the entire remaining non-k groups of data with your model and set parameters, then test the model on the k^{th} subset of data. You would ideally go through this process many times, each time tuning your hyperparameters to find the best values for each.

7. Describe the Train - Validate - Test framework.

The train-validate-test framework randomizes the observations, then splits the data into three groups. The general rule of thumb is that 60% is used to train, 20% is used to validate, and 20% is used to test. First, you solve for the weights using the training data only, then you tune a hyperparemeter with the validation data only, then you test the model on the testing data only. This process will have to be repeated multiple times for each hyperparemeter in order to find the best values for each one.

8. What does it mean for a data set to be "imbalanced," and how do we take this into consideration during cross validation?

If a set of data is imbalanced, it will have a relatively low percentage of the observations that represent either the target or a specific feature of the data. In order to rectify this, the entire data set is split into the different classifications and the stratified data is then equally split up into the different subsets of data used for the cross validation process. This ensures that the imbalanced data is represented through each step of the cross-validation process.

9. When applying L^2 Regularization to a model, how does λ_2 affect the model parameters? That is, what is the effect on the model parameters, and by what mechanism is this effect achieved?

In L^2 regulariztion, the λ_2 parameter smoothes the model over by scaling down the weights. This is acheived by tacking on a term to least squares function so that when it is optimized, it penalizes values with large weights. With the λ_2 term, we have

$$J = \frac{1}{2N}(y - \hat{y})^T(y - \hat{y}) + \frac{\lambda_2}{2N} w^T w$$

.

10. From the probabilistic perspective, explain how L^2 Regularization introduces bias into a model.

Both J and $F = e^{-J}$ will have the same minimum, because they are monotonic increasing functions. Because of this, you will find, when you expand F, that it becomes a product of two products,

$$F = \prod_{i=1}^{N} e^{-(y_i - \hat{y}_i)^2} \prod_{j=1}^{P} e^{-\lambda_2 w_j^2}$$

.

The factor with λ_2 looks an aweful lot like a Gaussian distribution with a mean of zero and a $\sigma^2 = \frac{1}{2\lambda_2}$. Because of this, $\sigma^2 \propto \frac{1}{\lambda_2}$, which means that increasing λ_2 will decrese σ^2 and decreasing λ_2 will increase σ^2. Interpreting it this way, the second factor could be seen as the probability of measuring a specific weight 'w', or $p(w)$.

The first factor also looks similar to a Gaussian distribution that could be interpreted as the probability of geting a y, given $\hat{y} = Xw$, or $p(y|X, w)$.

Putting these two factors together, we have $F \propto p(y|X, w)p(w)$. From Baye's theorem, we can now see that $F \propto p(w|X, y)$, which is the probability of geting weight values, given the features of the data and the target data.

If our total error, $E[(y - \hat{f})] = Irreducible\ Error + Var[\hat{f}] + Bias^2[\hat{f}]$, then we can see, if we kept the total error constant, that increasing the variance would reduce the bias and decreasing the variance would increase the bias. Therefore, if we increase λ_2, then the variance would decrease and the bias would increase. Vice versa, decreasing λ_2 would increase variance and decrease bias.