Discussion 8
Machine Learning, Spring 2019

1 Linear regression

1.1

Consider a dataset $\{(x_i, y_i)_{i=1}^n\}$, where $x, y \in \mathbf{R}$. We use a linear regression method to model this data. To test the linear regressor, we choose at random some data points to be a training set, and the remaining data points to be a test set. Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors?

Solution.

The training error tends to increase. As more examples have to be fitted, it becomes harder to 'hit', or even come close, to all of them.

The test error tends to decrease. As we take into account more examples when training, we have more information, and can come up with a model that better resembles the true behavior. More training examples lead to better generalization.

1.2

Fit the linear regression model to the following dataset:

x	-1	0	2
y	1	-1	1

1. Fit $Y = \beta_0$, find β_0
2. Fit $Y = \beta_1 x$, find β_1
3. Fit $Y = \beta_1 x + \beta_0$, find β_0, β_1

Provide the minimum of a loss function $J(\beta) = \sum_{i=1}^n (y_i - Y_i(\beta))^2$ for each of the models you computed and state which model achieves better fitting.

Solution.

1. $Y = \beta_0$
   $J(\beta) = \sum_{i=1}^n (y_i - \beta_0)^2 = 2(1-\beta_0)^2 + (1+\beta_0)^2 = 3(\beta_0 - \frac{1}{3})^2 + 2\frac{2}{3}$
   Therefore $\beta_0 = \frac{1}{3}$, $J(\beta_0) = 2\frac{2}{3}$
2. $Y = \beta_1 x$
   $J(\beta) = \sum_{i=1}^n (y_i - \beta_1 x_1)^2 = (1+\beta_1)^2 + 1 + (1-2\beta_1)^2 = 5(\beta_1 - \frac{1}{5})^2 + 2\frac{4}{5}$
   Therefore $\beta_1 = \frac{1}{5}$, $J(\beta_1) = 2\frac{4}{5}$
3. $Y = \beta_1 x + \beta_0$
   $\beta = (X^T X)^{-1} X^T y$, where

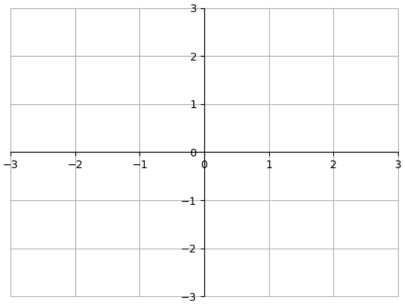

Figure 1: Problem 1.2

$$X = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 2 \end{pmatrix}; \quad y = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

Then

$$\beta = \frac{1}{7}\begin{pmatrix} 2 \\ 1 \end{pmatrix}; \quad J(\beta) = 2\frac{4}{7}$$

Model 3 produces the lowest least square loss, and model 2 produces the highest loss.

1.3

Consider following dataset of 3 points:

x	0	1	2
y	1	1	2

Figure 2 shows linear regression results with different regularization penalties. Linear regression problem solves following minimization problem:

$$\arg\min_{\theta_0,\theta_1} \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 + R(\theta_0, \theta_1)$$

where R represents a regularization penalty which could be L-1 or L-2.

However, instead of computing the derivatives to get a minimum value, we could adopt a geometric method. In this way, rather than letting the square error term and the regularization penalty term vary simultaneously as a function of θ_0 and θ_1, we can fix one and only let the other vary at a time. Having a upper-bound, r, on the penalty, we can replace $R(\theta_0, \theta_1)$ by r, and solve a minimization problem on the square error term for any non-negative value of r. Finally, we get the minimum value by enumerating over all possible value of r. That is,

$$\min_{\theta_0,\theta_1} \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 + R(\theta_0, \theta_1) = \min_{r \geq 0} \left(\min_{\theta_0,\theta_1} \{ \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 | R(\theta_0, \theta_1) \leq r \} + r \right)$$

The value of (θ_0, θ_1) corresponding to the minimum value of the object function can be got at the same time. In Figure 3, we plot the square error term, $\sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2$, by ellipse contours. The circle contours in Fig 3(a) plots a L-2 penalty with $\lambda = 5$, whereas the square contours in Fig 3(b) plots a L-1 penalty with $\lambda = 5$. To further explain how it works, the solution to

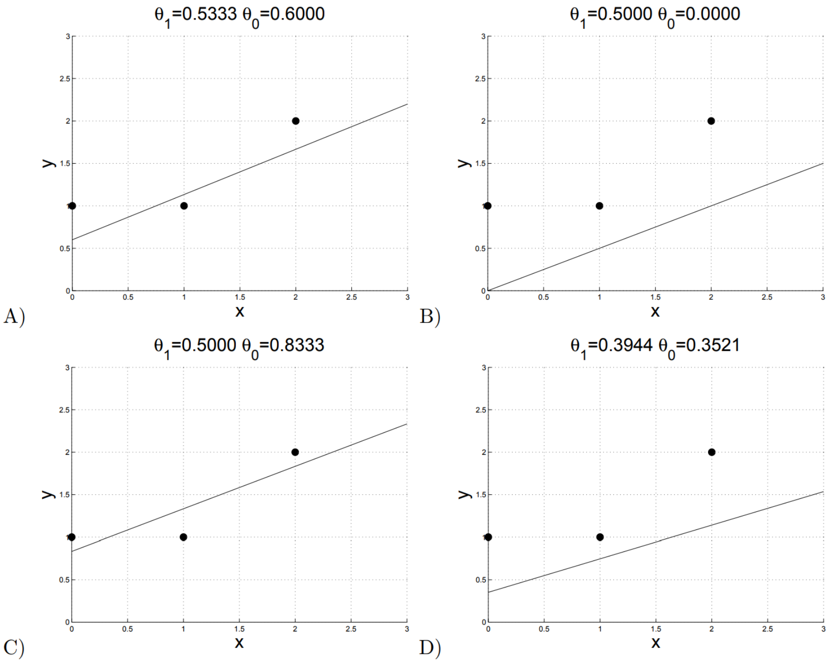

Figure 2: Problem 1.3. Plots of linear regression results with various regularization

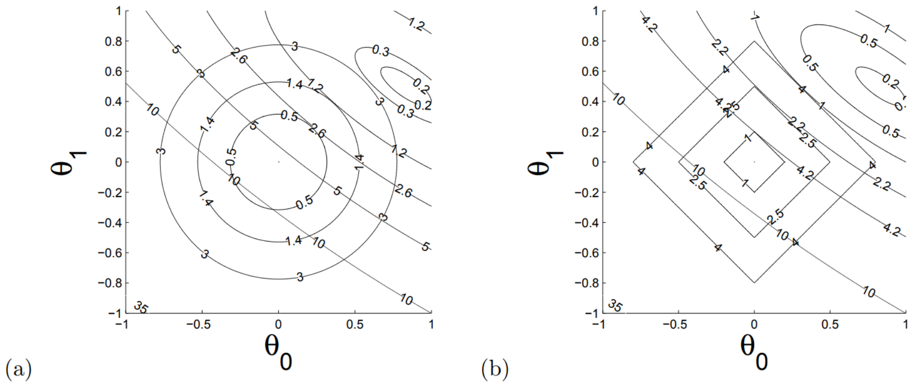

Figure 3: Problem 1.3. Contour plots of the decomposition for the linear regression problem with (a) L-2 regularization (b) L-1 regularization where the ellipsis correspond to the square error term, and circles/squares correspond to the regularization penalty term.

$$\min_{\theta_0,\theta_1}\{\sum_{i=1}^{n}(y_i-\theta_1 x_i-\theta_0)^2|R(\theta_0,\theta_1)\leq r\}$$

is the height of the smallest ellipse contour that is tangent with (or contained in) the contour that depict $R(\theta_0,\theta_1)=r$. The desired (θ_0,θ_1) are the coordinates of the tangent point.

Assign each plot in Figure 2 to one of the following regularization methods

1. No regularization
2. L-2 regularization with $\lambda=5$

3. L-1 regularization with $\lambda = 5$

4. L-2 regularization with $\lambda = 1$

Now, If we had more features and we want to perform feature selection while solving the linear regression problem, what regularization method is better to use?

Solution.

1. C

2. D

3. B

4. A

We will choose L-1, and we will use bigger λ when we want fewer effective features.

2 K-means and Hierarchical Clustering

Recall that in k-means clustering we attempt to find k cluster centers $c_j \in \mathbf{R}^d$, $j \in \{1, ..., k\}$ such that the total distance between each datapoint and the nearest cluster center is minimized. In other words, we attempt to find $c_1, ..., c_k$ that minimizes

$$\sum_{i=1}^{n} \min_{j \in \{1,...,k\}} ||x_i - c_j||^2$$

where n is the number of data points. To do so, we iterate between assigning xi to the nearest cluster center and updating each cluster center c_j to the average of all points assigned to the j-th cluster.

2.1

Instead of holding the number of clusters k fixed, one can think of minimizing over both k and c. Show that this is a bad idea. Specifically, what is the minimum possible value of $\sum_{i=1}^{n} \min_{j \in \{1,...,k\}} ||x_i - c_j||^2$? What values of k and c result in this value?

Solution.

The minimum objective value is 0. It is achieved when we have n clusters such that $c_i = x_i$.

2.2

Consider the one-dimensional ($d = 1$) case where $k = 3$ and we have 4 data points $x_1 = 1, x_2 = 2, x_3 = 5, x_4 = 7$. What is the optimal clustering for this data? What is the corresponding value of the objective?

Solution.

$c_1 = 1.5, c_2 = 5, c_3 = 7$, objective = 0.5.

2.3

Give advantages of hierarchical clustering over K-means clustering, and advantages of K-means clustering over hierarchical clustering.

Solution.

Advantages of hierarchical clustering:

1. Don't need to know how many clusters you're after

2. Can cut hierarchy at any level to get any number of clusters

3. Easy to interpret hierarchy for particular applications

Advantages of K-means clustering:

1. Can be much faster than hierarchical clustering, depending on data
2. Can incorporate new data and reform clusters easily