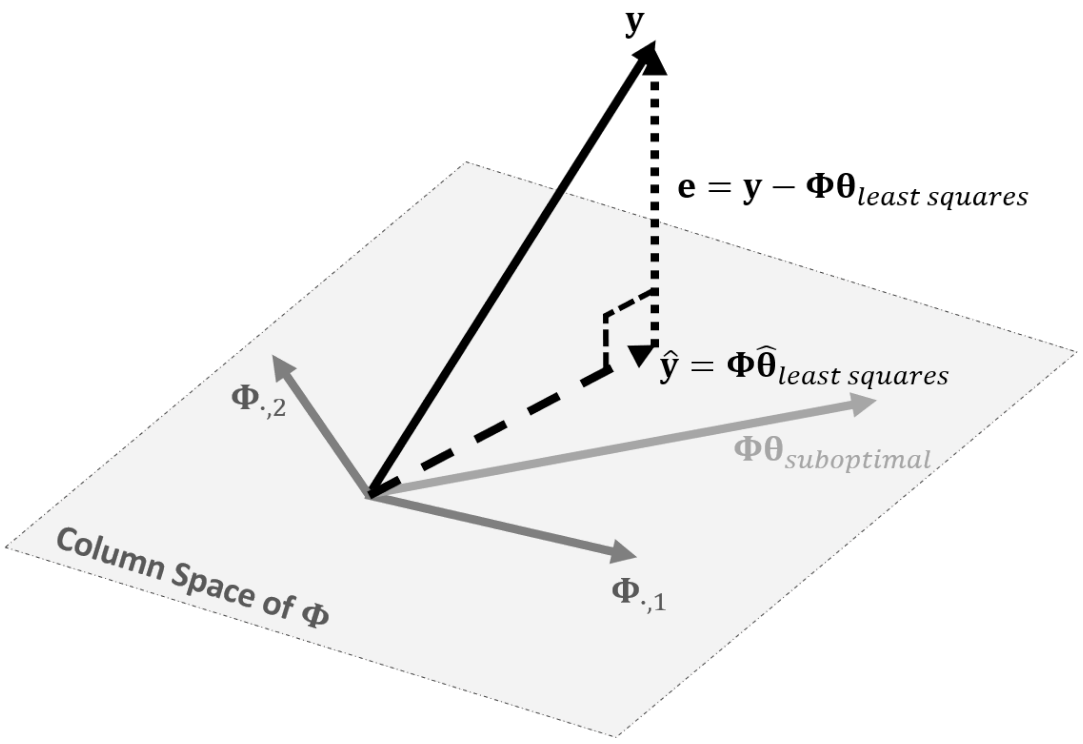
DS 100/200: Principles and Techniques of Data Science **Date: March 15, 2019**

Discussion #7 Solutions

Name:

Geometry of Least Squares

1. This diagram shows the geometry of 3 observations with 2 features. Φ_1 is the column vector of the three values for feature 1, and Φ_2 is the column vector of values for feature 2. We're fitting a model with parameters θ, a two-element vector, that determines a linear combination of the 2 features. A choice of θ gives fitted values for the 3 observations, and these fitted values are always in the column space of Φ. The observed y, a vector of the response values for the 3 observations, is not in the column space of Φ. The least-squares choice for θ is the one for which $\Phi\theta$ is closest to y. This diagram is analogous to a setting with more observations and more features.



(a) 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 using a calculus-based argument about minimizing the linear regression loss function.

Solution: We can say that the residuals are orthogonal to the column space of $\mathbf{\Phi}$. Mathematically $\mathbf{\Phi}^T\mathbf{e} = \mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\theta) = 0$. Note that we use $\mathbf{\Phi}^T$ to match dimensions because $\mathbf{\Phi}$ is 3x2 and $\mathbf{y}$ is 3x1. This equation implies that the dot product of each column of $\mathbf{\Phi}$ and $\mathbf{e}$ is 0 since $\mathbf{\Phi}_{:,i}^T\mathbf{e} = 0$ for i between 1 and 3.

Recall that the estimator $\hat{\theta}$ linear regression is trying to solve this problem:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{\Phi}\theta\|_2^2 = \underset{\theta}{\operatorname{argmin}} (\mathbf{y} - \mathbf{\Phi}\theta)^T (\mathbf{y} - \mathbf{\Phi}\theta)$$

Since this is a convex function, the minimizing θ is where the gradient is zero. Note that the θ in this solution is the same as $\hat{\theta}_{\text{least squares}}$ in the diagram.

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

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 $\hat{\mathbf{y}} = \mathbf{\Phi}\hat{\theta}$. Note that we know $\|a + b\|_2^2 = \|a\|_2^2 + \|b\|_2^2$ because a and b are orthogonal.

(b) Show that $\theta = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T Y$. from the fact above for the least squares solution $\mathbf{\Phi}$.

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: $\theta = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T Y$.

(c) Let $\mathbf{\Phi}$ be a $n \times p$ design matrix with full column rank (the rank is equal to the number of columns). 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. 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 $\hat{y}_i$.
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 y_i has on the regression prediction $\hat{y}_i$.

(d) We can show that $rank(\mathbf{\Phi}) = rank(\mathbf{\Phi}^T\mathbf{\Phi})$ by showing that these two matrices have the same null space. List some reasons why $\mathbf{\Phi}$ might not have full column rank, which would make $\mathbf{\Phi}^T\mathbf{\Phi}$ not invertible.

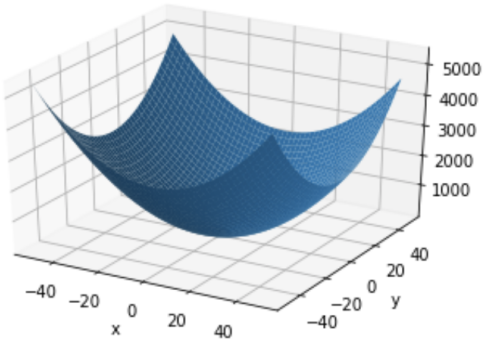
Solution:

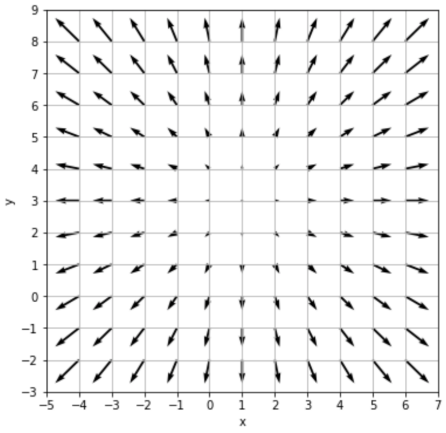
- There are fewer observations than features i.e. $n < d$

- Some features are linear combinations of others e.g. gross income, expenses, and net profit are all included in the design matrix

Gradients

2. On the left is a 3D plot of $f(x, y) = (x-1)^2 + (y-3)^2$. On the right is a plot of its gradient field. Note that the arrows show the relative magnitudes of the gradient vector.



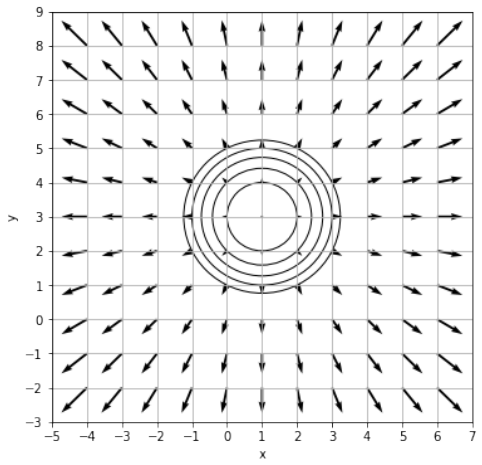


(a) Is this function convex? Make a visual argument—it doesn't have to be formal.

Solution: Yes. From looking at the surface, connecting any two points remains above the surface. Additionally, the second derivative is positive in x and y.

(b) Superimpose a contour plot of this function for $f(x, y) = 0, 1, 2, 3, 4, 5$ onto the gradient field.

Solution: The contour plots are concentric circles centered at $(1, 3)$ with radii of $0, \sqrt{1}, \sqrt{2}, \sqrt{3}, \sqrt{4}, \sqrt{5}$.



(c) What do you notice about the relationship between the level curves and the gradient vectors?

Solution: The gradient vectors increase in magnitude as the level curves increase. Additionally, the gradient vectors always lie perpendicular to the level curves, as they represent the direction in which the function curves away from each level set.

(d) From the visualization, what do you think is the minimal value of this function and where does it occur?

Solution: Since $(x-1)^2$ and $(y-3)^2$ are both nonnegative, the minimum function value of $f(x,y)$ is attained when both are equal to zero. This occurs at $(1,3)$ where the gradient field shows the smallest (in magnitude) vectors.

(e) Calculate the gradient $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T$.

Solution:

$$\begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T = \begin{bmatrix} 2(x-1) & 2(y-3) \end{bmatrix}^T.$$

(f) When $\nabla f = \mathbf{0}$, what are the values of x and y?

Solution:

$$\nabla f = \mathbf{0} \implies 2(x-1) = 2(y-3) = 0 \implies x = 1,\ y = 3.$$

If the gradient is equal to zero, then the function must be at a local minima. The only minima in this case is the global minima, meaning it must be at $(1,3)$, due to part (e).

3. In this question, we will explore some basic properties of the gradient.

   Note: In this class, we use the following conventions:

   - x represents a scalar
   - X represents a random variable
   - $\mathbf{x}$ represents a vector
   - $\mathbf{X}$ represents a matrix or a random vector (context will tell)

   (a) Determine the derivative of $f(x) = a_0 + a_1 x$ and gradient of $g(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2$.

   Solution:

   $$\frac{df}{dx} = a_1$$

   $$\nabla g = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 \end{bmatrix}^T$$

   (b) Suppose $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix}^T$, and $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$, where $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$. Determine ∇h.

   Solution: Note that $h(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is just a concise way of writing

   $$h(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i = a_1 x_1 + a_2 x_2 + ... + a_n x_n$$

   So as in (a), we have

   $$\nabla h = \begin{bmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & ... & \frac{\partial h}{\partial x_n} \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 & ... & a_n \end{bmatrix}^T = \mathbf{a}$$

   (c) Determine the gradient of $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$. *(Hint: f is a scalar-valued function. How can you write $\mathbf{x}^T\mathbf{x}$ as a sum of scalars?)*

   Solution: $f(\mathbf{x})$ can also be expanded as $\sum_{i=1}^{n} = x_i^2 = x_1^2 + x_2^2 + ... + x_n^2$

   $$\nabla f = \begin{bmatrix} 2x_1 & 2x_2 & ... & 2x_n \end{bmatrix}^T = 2\mathbf{x}$$