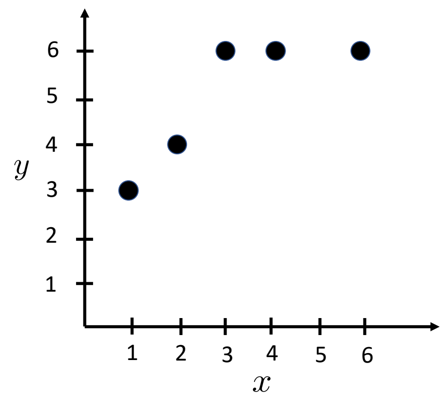
Name: ____________________

Beatriz and mysteries of regression

1. (9 points) Recall that ridge regression is a special case of a general recipe for constructing ML objectives,

$$J(\Theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(h(x^{(i)};\Theta), y^{(i)})\right) + \lambda\mathcal{R}(\Theta),$$

where the hypothesis is $h(x^{(i)};\Theta) = \theta^T x^{(i)} + \theta_0$, the loss is $\mathcal{L}(\hat{y}, y) = (\hat{y} - y)^2$ (where $\hat{y}$ is the prediction, y the observed value), and the regularizer is $\mathcal{R}(\Theta) = ||\theta||^2$ (λ always assumed to be ≥ 0). Consider the following 1-D data set:



(a) What is the mean-squared error (MSE) on this data for the hypothesis $h(x^{(i)}) = 2x^{(i)}$?

Solution: MSE $= \frac{1}{5}(1^2 + 0^2 + 0^2 + 2^2 + 6^2)$ or $\frac{41}{5}$.

(b) Beatriz decides that for her application, small errors in the predicted y-values are irrelevant, and so she designs a new loss function $\mathcal{L}_{tol}(\hat{y}, y)$ which is 0 if $y - 2 \leq \hat{y} \leq y + 2$, and $(|y - \hat{y}| - 2)^2$ otherwise. In words, Loss(guess, actual) is 0 if guess is within 2 units of actual and the difference minus 2, squared, if guess is at least 2 units away from actual. What is the average loss using $\mathcal{L}_{tol}$ on the same data set as the previous question, assuming again the hypothesis $h(x^{(i)}) = 2x^{(i)}$?

Solution: We have

$$\mathcal{L}_{tol}(\hat{y}, y) = \begin{cases} 0, & \text{if } y - 2 \leq \hat{y} \leq y + 2 \\ (|y - y| - 2)^2, & \text{otherwise} \end{cases}$$

Therefore,

$$\frac{1}{5}\sum_{i=1}^{5}\mathcal{L}_{tol}(h(x^{(i)}), y^{(i)}) = \frac{1}{5}(0^2 + 0^2 + 0^2 + 0^2 + 4^2), \tag{1}$$

$$= \frac{16}{5}. \tag{2}$$

(c) In reviewing her 6.036 notes, Beatriz wonders why the regularizer shouldn't instead be $\mathcal{R}(\Theta) = -||\theta||^2$. Explain why this is this a bad idea.

Solution: This is a bad idea because we know that $\lambda \geq 0$ and for an optimization problem where we are looking to minimize the objective, the term $-\lambda||\theta||^2$ can be made to be arbitrarily large and negative (by setting θ to be larger and larger without any constraints).