DS 100: Principles and Techniques of Data Science **Date: March 2, 2018**

Discussion #6

Name:

Loss Functions

1. Recall the loss functions discussed during lecture.Discuss the advantages and drawbacks of each of the following loss functions:

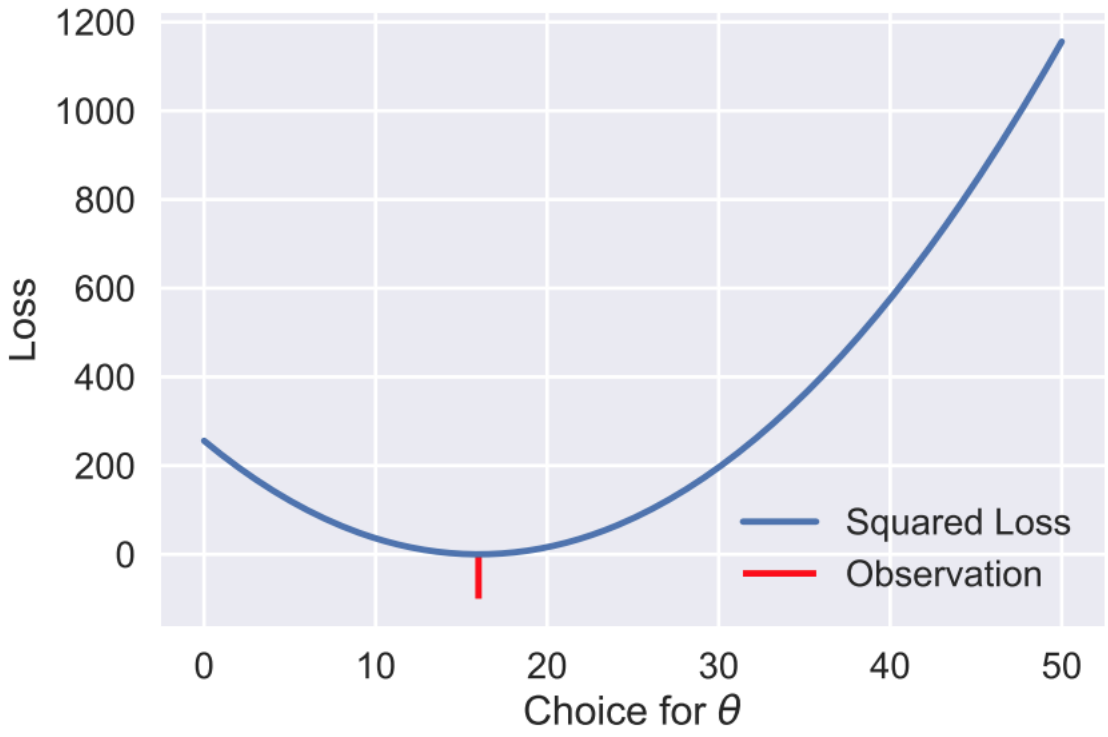
   a Squared loss: $L(\theta, y) = (y - \theta)^2$

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

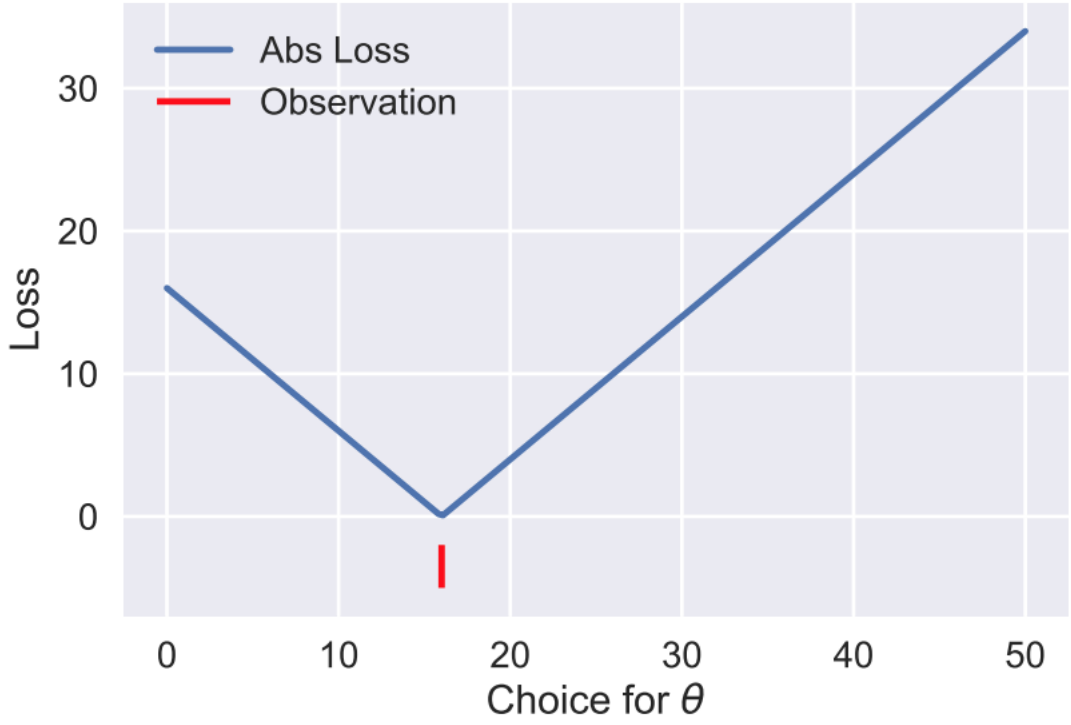
   c Huber Loss:

   $$L(\theta, y) = \begin{cases} (y-\theta)^2 & |y-\theta| < \alpha \\ \alpha(|y-\theta| - \alpha/2) & \text{otherwise} \end{cases}$$

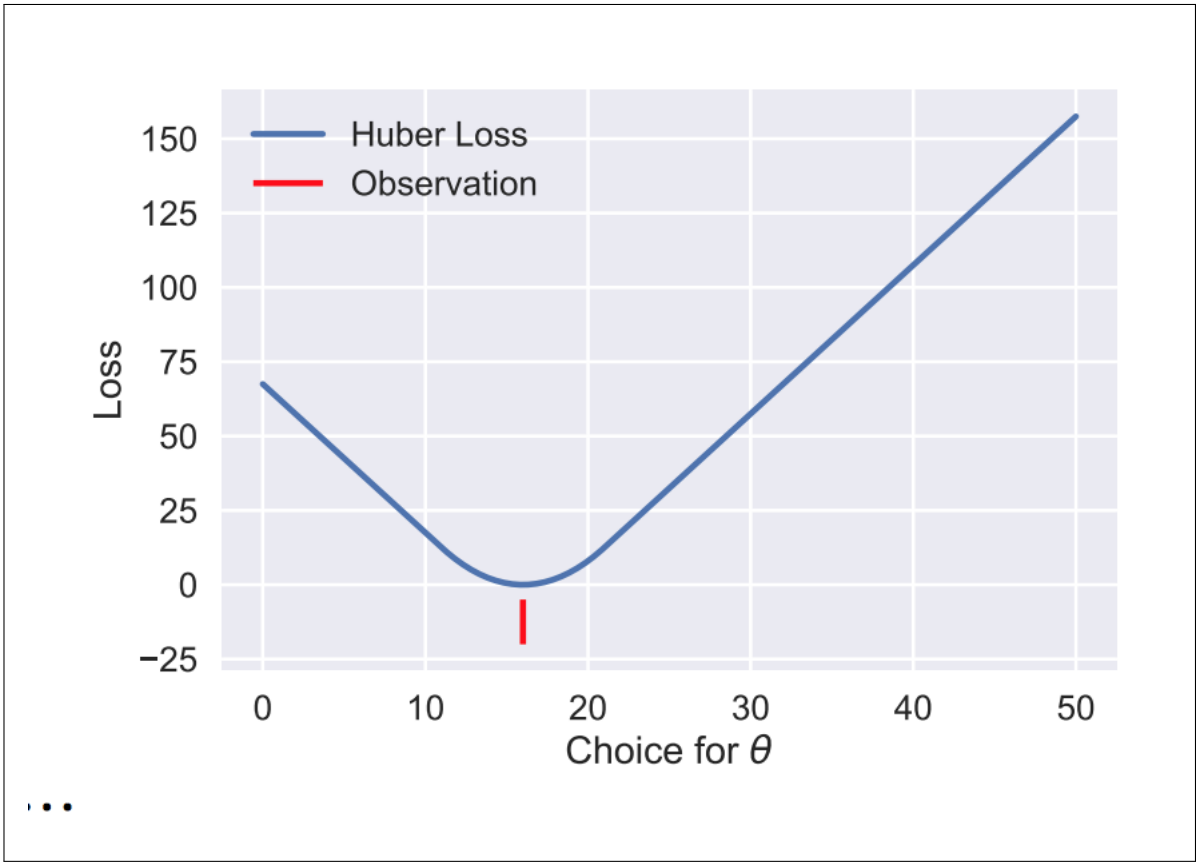
Solution: Squared Loss: Commonly used loss function. Sensitive to outliers. Differentiable, hence we can find solution analytically.



Absolute Loss: Less sensitive to outliers. Not differentiable; hence, no analytic closed form solution.



Huber Loss: Best of both worlds! Differentiable; however, we cannot find an analytic solution because we cannot isolate the θ term



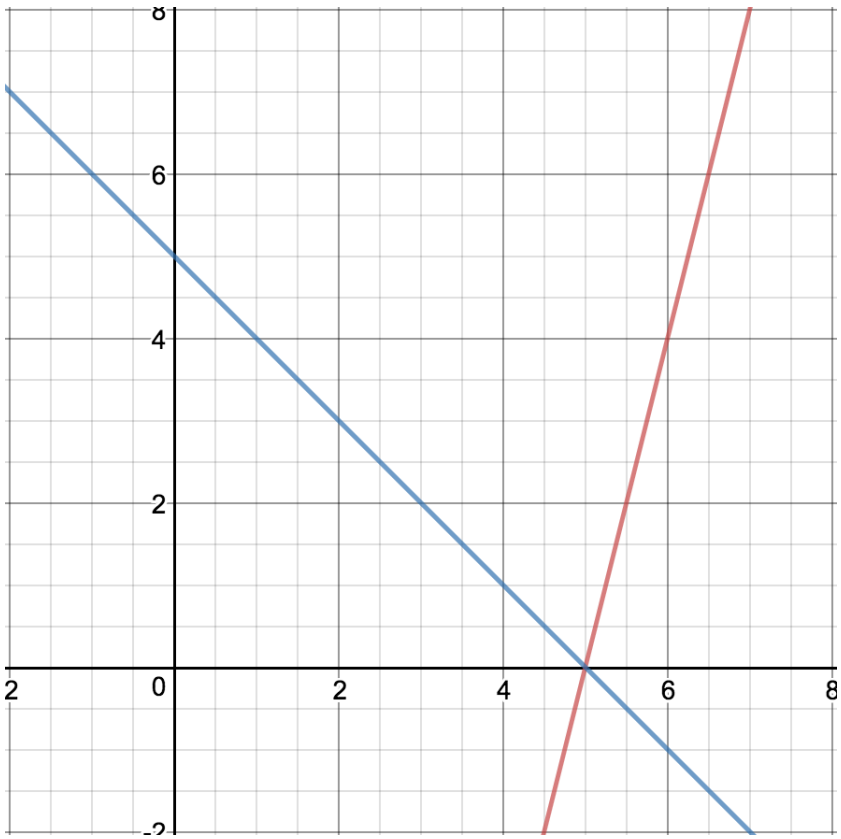
Loss Minimization

Consider the following loss function.

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

2. Draw out this loss function about a point x. Is it convex?

Solution: Plot shown is loss function about x=5. Show convexity by drawing lines. (ignore the lines below 0)



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

Solution:

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

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

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

Gradient Descent

4. Given the following loss function and x, y, θ^t, write out the update equation for θ^{t+1}.

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

Solution: The update equation is given as follows, where α is the step size.

$$\theta^{t+1} \leftarrow \theta^t - \alpha L'(\theta^t)$$
$$L'(\theta) = \sum_{i=1}^{n} 2\theta x_i^2$$

Modeling

5. We wish to model exam grades for DS100 students. We collect various information about student habits, such as how many hours they studied, how many hours they slept before the exam, how many lectures they attended and observe how well they did on the exam. Propose a model to predict exam grades and a loss function to measure the performance of your model.

Solution: Example solution: Let x_1, x_2, x_3 correspond to hours studied, hours slept and number of lectures attended respectively. Let y be their score on the exam.

$$f(x) = \theta_1 * x_1 + \theta_2 * x_2 + \theta_3 * x_3$$
$$L(\theta, x, y) = (f(x) - y)^2$$

6. Suppose we collected even more information about each student, such as their eye color, height, and favorite food. Do you think adding these variables as features would improve our model?

Solution: These features are most likely not going to improve our model. This problem is meant to emphasize overparameterization/overfitting using too many features that do not contribute to the performance of the model. Overfitting, bias variance and feature engineering will be discussed later on in the semester.

Convexity

7. Convexity is a very important and has many useful properties that apply to many relevant areas, including machine learning. It allows optimization problems to be solved more efficiently and for global optimums to be realized. This question will explore the notion of convexity. There are three ways to define convexity.

   a For all $\lambda \in [0, 1], x, y$, we have $f(\lambda x + (1 - \lambda)y \leq \lambda f(x) + (1 - \lambda)f(y)$

   b For all x, y we have $f(y) >= f(x) + f'(x)(y - x)$

   c For all x, we have $f''(x) >= 0$

   Describe the geometric interpretation/meaning of each of the definitions. What assumptions do each of the definitions above make on the function?

   Solution: First definition: Walking in a straight line between points on the function keeps you above the function. The first definition works for any function.
   Second definition: The tangent line at any point lies below the function (globally). The second function requires that the function is differentiable, and the third function requires twice differentiability. Third Definition: The second derivative is non-negative everywhere (aka the function is "concave up" everywhere).

8. 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.

9. Prove that the composition of a convex function f and a convex non decreasing function g is also convex. $h = g(f(x))$
   Hint: Show the following is true

$$(g \circ f)(\lambda x + (1 - \lambda)y) \leq \lambda(g \circ f)(x) + (1 - \lambda)(g \circ f)(y)$$

Solution: We show that the following is true:

$$(g \circ f)(\lambda x + (1-\lambda)y) \leq \lambda(g \circ f)(x) + (1-\lambda)(g \circ f)(y)$$

$$\begin{aligned}
(g \circ f)(\lambda x + (1-\lambda)y) &= g(f(\lambda x + (1-\lambda)y)) \\
&\leq g(\lambda f(x) + (1-\lambda)f(y)) \text{ f convex and g non decreasing} \\
&\leq \lambda g(f(x)) + (1-\lambda)g(f(y)) \text{ g convex} \\
&= \lambda(g \circ f)(x) + (1-\lambda)(g \circ f)(y)
\end{aligned}$$

10. Show that the union of two convex sets is not necessarily convex.

Solution: Draw a venn diagram and show that a line does not lie in the set.