6.390 Introduction to Machine Learning

Recitation Week #2

Issued February 13, 2023

Suppose that you are given the task to train a model to predict the pollution level in different cities given data points from satellite readings. You are given a data set $D_n = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{n}$ comprised of pairs of satellite readings and the pollution level recorded. The satellite readings include many features, e.g., temperature readings, building density, local population, etc. The feature vectors and labels are both assumed to be real-valued.

You hypothesize that the pollution level will be equal to a linear combination of the different features measured by the satellites, i.e., your hypotheses are linear regressors with parameters $\Theta = (\theta, \theta_0)$ of the form $h(x; \Theta) = \theta^\top x + \theta_0$. The goal is to apply your models to locations where you have the satellite readings, but you do not have the pollution level readings.

(a) Consider the ordinary least squares objective function

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

Suppose for now you just hope to find a solution to θ and ignore the offset θ_0. You construct matrices $\tilde{X}, \tilde{Y}$ from the training data set as follows:

$$\tilde{X} = \begin{bmatrix} x_1^{(1)} & \dots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \dots & x_d^{(n)} \end{bmatrix}, \quad \tilde{Y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}.$$

and attempt compute the analytical solution to θ as,

$$\theta = \left(\tilde{X}^\top \tilde{X}\right)^{-1} \tilde{X}^\top \tilde{Y}.$$

However, you are not able to invert the matrix $\tilde{X}^\top \tilde{X}$.

i. Brainstorm some potential causes for this predicament. Is there an issue with the satellite readings, the pollution level readings, both, or neither? Is it at all possible to obtain a solution to minimize the ordinary least squares objective function in this regime?

ii. What might happen if you employ the random regression algorithm (from Lab 1 and Homework 1) on this dataset, instead of computing the analytical solution?

iii. If you indeed are interested in finding a solution to θ_0 as well, how should you construct your data matrices?

(b) You shift gears to a different objective function known as ridge regression,

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} (h(x^{(i)}; \theta, \theta_0) - y^{(i)})^2 + \lambda ||\theta||^2.$$

You construct matrices $\tilde{X}, \tilde{Y}$ from the training data set as follows:

$$\tilde{X} = \begin{bmatrix} x_1^{(1)} & \dots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \dots & x_d^{(n)} \end{bmatrix}, \quad \tilde{Y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}.$$

You pick a positive λ and find the analytical solution to the ridge regression objective function, ignoring the offset θ_0 for now, as follows:

$$\theta = \left(\tilde{X}^\top \tilde{X} + n\lambda I \right)^{-1} \tilde{X}^\top \tilde{Y}.$$

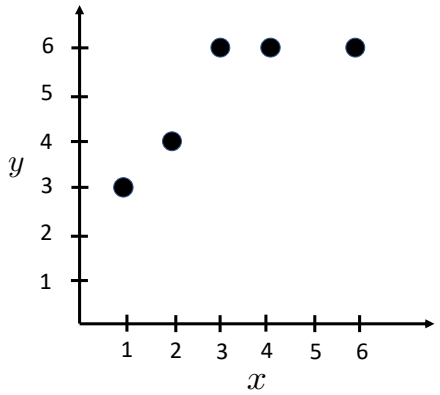
i. What is the role of the hyperparameter, λ, in this objective function? What if different features are at different scales (e.g., parts per million of pollution vs. population density vs. distance of sea level in miles)? How does ridge regression deal with the case where one feature is a linear function of another? How could you interpret the impact of different features from the values of θ?

ii. What would happen if the offset parameter θ_0 were to be included in the regularization term? Why would we want the parameters θ to have a small magnitude, but not θ_0?

(c) Recall that least-squares regression is a special case of a general recipe for constructing ML objectives,

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

the loss is $\mathcal{L}(g, y) = (g - y)^2$ (where $g = h(x;\Theta)$ is the prediction, y the observed value). Consider the following 1-D data set:



i. What is the mean-squared error (MSE) on this data for the hypothesis: $h(x) = 2x$?

ii. Suppose that, for this application, small errors in the predicted y-values are irrelevant, and so you design a new loss function $\mathcal{L}_{tol}(g, y)$ as,

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

What is the average loss using $\mathcal{L}_{tol}$ on the same data set as the previous question, assuming again the hypothesis $h(x) = 2x$?