DS 100: Principles and Techniques of Data Science **Date: July 24, 2019**

Discussion #9

Name:

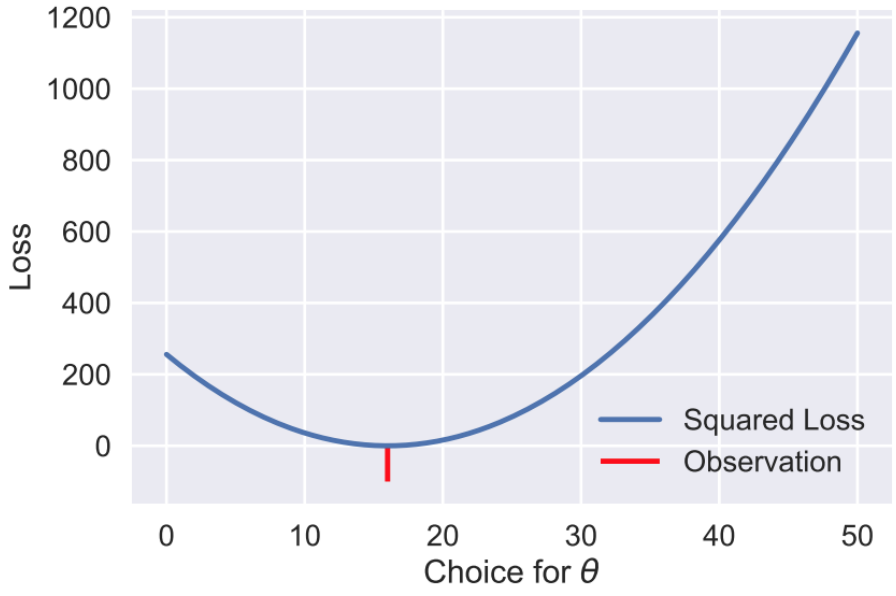
Loss Functions

1. Recall the loss functions discussed during lecture. θ represents our estimate of our parameter, and y represents a data point. Discuss the advantages and drawbacks of each of the following loss functions:

   (a) Squared loss: $L(\theta, y) = (y - \theta)^2$

   Solution:

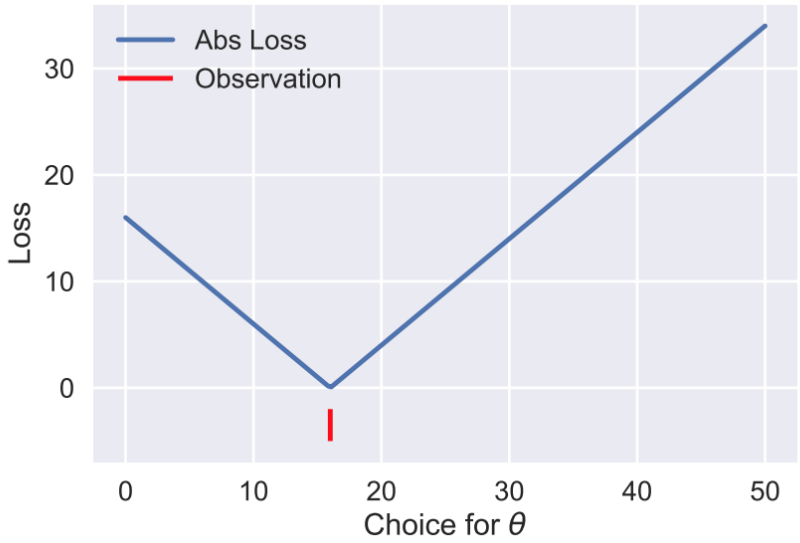
   - Commonly used loss function.
   - Sensitive to outliers.
   - Differentiable, hence we can find solution analytically.



   (b) Absolute Loss: $L(\theta, y) = |y - \theta|$

Solution:

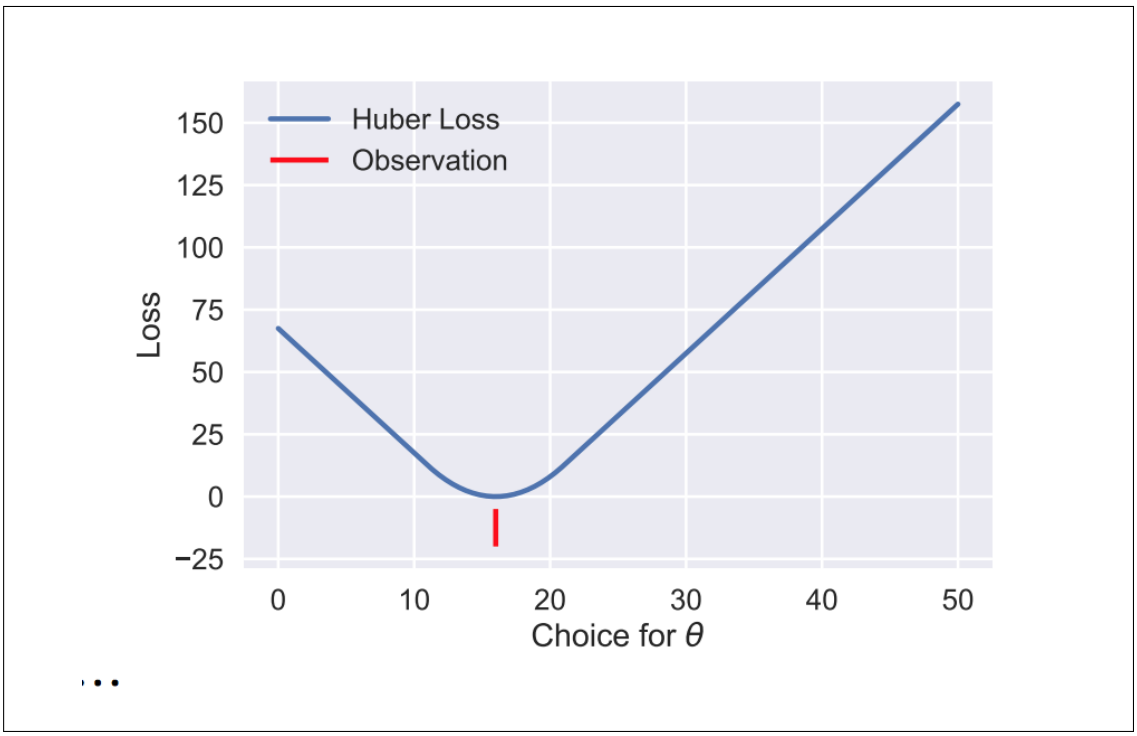
- Less sensitive to outliers.
- Not differentiable; hence, no analytic closed form solution.



(c) Huber Loss: $L_\alpha(\theta, y) = \begin{cases} (y-\theta)^2 & |y-\theta| < \alpha \\ \alpha(|y-\theta| - \alpha/2) & \text{otherwise} \end{cases}$

Solution:

- Melding of both square and absolute loss. More sensitive to small deviations from θ than squared loss but less sensitive to outlier values.
- Differentiable; however, we cannot find an analytic solution because we cannot isolate the θ term
- Must choose where the hinges occur (another tuning parameter!)



Loss Minimization

2. Consider the following loss function:

$$L(\theta, x) = \begin{cases} 4(\theta - x) & \theta \geq x \\ x - \theta & \theta < x \end{cases}$$

Given a sample of $x_1, ..., x_n$, find the optimal θ that minimizes the the average loss.

Solution:

$$\frac{\partial L(\theta, x)}{\partial \theta} = \begin{cases} 4 & \theta \geq x \\ -1 & \theta < x \end{cases}$$

$$\sum_{i=1}^{n} \frac{\partial L(\theta, x_i)}{\partial \theta} = \sum_{\theta < x_i} -1 + \sum_{\theta \geq x_i} 4 = 0$$

Which implies that

$$\#\{x_i > \theta\} = 4\#\{x_i \leq \theta\}$$

We also know that

$$\#\{x_i > \theta\} + \#\{x_i \leq \theta\} = n$$

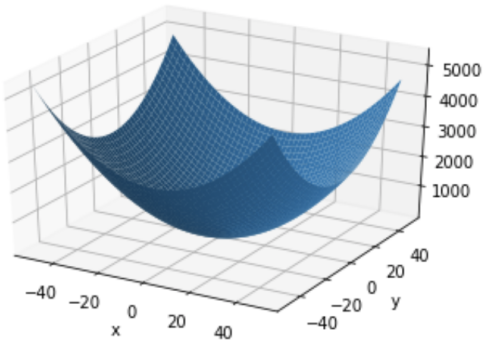
Substituting the first equation into the second, we get

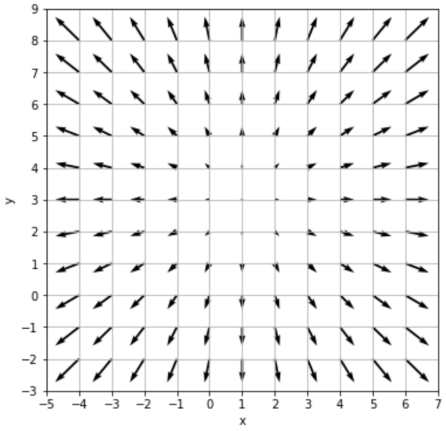
$$n - \#\{x_i \leq \theta\} = 4\#\{x_i \leq \theta\} \implies 0.20n = \#\{x_i \leq \theta\}$$

Thus, θ is the 20th percentile of $x_1, ..., x_n$.

Gradients

3. 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) 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.

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

Solution:

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

(c) 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).

4. 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 $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^T$, 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

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

   So as in (a), we have

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

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

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

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

Gradient Descent Algorithm

5. Given the following loss function and $\mathbf{x} = (x_i)_{i=1}^n$, $\mathbf{y} = (y_i)_{i=1}^n$, θ^t, explicitly write out the update equation for θ^{t+1} in terms of x_i, y_i, θ^t, and α, where α is the step size.

$$L(\theta, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\theta^2 x_i^2 - log(y_i) \right)$$

Solution:

$$\theta^{t+1} \leftarrow \theta^t - \alpha \frac{\partial L}{\partial \theta}\bigg|_{\theta=\theta^t}$$

$$\frac{\partial L}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} 2\theta x_i^2$$

6. (a) In your own words, describe how to use the update equation in the gradient descent algorithm.

   (b) Say that x and y are your model parameters and f as defined in question 1 is your loss function. Describe in your own words what happens "visually" as the gradient descent algorithm runs.

Convexity

Convexity allows optimization problems to be solved more efficiently and for global optimums to be realized. Mainly, it gives us a nice way to minimize loss (i.e. gradient descent). There are three ways to informally define convexity.

a. Walking in a straight line between points on the function keeps you above the function. This works for any function.

b. The tangent line at any point lies below the function (globally). The function must be differentiable.

c. The second derivative is non-negative everywhere (aka "concave up" everywhere). The function must be twice differentiable.

7. Find a counterexample for the claim that the composition of two convex functions is also convex. $h = g(f(x))$

Solution: Let $f(x) = x^2, g(x) = -x$. $g(f(x)) = -x^2$ which is not convex.