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

Discussion #6

Name:

Bias-Variance Tradeoff

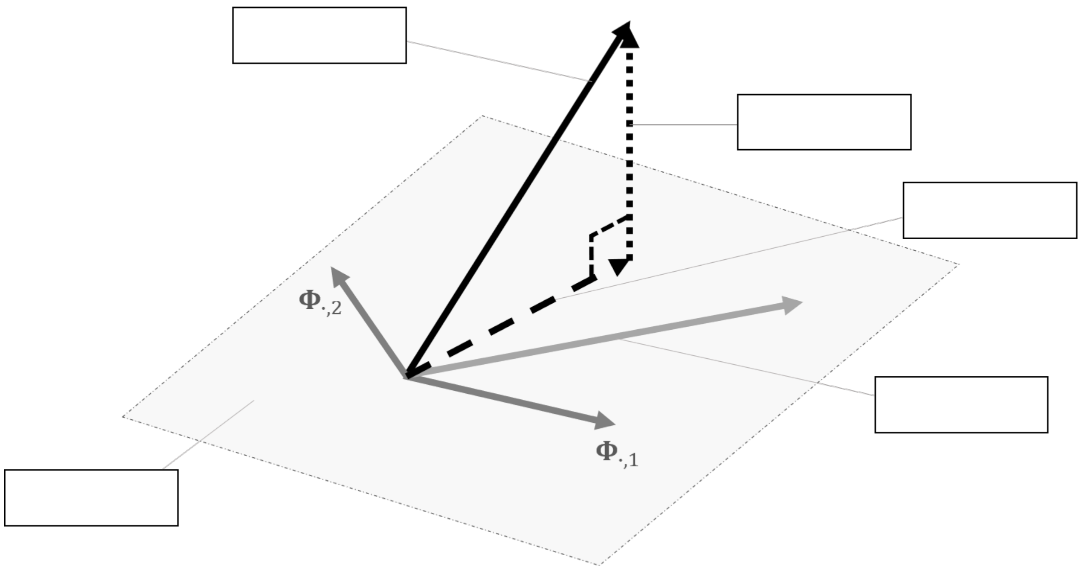
1. Let X be a random variable with mean $\mu = \mathbb{E}[X]$. Using the definition $\text{Var}(X) = \mathbb{E}[(X-\mu)^2]$, show that for any constant c,

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

2. Use the above result to prove that
   - $\text{Var}(X) \leq \mathbb{E}[(X-c)^2]$ for any c
   - $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

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

(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.

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

(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.

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 $\mathbf{Hy} = \hat{\mathbf{y}}$

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

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

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

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.

(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 $\mathbf{\Phi}^T\mathbf{\Phi}$.

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

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.

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.

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