DS 100: Principles and Techniques of Data Science **Date: March 23, 2018**

Discussion #8

Name:

Bias Variance Tradeoff and Regularization

1. What happens to the bias, validation error and test error as the regularization parameter λ increases? Draw a picture.

2. As model complexity increases, what happens to the bias-variance tradeoff?

3. Ridge regression is a variant of least squares that involves regularization. It is defined as follows:

$$\min_{\vec{\theta}} L(\vec{\theta}) = \min_{\vec{\theta}} ||\vec{y} - X\vec{\theta}||_2^2 + \lambda||\vec{\theta}||_2^2 = \min_{\vec{\theta}} \sum_{i=1}^{n}(y_i - \vec{x_i}^T\vec{\theta})^2 + \lambda\sum_{j=1}^{d}\theta_j^2$$

Here, λ is a hyper parameter that determines the impact of the regularization term. X is a $n \times d$ matrix, $\vec{\theta}$ is a $d \times 1$ vector and $\vec{y}$ is a $n \times 1$ vector. Find the optimal $\vec{\theta}^*$.

4. How does the bias-variance tradeoff of a ridge regression estimator compare with that of ordinary least squares regression?

5. In ridge regression, what happens if we set $\lambda = 0$? What happens as λ approaches ∞?

6. If we have a large number of features (10,000+) and we suspect that only a handful of features are useful, which type of regression (Lasso vs Ridge) would be more helpful in interpreting useful features?

7. What are the benefits of using ridge regression?

Cross Validation

8. Describe the k-fold cross validation procedure and why we might use it in developing models.

9. We are computing the loss over our data/predictions using squared loss with the Lasso regularization function:

$$\min_{\vec{\theta}} \sum_{i=1}^{n}(y_i - \vec{x_i}^T\vec{\theta})^2 + \lambda\sum_{j=1}^{d}|\theta_j|$$

In order to implement k-fold cross validation, we can run the following pseudocode:

```
for lambda in lambdas:
    for fold in folds:
        calculate MSE Lasso(X_test[fold], X_train[fold],
            Y_train[fold], Y_train[fold], lambda)
```

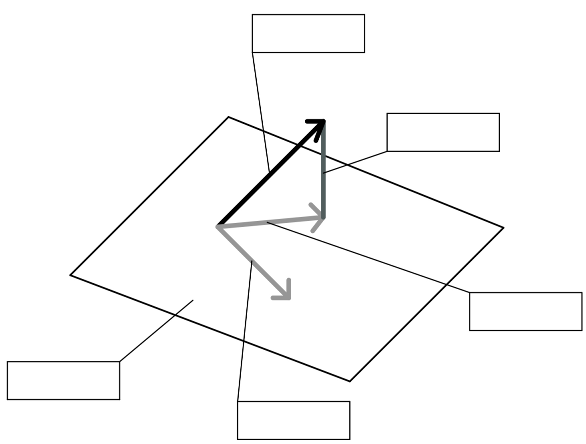
After running k-fold cross validation, we get the following mean squared errors for each fold and value of λ:

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	80.2	70.2	91.2	91.8	83.4
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0
4	79.4	68.4	92.3	92.4	83.1
5	77.3	67.3	93.4	94.3	83.0
Col Avg	79.0	68.8	90.4	93.2	

Based on these results, what parameter for λ should we use? Explain.

Geometric Interpretation of Linear Regression

10. Draw the geometric interpretation of the column space of the design matrix, the response vector ($\vec{y}$), the residuals, and the predictions.



11. From the image above, what can we say about the residuals and the column space of X? Write this mathematically and prove this statement (note: we can use linear algebra or summations)

12. Derive the normal equations from the fact above.

13. What must be be true about ϕ for the normal equation to be solvable? What does this imply about the features we select?

14. What does this imply about the dimension of the design matrix?