DS 100: Principles and Techniques of Data Science **Date: October 3, 2018**

Discussion #6 Solutions

Name:

Bias-Variance Tradeoff

1. Let X be a random variable with mean $\mu = \mathbb{E}[X]$. Using the definition $\text{Var}(X) = \mathbb{E}[(X-\mu)^2]$, show that for any constant c,

$$\mathbb{E}[(X-c)^2] = (\mu - c)^2 + \text{Var}(X).$$

Solution: One way to show this is to write $X - c = X - \mu + \mu - c$. Squaring both sides,

$$\mathbb{E}[(X-c)^2] = \mathbb{E}[(X-\mu)^2 + (\mu-c)^2 + 2(X-\mu)(\mu-c)]$$

Now using linearity of expectation and pulling out the constants,

$$\begin{aligned}\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mu)^2] + (\mu-c)^2 + 2\underbrace{\mathbb{E}[X-\mu]}_{=0}(\mu-c) \\ &= \text{Var}(X) + (\mu-c)^2.\end{aligned}$$

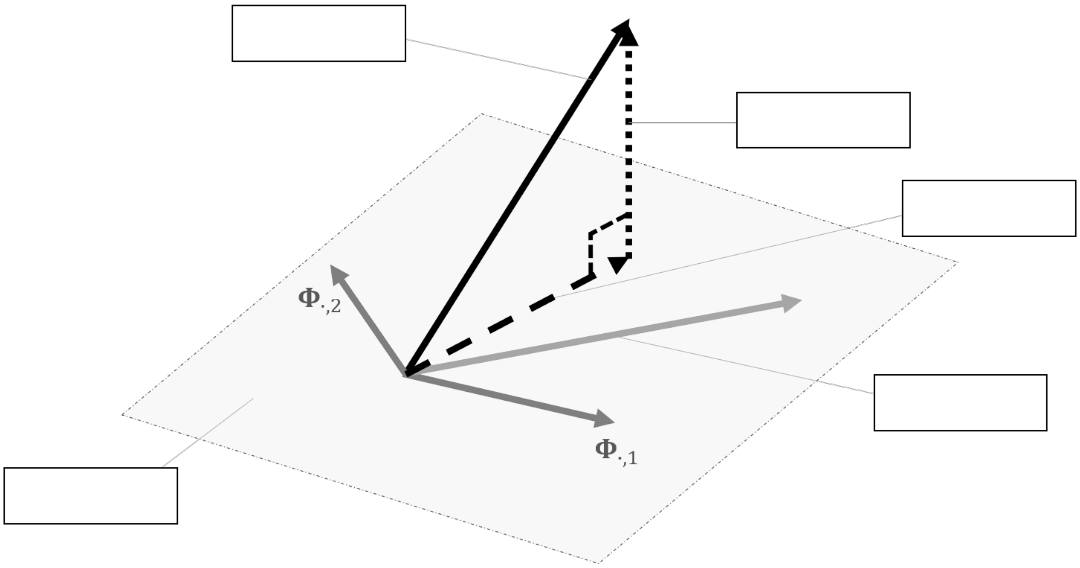
2. Use the above result to prove that

- $\text{Var}(X) \leq \mathbb{E}[(X-c)^2]$ for any c
- $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

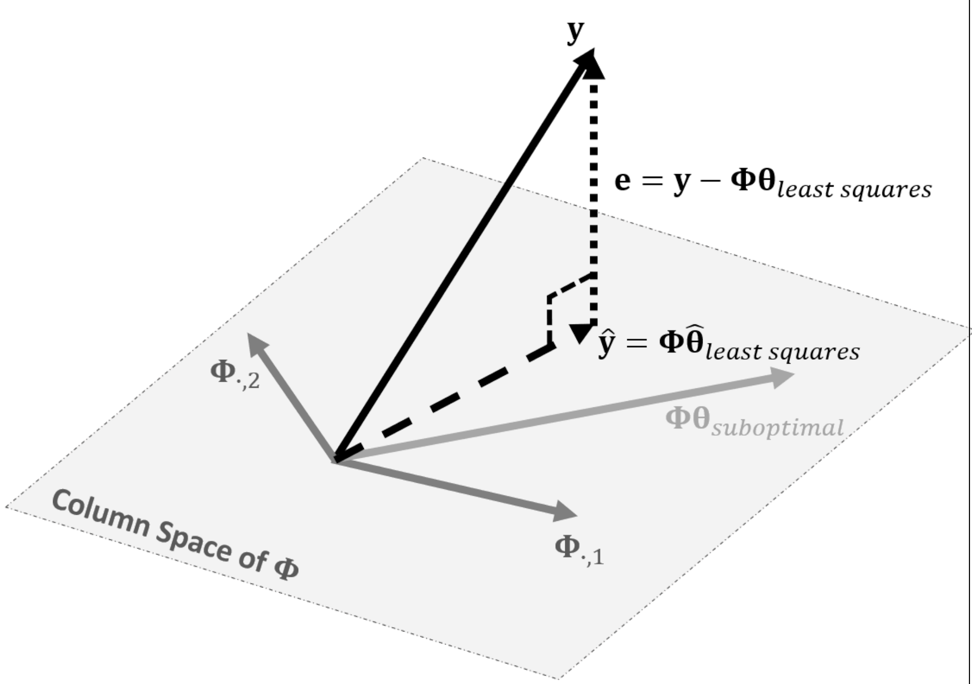
Solution: The first bullet follows from using $(\mu - c)^2 \geq 0$, and the second bullet follows from plugging in $c = 0$.

Geometry of Least Squares

3. The following question will refer to the diagram below:



(a) Fill in the diagram of the geometric interpretation of 1) the column space of the design matrix, 2) the response vector ($\mathbf{y}$), 3) the residuals and 4) the predictions



Solution:

(b) From the image above, what can we say about the residuals and the column space of $\mathbf{\Phi}$? Write this mathematically and prove this statement with a calculus-based argument and a linear-algebra-based argument.

Solution: We can say that the residuals are orthogonal to the column space of $\mathbf{\Phi}$. Mathematically $\mathbf{\Phi}^T\mathbf{e} = \mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\theta) = 0$.
Recall that the estimator $\hat{\theta}$ linear regression is trying to solve this problem:

$$\hat{\theta} \in \underset{\theta}{\operatorname{argmin}}\, (\mathbf{y} - \mathbf{\Phi}\theta)^T\, (\mathbf{y} - \mathbf{\Phi}\theta)$$

Since this is a convex function, the minimizing θ is where the gradient is zero.

$$\begin{aligned}
0 &= \nabla \left[(\mathbf{y} - \mathbf{\Phi}\theta)^T\, (\mathbf{y} - \mathbf{\Phi}\theta)\right] \\
&= \nabla \left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{\Phi}\theta - \theta^T\mathbf{\Phi}^T\mathbf{y} + \theta^T\mathbf{\Phi}^T\mathbf{\Phi}\theta\right) \\
&= \nabla \left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{\Phi}\theta + \theta^T\mathbf{\Phi}^T\mathbf{\Phi}\theta\right) \\
&= 0 - 2(\mathbf{y}^T\mathbf{\Phi})^T + \left[(\mathbf{\Phi}^T\mathbf{\Phi}) + (\mathbf{\Phi}^T\mathbf{\Phi})^T\right]\theta \\
&= -2\mathbf{\Phi}^T\mathbf{y} + 2\mathbf{\Phi}^T\mathbf{\Phi}\theta \\
&\implies \mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\theta) = 0
\end{aligned}$$

Alternatively, let V be a vector space equipped with an inner product $(\cdot, \cdot)$ and $W \subseteq V$ be a subspace of V. In the linear regression setting: $V = \mathbb{R}^k$ for some $k \in \mathbb{N}$. $W = colspace(\mathbf{\Phi})$. Consider a vector $\mathbf{y} \in V$
Suppose we have $\hat{\mathbf{y}} \in W$ such that $\mathbf{y} - \hat{\mathbf{y}} \perp W$. Given an arbitrary $\mathbf{w} \in W$, we have:

$$\|\mathbf{y} - \mathbf{w}\|^2 = \|(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \mathbf{w})\|^2 = \|(\mathbf{y} - \hat{\mathbf{y}})\|^2 + \|(\hat{\mathbf{y}} - \mathbf{w})\|^2 \geq \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

Since this is true for any $\mathbf{w}$, we have that $\hat{\mathbf{y}}$ is the minimizer. Furthermore, since $\hat{\mathbf{y}} \in W$, then there exists some $\hat{\theta} \in \mathbb{R}^d$ such that $\hat{\mathbf{y}} = \mathbf{\Phi}\hat{\theta}$.

(c) Derive the normal equations from the fact above.

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

(d) Let $\mathbf{\Phi}$ be a $n \times p$ design matrix with full column rank. In this question, we will look at properties of matrix $H = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ that appears in linear regression.

i. Recall for a vector space V that a projection $\mathbf{P} : V \to V$ is a linear transformation such that $\mathbf{P}^2 = \mathbf{P}$. Show that $\mathbf{H}$ is a projection matrix.

Solution:

$$\begin{aligned}\mathbf{H}^2 &= \left(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\right)\left(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\right)\\ &= \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}(\mathbf{\Phi}^T\mathbf{\Phi})(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\\ &= \mathbf{\Phi}\mathbf{I}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\\ &= \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\\ &= \mathbf{H}\end{aligned}$$

ii. This is often called the "hat matrix" because it puts a hat on $\mathbf{y}$, the observed responses used to train the linear model. Show that $\mathbf{Hy} = \hat{\mathbf{y}}$

Solution: $\mathbf{Hy} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T y = \mathbf{\Phi}\left[(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T y\right] = \mathbf{\Phi}\hat{\theta} = \hat{\mathbf{y}}$

iii. Show that $\mathbf{M} = \mathbf{I} - \mathbf{H}$ is a projection matrix.

Solution: $\mathbf{M}^2 = (\mathbf{I}-\mathbf{H})^2 = \mathbf{I}^2-\mathbf{IH}-\mathbf{HI}+\mathbf{H}^2 = \mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{H} = \mathbf{I}-\mathbf{H} = \mathbf{M}$

iv. Show that $\mathbf{My}$ results in the residuals of the linear model.

Solution: $\mathbf{My} = (\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{y} - \mathbf{Hy} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}$

v. Prove that $\mathbf{H} \perp \mathbf{M}$

Solution: $\mathbf{HM} = \mathbf{H}(\mathbf{I}-\mathbf{H}) = \mathbf{H} - \mathbf{H}^2 = \mathbf{H} - \mathbf{H} = \mathbf{0}$

vi. Notice that the hat matrix is a function of our observations $\mathbf{\Phi}$ rather than our response variable $\mathbf{y}$. Intuitively, what do the values in our hat matrix represent? It might be helpful to write $\hat{y}_i$ as a summation.

Solution: Using the fact that $\hat{\mathbf{y}} = \mathbf{Hy}$, we can write the ith prediction as $\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j$. From this, we can see that the values of the hat matrix allow us to understand how much each observation influences our prediction of $\hat{y}_i$. Moreover, $h_i i$ is a measure of
The diagonal elements of $\mathbf{H}$ denoted as h_{ii} are called leverages. The larger a point's leverage is, the more influence the point has on the regression predictions.

(e) Suppose $\mathbf{\Phi} \in \mathbb{R}^{n \times d}$ does not have full column rank. Then $\mathbf{\Phi}^T\mathbf{\Phi}$ is not invertible. Why is that? Complete the argument below:

i. Recall that the null space $N(\mathbf{\Phi})$ of a matrix $\mathbf{\Phi}$ is defined as all the vectors that get sent to 0 by $\mathbf{\Phi}$ i.e.

$$N(\mathbf{\Phi}) = \{\mathbf{x} \mid \mathbf{\Phi}\mathbf{x} = \mathbf{0}\}$$

Show that the null space of $\mathbf{\Phi}$ is a subset of the null space of $\mathbf{\Phi}^T\mathbf{\Phi}$.

Solution: Let $\mathbf{x} \in N(\mathbf{\Phi})$. Then

$$\begin{aligned} & \mathbf{\Phi}\mathbf{x} = \mathbf{0} \\ \Longrightarrow\ & \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{x} = \mathbf{0} \\ \Longrightarrow\ & \mathbf{x} \in N(\mathbf{\Phi}^T\mathbf{\Phi}) \end{aligned}$$

ii. Show that the reverse inclusion is also true i.e. that $N(\mathbf{\Phi}^T\mathbf{\Phi}) \subseteq N(\mathbf{\Phi})$

Solution: Let $\mathbf{x} \in N(\mathbf{\Phi}^T\mathbf{\Phi})$. Then

$$\begin{aligned} & \mathbf{\Phi}^T\mathbf{\Phi}\mathbf{x} = \mathbf{0} \\ \Longrightarrow\ & \mathbf{x}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{x} = 0 \\ \Longrightarrow\ & (\mathbf{\Phi}\mathbf{x})^T(\mathbf{\Phi}\mathbf{x}) = \|\mathbf{\Phi}\mathbf{x}\|^2 = 0 \\ \Longrightarrow\ & \mathbf{\Phi}\mathbf{x} = \mathbf{0} \\ \Longrightarrow\ & \mathbf{x} \in N(\mathbf{\Phi}) \end{aligned}$$

We can then conclude that $N(\mathbf{\Phi}^T\mathbf{\Phi}) = N(\mathbf{\Phi})$, which implies $dim(N(\mathbf{\Phi}^T\mathbf{\Phi}) = dim(N(\mathbf{\Phi}))$. By the rank-nullity theorem, $rank(\mathbf{\Phi}^T\mathbf{\Phi}) = rank(\mathbf{\Phi})$. Thus if $rank(\mathbf{\Phi}) < d$, then $rank(\mathbf{\Phi}^T\mathbf{\Phi}) < d$. But $\mathbf{\Phi}^T\mathbf{\Phi} \in \mathbb{R}^{d \times d}$, so there's no hope for invertibility.

iii. List some reasons why $\mathbf{\Phi}$ might not have full column rank.

Solution:

- There are fewer observations than features i.e. $n < d$
- Some features are linear combinations of others e.g. gross income, expenses, and net profit are all included in the design matrix

Regularization

4. In a petri dish, yeast populations grow exponentially over time. In order to estimate the growth rate of a certain yeast, you place yeast cells in each of n petri dishes and observe the population y_i at time x_i and collect a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Because yeast populations are known to grow exponentially, you propose the following model:

$$\log(y_i) = \beta x_i \tag{1}$$

where β is the growth rate parameter (which you are trying to estimate). We will derive the L_2 regularized estimator least squares estimate.

(a) Write the *regularized least squares loss function* for β under this model. Use λ as the regularization parameter.

Solution:

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i) - \beta x_i)^2 + \lambda\beta^2 \tag{2}$$

(b) Solve for the optimal $\widehat{\beta}$ as a function of the data and λ.

Solution: Taking the partial derivative of the regularized loss function function:

$$\begin{aligned}
\frac{\partial}{\partial\beta}L(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}(\log(y_i) - \beta x_i)^2 + \frac{\partial}{\partial\beta}\lambda\beta^2 \\
&= -\frac{2}{n}\sum_{i=1}^{n}(\log(y_i) - \beta x_i)x_i + 2\lambda\beta \\
&= -\frac{2}{n}\sum_{i=1}^{n}\log(y_i)x_i + \frac{2}{n}\sum_{i=1}^{n}\beta x_i^2 + 2\lambda\beta \\
&= -\frac{2}{n}\sum_{i=1}^{n}\log(y_i)x_i + \frac{2}{n}\beta\left(\sum_{i=1}^{n}x_i^2 + \lambda n\right)
\end{aligned}$$

Setting the derivative equal to zero and solving for β:

$$\begin{aligned}
0 &= -\frac{2}{n}\sum_{i=1}^{n}\log(y_i)x_i + \frac{2\beta}{n}\left(\lambda n + \sum_{i=1}^{n} x_i^2\right) \\
\beta\left(\lambda n + \sum_{i=1}^{n} x_i^2\right) &= \sum_{i=1}^{n}\log(y_i)x_i \\
\beta &= \left(\lambda n + \sum_{i=1}^{n} x_i^2\right)^{-1}\sum_{i=1}^{n}\log(y_i)x_i
\end{aligned}$$