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

Discussion #6

Name: Kagan

surajrampure.com

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,

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

$$E[(X-c)^2] = E[X^2 - 2Xc + c^2]$$
$$= E[X^2] - 2c\,\underbrace{E[X]} + c^2$$
$$= \underbrace{E[X^2] - \mu^2} + \underbrace{\mu^2 - 2c\mu + c^2}$$
$$= var(X) + (\mu - c)^2$$

$c = \mu$ is optimal value that minimizes $E[(X-c)^2]$

2. Use the above result to prove that

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

both non-neg.

$$E[(X-c)^2] = var(X) + (\mu - c)^2$$
$$\geq var(X)$$

alt: $(\mu - c)^2 \geq 0$

$$\underbrace{var(X) + (\mu-c)^2} \geq var(X)$$
$$E[(X-c)^2] \geq var(X)$$

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

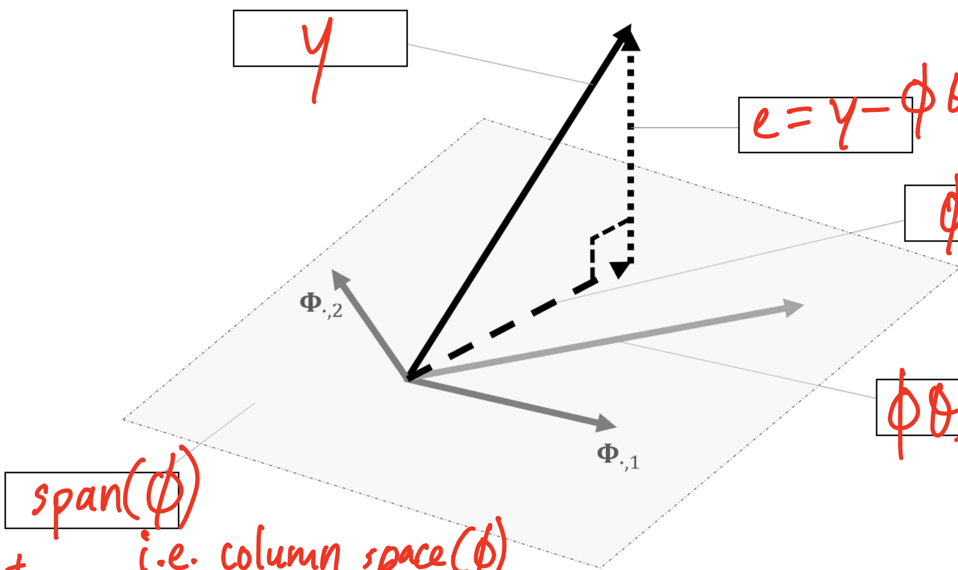
Geometry of Least Squares

$X = \begin{bmatrix} 1 & 4 \\ 2 & 7 \end{bmatrix}$

$\text{span}(X): a\begin{bmatrix} 1 \\ 2 \end{bmatrix} + b\begin{bmatrix} 4 \\ 7 \end{bmatrix}$

$a, b \in \mathbb{R}$

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



projection of y onto col space of ϕ

span(ϕ)

$\phi\theta_{suboptimal}$

$\phi\theta_{optimal}$

y

$e = y - \phi\theta_{opt}$

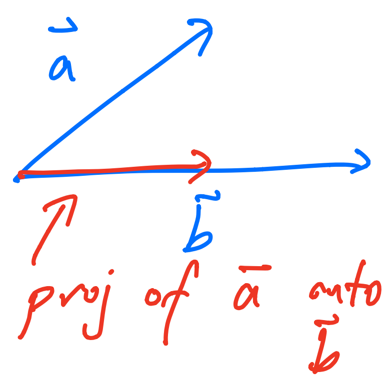
(a) Fill in the diagram of the geometric interpretation of 1) the column space of the design matrix, 2) the response vector ($\mathbf{y}$), 3) the residuals and 4) the predictions

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

two vec.s orthogonal:
$a^T b = 0$
$(a \cdot b = 0)$

error is $\perp$ to all ϕ_i

$$\phi_1^T(y - \phi\theta) = 0$$
$$\phi_2^T(y - \phi\theta) = 0$$
$$\vdots$$
$$\phi_n^T(y - \phi\theta) = 0$$



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

$$\rightarrow \quad \phi^T(y - \phi\theta) = 0$$

ϕ_i : vector

ϕ : matrix

$$\phi^T y - \phi^T\phi\,\theta = 0$$

$$\phi^T\phi\,\theta = \phi^T y \longrightarrow \boxed{\theta_{opt} = (\phi^T\phi)^{-1}\phi^T y}$$

calculus derivation on last page

LOOK AT SOLUTIONS FOR THIS PAGE

(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 $\mathbf{H}\mathbf{y} = \hat{\mathbf{y}}$

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

iv. Show that $\mathbf{M}\mathbf{y}$ 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 $\mathbf{\Phi}^T\mathbf{\Phi}$.

$x \in N(\phi) \rightarrow \phi x = 0$

$\phi^T \phi x = 0$

$\rightarrow x \in N(\phi^T\phi)$

$\therefore N(\phi) \subseteq N(\phi^T\phi)$

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

$x \in N(\phi^T\phi) \rightarrow \phi^T\phi x = 0$

$\cancel{(\phi^T)^{-1}}\,\cancel{\phi^T}\phi x = (\phi^T)^{-1} 0$

$\phi x = 0$

$\rightarrow x \in N(\phi)$

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.

training data → create model

→ use model to make predictions

model is sensitive to outliers in training

Regularization

overfitting : model won't generalize well to other data

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:

$$\log(y_i) = \beta x_i \qquad \hat{y} = \beta x \tag{1}$$

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.

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

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

$$\frac{\partial L}{\partial \beta} = \frac{1}{n}\sum_i 2\left(\log y_i - \beta x_i\right)(-x_i) + 2\lambda\beta = 0$$

$$\vdots$$

$$\boxed{\beta_{opt} = \frac{\sum \log(y_i)\, x_i}{\lambda n + \sum x_i^2}}$$

"ordinary least squares"

OLS

$$L(\theta) = \|y - X\theta\|^2$$

$$\theta_{opt} = (X^T X)^{-1} X^T y$$

"hyperparameter"

Regularized LS

$$L(\theta) = \|y - X\theta\|^2 + \lambda \|\theta\|^2$$

prevents overfitting

Φ instead of X

y : vec

X : matrix

θ : vec

Calculus Derivation of $\theta_{opt} = (X^T X)^{-1} X^T y$

Recall from last week:

$\nabla_x a^T x = a$ ①

$\nabla_x x^T A x = (A + A^T) x$

(if $A = A^T$: $= 2Ax$) ②

Also: $\|a\|^2 = a^T a$

$$L(\theta) = \|y - \phi\theta\|^2 = (y - \phi\theta)^T (y - \phi\theta)$$

$$= (y^T - (\phi\theta)^T)(y - \phi\theta)$$

$$= y^T y - \underbrace{y^T \phi\theta - (\phi\theta)^T y}_{} + (\phi\theta)^T \phi\theta$$

since $y^T(\phi\theta)$ and $(\phi\theta)^T y$ are both dot products of the same two vectors, they're equal

(think $a^T b = b^T a$)

$$= y^T y - 2y^T \phi\theta + \theta^T \phi^T \phi\theta$$

$$= \underbrace{y^T y}_{\text{ind. of } \theta} - 2(\phi^T y)^T \theta + \theta^T \phi^T \phi \theta$$

① ②

look above to see which grad. rules used

(2) $\phi^T\phi$ is symmetric, $\therefore \nabla_x \theta^T \phi^T \phi \theta = 2\phi^T \phi\theta$

$$\frac{\partial L}{\partial \theta} = -2\phi^T y + 2\phi^T \phi\theta = 0$$

$$\rightarrow \phi^T \phi \theta = \phi^T y$$

$$\rightarrow \boxed{\theta = (\phi^T \phi)^{-1} \phi^T y}$$

as we saw earlier! two derivations of same thing