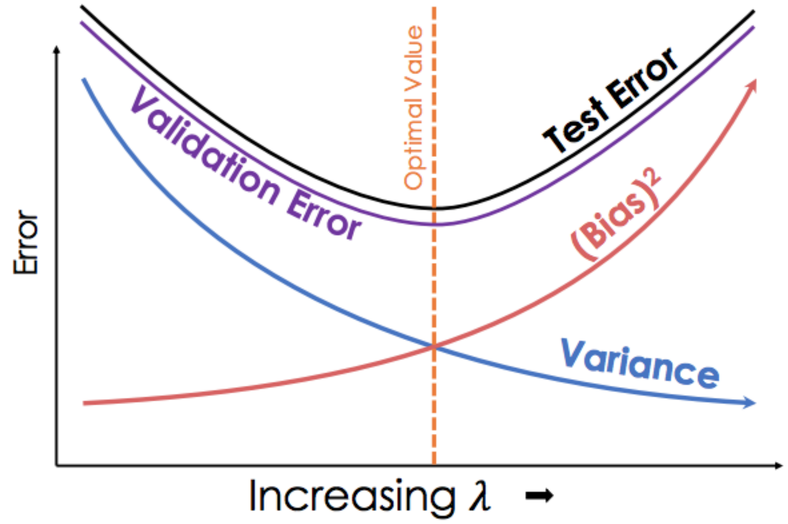
DS 100: Principles and Techniques of Data Science **Date: March 23, 2018**

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.



Solution:

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

Solution: Model complexity is inversely related to the regularization parameter λ. Bias tends to decrease and variance tends to increase.

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 $\vec{\theta}^*$.

Solution:

Notation:

$$\text{X} = \begin{bmatrix} — & x_1^T & — \\ — & x_2^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix}$$

$$\begin{aligned} L(\vec{\theta}) &= ||\vec{y} - X\vec{\theta}||_2^2 + \lambda||\vec{\theta}||_2^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \vec{x}_i^T\vec{\theta})^2 + \lambda\sum_{j=1}^{d}\theta_j^2 \\ \nabla\sum_{i=1}^{n}(y_i - \vec{x}_i^T\vec{\theta})^2 &= -2\sum_{i=1}^{n}(y_i - \vec{x}_i^T\vec{\theta})\vec{x}_i \\ &= -2\sum_{i=1}^{n} y_i\vec{x}_i - \vec{x}_i^T\vec{\theta}\vec{x}_i \\ &= -2X^T\vec{y} + 2X^TX\vec{\theta} \\ \nabla\sum_{j=1}^{d}\theta_i^2 &= 2\vec{\theta} \\ \nabla L(\vec{\theta}) &= -2X^T\vec{y} + 2X^TX\vec{\theta} + 2\lambda\vec{\theta} = 0 \end{aligned}$$

Solving for theta, we see that $\theta = (X^TX + \lambda I)^{-1}X^T\vec{y}$

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

Solution: Ridge regression has higher bias and lower variance relative to ordinary least squares regression.

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

Solution: If we set $\lambda = 0$ we end up with OLS. As λ approaches ∞ then θ goes to 0.

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?

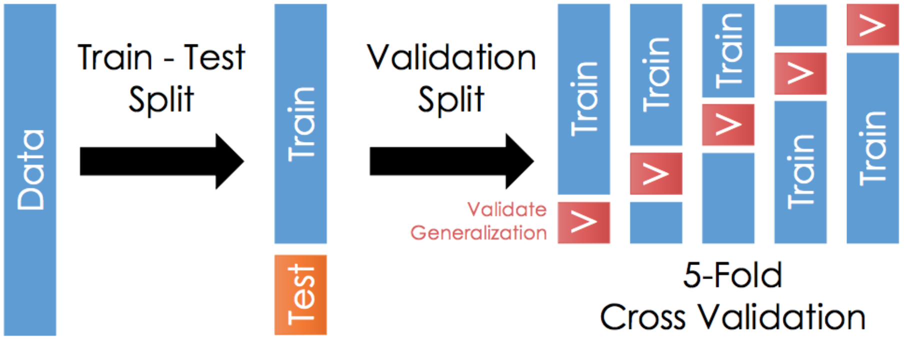
Solution: LASSO would be better as it sets many values to 0, so it would be effectively selecting useful features and forgetting bad ones.

7. What are the benefits of using ridge regression?

Solution: If multiple features are correlated, weight can be shared across those features. If X^TX is not full rank (not invertible), then we end up with infinitely many solutions for least squares. But if we use ridge regression, $\theta = (X^TX + \lambda I)^{-1}X^TY$. This guarantees invertbility and a unique solution, for $\lambda > 0$.
If you're interested why there are infinitely many solutions (underdetermined systems):
http://pages.cs.wisc.edu/ amos/412/lecture-notes/lecture17.pdf

Cross Validation

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



Solution:

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:

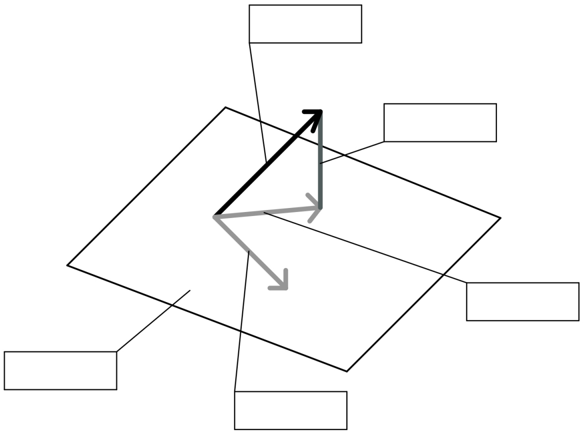
Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	80.2	70.2	91.2	91.8	83.4
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0
4	79.4	68.4	92.3	92.4	83.1
5	77.3	67.3	93.4	94.3	83.0
Col Avg	79.0	68.8	90.4	93.2	

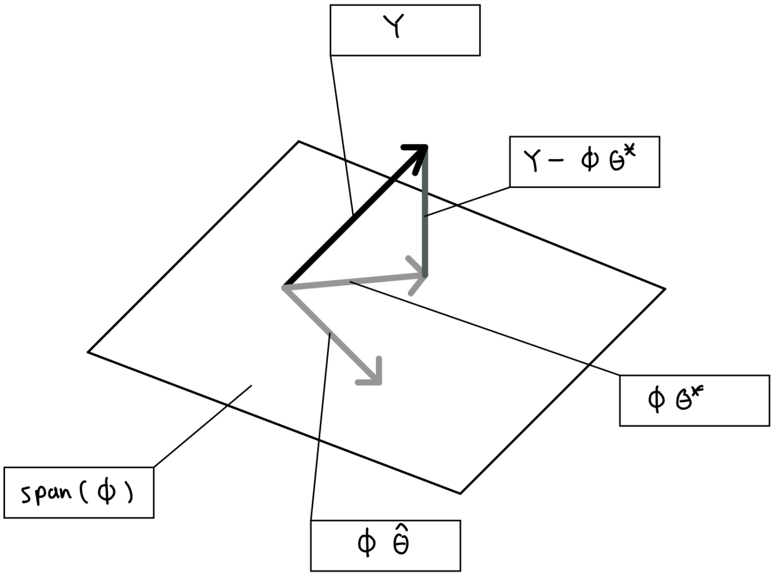
Based on these results, what parameter for λ should we use? Explain.

Solution: We should use $\lambda = 0.2$ because this value has the least average MSE across all folds.

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.





Solution:

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)

Solution: We can say that the residuals are orthogonal to the column space of ϕ Mathematically $\phi^T(Y - \phi\theta) = 0$. We can prove this fact using summations by taking the derivative of the squared loss for some simple linear regression model ax + b. The fitted/predicted y-values are given by:

$$\hat{y}_i = f_\theta(x_i) = \theta_1 + \theta_2 x_i$$

We take the derivative with respect to θ_2 and set it equal to 0:

$$0 = \frac{\partial}{\partial\theta_2}\sum_{i=1}^{n}(y_i - (\theta_1 + \theta_2 x_i))^2 = -2\sum_{i=1}^{n}(y_i - (\theta_1 + \theta_2 x_i)) * x_i$$

We know that $(y_i - (\theta_1 + \theta_2 x_i))$ is the residual; hence, the residuals are orthogonal to the span of the x values.

12. Derive the normal equations from the fact above.

Solution: $\phi^T(Y - \phi\theta) = 0$ We can distribute and move rearrange terms: $\phi^T Y = \phi^T\phi\theta$ We can left multiply by $(\phi^T\phi)^-1$ in order to get the least squares estimator for θ :
$\theta = (\phi^T\phi)^(-1)\phi^T Y$

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

Solution: The design matrix must be invertible; hence, no linearly dependent features.

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

Solution: In order for the columns to be linearly independent, it must be true that the number of columns of the design matrix must be less than or equal to the number of rows. Mathematically, $n >= d$ where n is the number of data points and d is the number of features.