

6.390 Introduction to Machine Learning
Recitation Week #3
Issued September 19, 2022

1. All Greek to Me! (Based on Fall 2018 Midterm, Question 2)

Consider solving a regularized linear regression problem. For simplicity, we will ignore the offset. Our hypothesis has the form, $h(x; \theta) = \theta^\top x$. Our objective function has the form:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n (h(x^{(i)}; \theta) - y^{(i)})^2 + \lambda R(\theta).$$

(a) Consider these two scenarios:

- $R_1(\theta) = \|\theta - b\|^2$, where λ is very **large**.
- $R_2(\theta) = \|\theta\|^2$, where λ is very **small**.

Here, b is a vector—with compatible dimensions—of all 5's. What benefits and drawbacks may come from these forms of regularization?

(b) We will compute T steps of gradient descent using an update rule of the form,

$$\theta^{t+1} = \theta^t - \eta \nabla_{\theta} J(\theta)|_{\theta=\theta^t},$$

for, $t = 0, \dots, T - 1$ where η is a fixed value throughout execution.

Moving forward, you decide to instead use ridge regression, i.e., $R(\theta) = \|\theta\|^2$. Which parameter(s)/hyperparameters(s) would be included when using the hypothesis to make predictions? Which parameter(s)/hyperparameters(s) are primarily intended to improve generalization? Can T play a similar role to λ ? Can η (for fixed T) play a similar role to λ ?

2. Regression with Standardization (Based on Fall 2019 Midterm, Question 3)

Consider solving a regularized linear regression problem. For simplicity, suppose that your features are one-dimensional. (But, we will consider the offset in this question!) Our hypothesis has the form, $h(x; \theta) = \theta x + \theta_0$.

Suppose that, prior to running your learning algorithm, you decide to standardize your data. With training data, $D_n = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$, and, $x^{(i)}, y^{(i)} \in \mathbb{R}$, you transform the data as:

$$x_r^{(i)} = \frac{x^{(i)} - \mu(X)}{\text{SD}(X)}, \quad y_r^{(i)} = \frac{y^{(i)} - \mu(Y)}{\text{SD}(Y)},$$

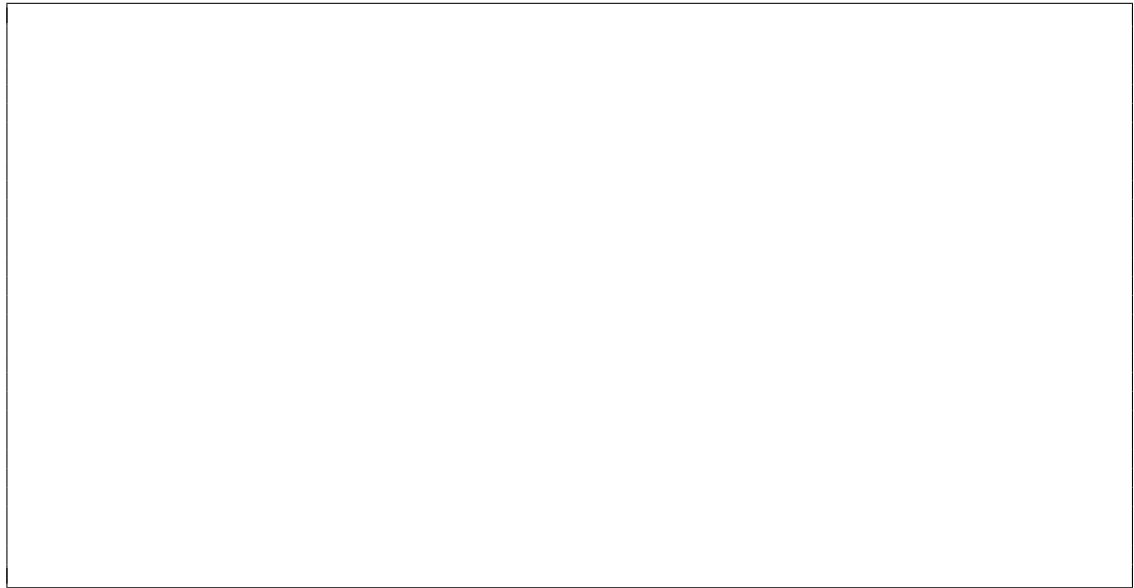
where $X \in \mathbb{R}^n$ is a vector containing all n features, $\mu(X)$ is the mean of X and $\text{SD}(X)$ is the standard deviation; let the same apply for Y , respectively. You then perform ordinary least squares regression using the $(x_r^{(i)}, y_r^{(i)})$ data points, and get the parameter θ . Now, find the relationship between θ^* and θ , where θ^* is the parameter of the optimal fit hyperplane on the non-standardized dataset, and θ is the optimal parameter using the standardized data. Write the expression in terms of: $\theta, x^{(i)}, y^{(i)}, \mu(X), \mu(Y), \text{SD}(X), \text{SD}(Y)$.

3. New Hypothesis, Same Gradient Descent? (Based on Spring 2019 Midterm, Question 5)
Ben develops a new hypothesis class:

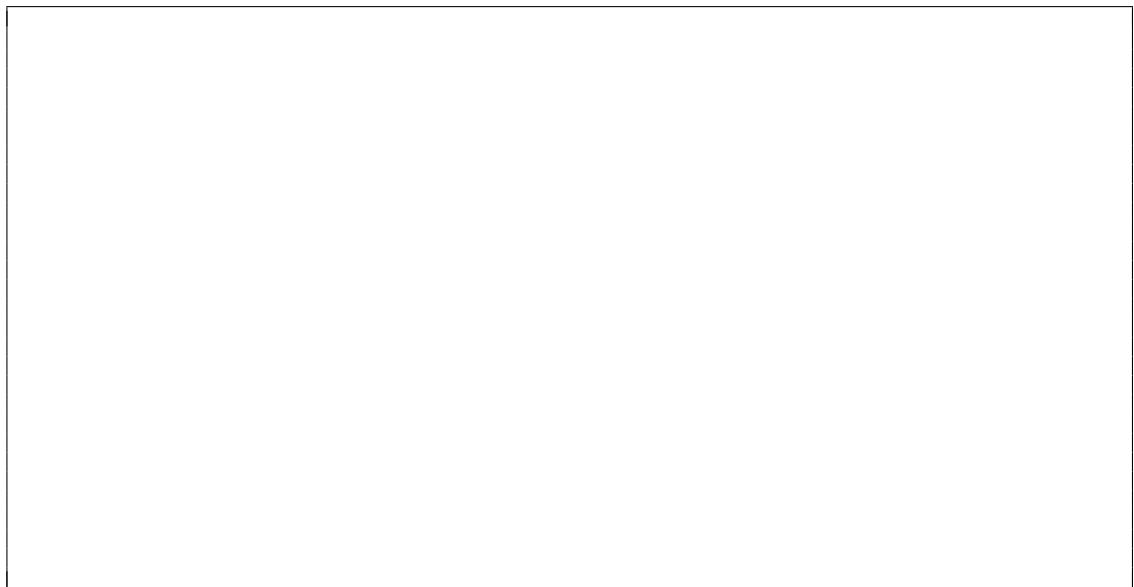
$$h(x; \theta) = \theta_1 x_1 + \theta_1 x_1^2 + \theta_2 x_2 + \theta_2 x_2^2,$$

where, $x = (x_1, x_2)$. He plans to use it for a regression problem on the data set, $D_n = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$.

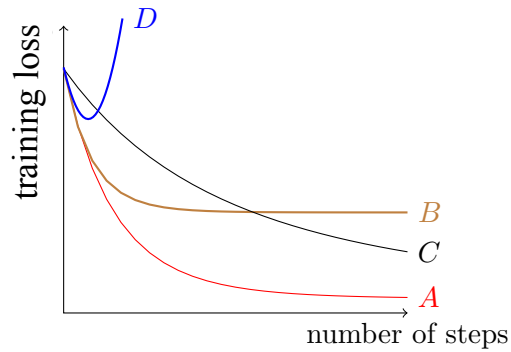
- (a) Ben will use gradient descent to compute model parameters θ_1, θ_2 . His loss function is squared error. Derive an update rule for θ_1 given the learning rate η .



- (b) Describe the shape of the objective function, $J(\theta_1, \theta_2)$ with $\lambda = 0$. How many minima will it have? Assume that the data set D_n is fixed.



- (c) Ben tries different settings of the learning rate η , during **training**. Depending on the setting he obtains different behavior of the gradient descent algorithm. Match each plot (A,B,C,D) to the best fitting description (assume MSE loss): (i) Learning rate too low, (ii) learning rate about right, (iii) learning rate too high, (iv) learning rate much too high.



- (d) Alyssa suggests using absolute error, instead, defined by:

$$L_{\text{AE}} = |y^{(i)} - h(x^{(i)}; \theta)|.$$

What advantages (or disadvantages) would this loss function have over squared error?