CS2109s Tutorial 6

by Lee Zong Xun

Recap

Support Vector Machines 💀

Question 1

L1 Regularization, also called a **lasso regression**, adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function.

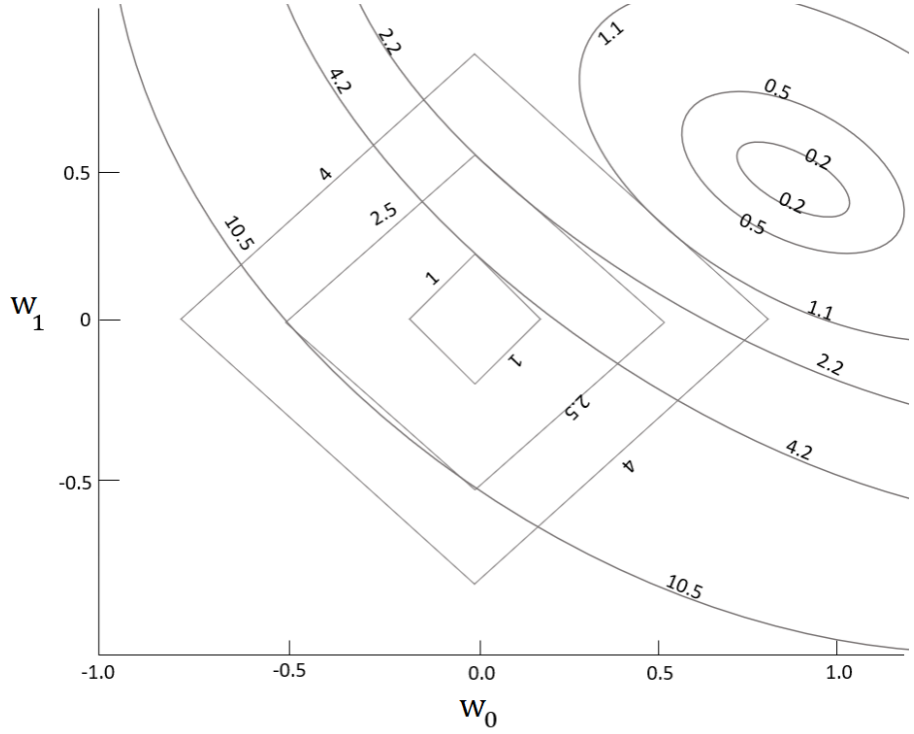
$$J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{n} |w_i|\right]$$

L2 Regularization, also called a **ridge regression**, adds the “squared magnitude” of the coefficient as the penalty term to the loss function.

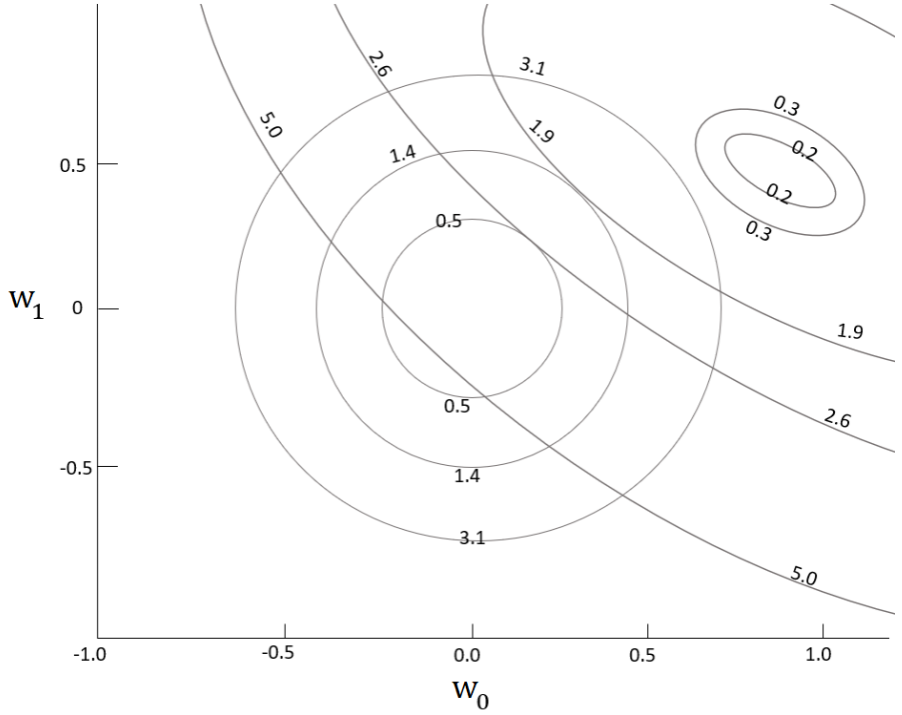
$$J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{n} {w_i}^2\right]$$

Question 1

The figures below shows contour plots for a linear regression problem with the 2 different regularizers. The elliptic contours represent the squared error term. The diamond (Figure 1) and circle (Figure 2) contours represent the regularisation penalty term when $\lambda = 5$.



L1



L2

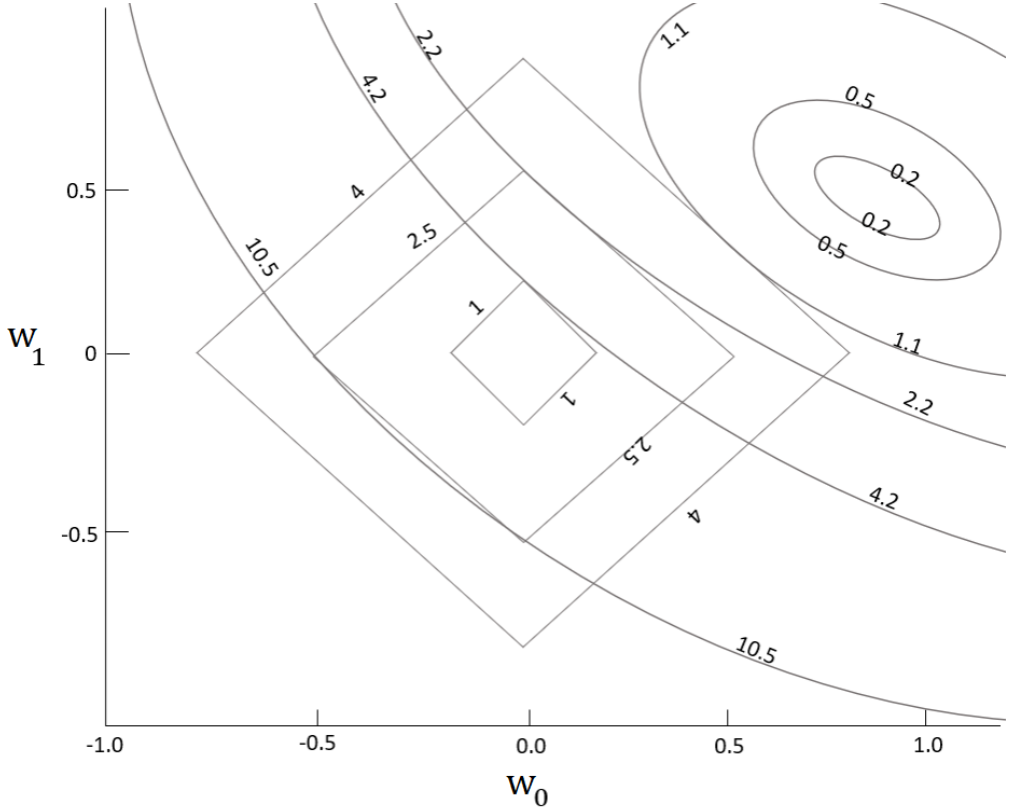
Along a contour, the corresponding loss remains the same as w_0 and w_1 vary. Intersections between the 2 different contours represent points with possible values for w_0 and w_1. The total value of the loss function at such a point is the sum of the two contours values. For example, for the point $(w_0 = 0.0,\ w_1 = -0.5)$ in Figure 1, the total loss (regularization and error loss) is $2.5 + 10.5 = 13$.

For each of the following cases, provide an estimate of the optimal values of w_0 and w_1 using the figures as reference.

- No regularisation.
- L1 regularisation with $\lambda = 5$.
- L2 regularisation with $\lambda = 5$.

Question 1

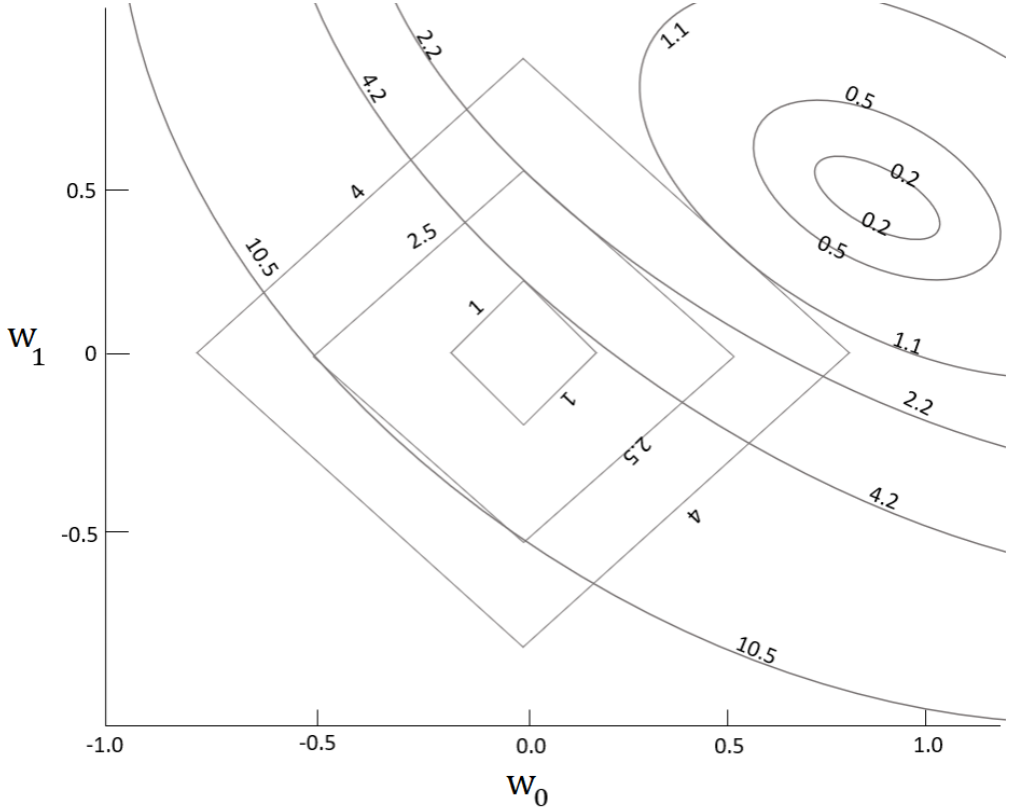
No regularization



$w_0 = 0.9, w_1 = 0.5$. Cost: approx 0 (no MSE and no regularization penalty).

Question 1

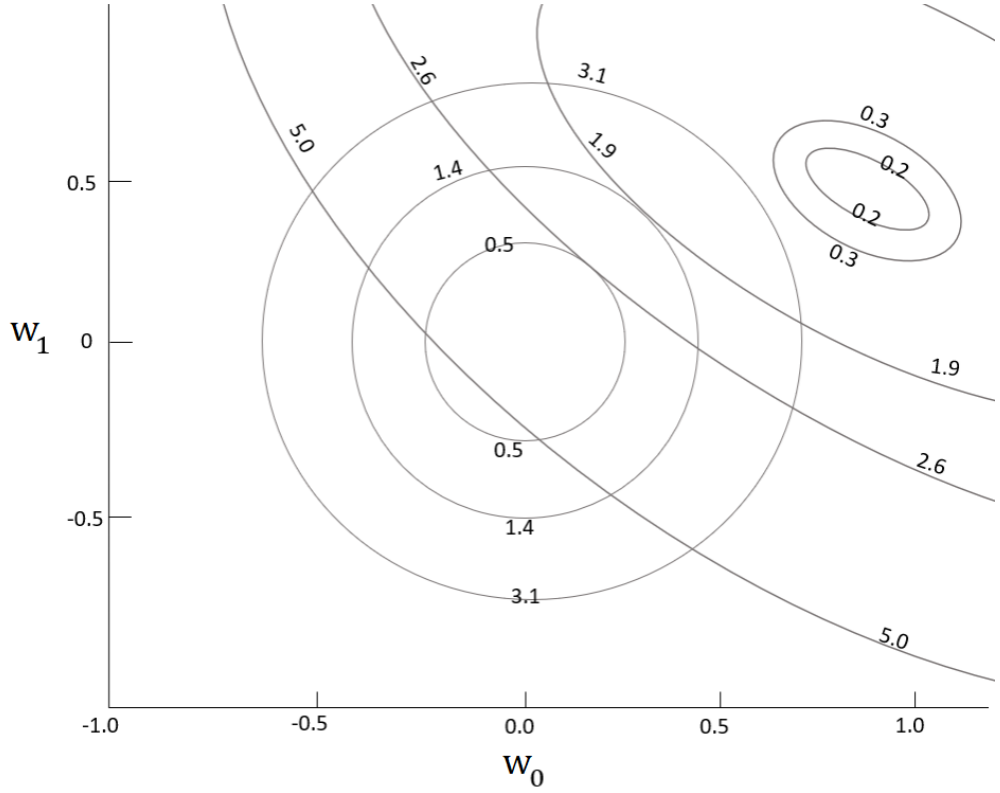
L1 regularization



$w_0 = 0.0, w_1 = 0.55$. Cost: 4.7 = 2.2 (MSE) + 2.5 (L1 penalty).

Question 1

L2 regularization



$w_0 = 0.2, w_1 = 0.25$. Cost: 3.1 = 2.6 (MSE) + 0.5 (L2 penalty).

Question 1

What makes L2 Regularisation different from L1 Regularisation in terms of what they do to the parameters?

- L2 norm regularization heavily penalizes larger parameters. L2 thus attempts to pull **all** parameters towards small values.
- L1 regularisation may set values of certain parameters, which are multiplied with less important features, to zero, while others are set to some non-zero value (that are possibly large).
    - By setting some parameter values to zero, L1 regularization implicitly selects features to be 'excluded'.

Question 2

In class, we formulated SVMs as the following minimization problem:

$$\min_w C \left[\sum_{i=1}^{m} y^{(i)} \, Cost_1(w^T x^{(i)}) + (1 - y^{(i)}) \, Cost_0(w^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} w_i^2$$

The summand term,

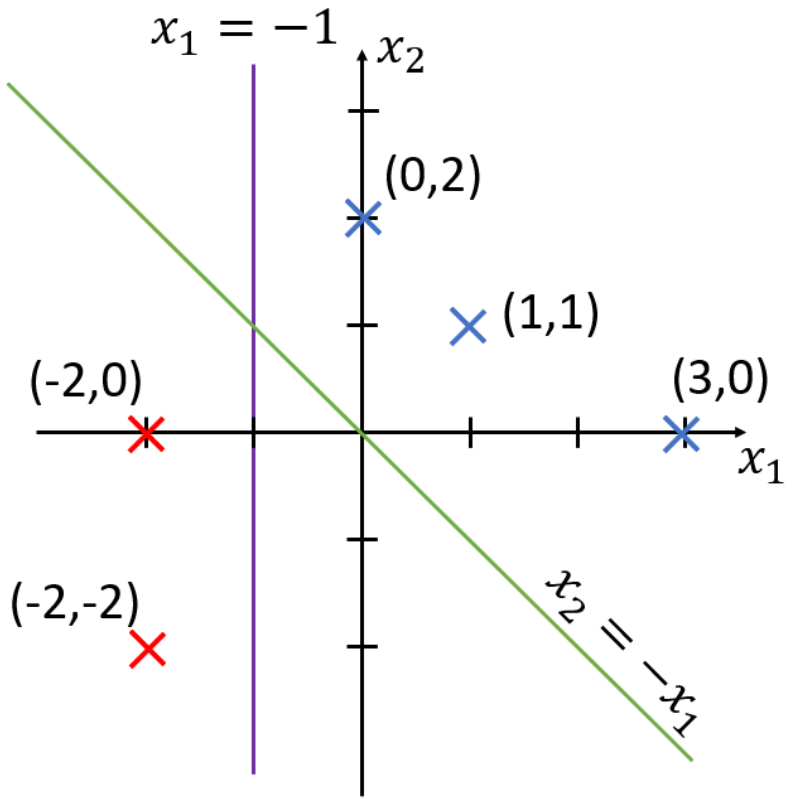
$$y \cdot Cost_1(w^T x) + (1 - y) \, Cost_0(w^T x)$$

is often referred to as the *hinge loss* and expressed as follows:

$$y \max(0, 1 - w^T x) + (1 - y) \max(0, 1 + w^T x)$$

Question 2

x_1	x_2	y
-2	-2	0
-2	0	0
0	2	1
1	1	1
3	0	1



The red and blue points correspond to points with $y = 0$ and $y = 1$ respectively. The decision boundary for a linear model on this data would be some function like

$$h(x_1, x_2) = w_0 + w_1 x_1 + w_2 x_2$$

Question 2

The two lines (green and purple) represent decision boundaries of 2 different linear models. How can we parametrize the lines, i.e. what are the values for w_0, w_1, w_2 for the 2 lines?

Note that the choice of w_0, w_1, w_2 is not unique.

- Green is $x_2 = -x_1$ which can be written as $h(x_1, x_2) = 0 + x_1 + x_2$ or more generally as $h(x_1, x_2) = k \cdot x_1 + k \cdot x_2$. So can be parametrized as any $w = (0, k, k)$ for some real number k.
- Purple is $x_1 = -1$ which can be written as $h(x_1, x_2) = 1 + x_1 + 0 \cdot x_2$ or more generally as $h(x_1, x_2) = k + k \cdot x_1$. So can be parametrized as any $w = (k, k, 0)$ for some real number k.

Example

Lets now compute the total losses associated with the green and purple lines, with $C = 1$. For the purple line, the support vectors are the points $(-2, -2)$, $(-2, 0)$ and $(0, 2)$. Assuming $\boldsymbol{w} = (1, 1, 0)$, the hinge loss is:

$$\begin{aligned}\max(0, 1 + (k - 2k + 0)) + \max(0, 1 + (k - 2k + 0)) + \max(0, 1 - (k + 0 + 0)) \\ + \max(0, 1 - (k + k + 0)) + \max(0, 1 - (k + 3k + 0)) \\ = 3 \cdot \max(0, 1 - k) + \max(0, 1 - 2k) + \max(0, 1 - 4k)\end{aligned}$$

Hence, the total loss is $3 \cdot \max(0, 1 - k) + \max(0, 1 - 2k) + \max(0, 1 - 4k) + \frac{k^2 + 0^2}{2}$. This is minimized when $k = 1$, which results in a total loss of 0.5.

Question 2

Calculate the total loss for the green line in in a similar manner. Also find the parameter(s), that result in the least loss.

Support vectors are $(-2, 0)$, $(0, 2)$ and $(1, 1)$.

Hinge loss is:

$$\begin{aligned}\max(0, 1-4k) + \max(0, 1-2k) + \max(0, 1-2k) + \max(0, 1-2k) + \max(0, 1-3k) \\ = 3 \cdot \max(0, 1-2k) + \max(0, 1-3k) + \max(0, 1-4k)\end{aligned}$$

and total loss is $3 \cdot \max(0, 1-2k) + \max(0, 1-3k) + \max(0, 1-4k) + k^2$.
Minimized at 0.25 when $k = 0.5$.

Notice how, while minimizing, the loss can be reduced to only the loss on the support vectors; the other points don't contribute to the loss. That's why the model is called Support Vector Machine!

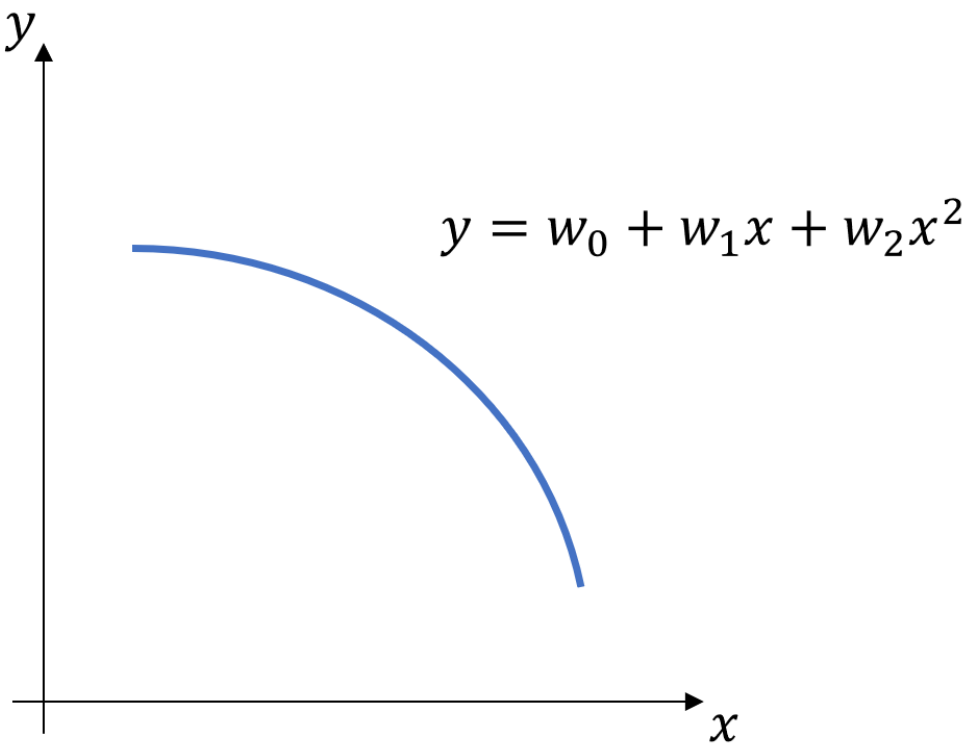
Question 2

Which line is a better solution to the SVM?

- The green line is a better solution than the purple line as it has a lower associated loss as computed in part (b)
- **An intuitive way**: Both lines completely separate the sets without mislabels, then we can choose the line that has the maximum margin.

Question 3

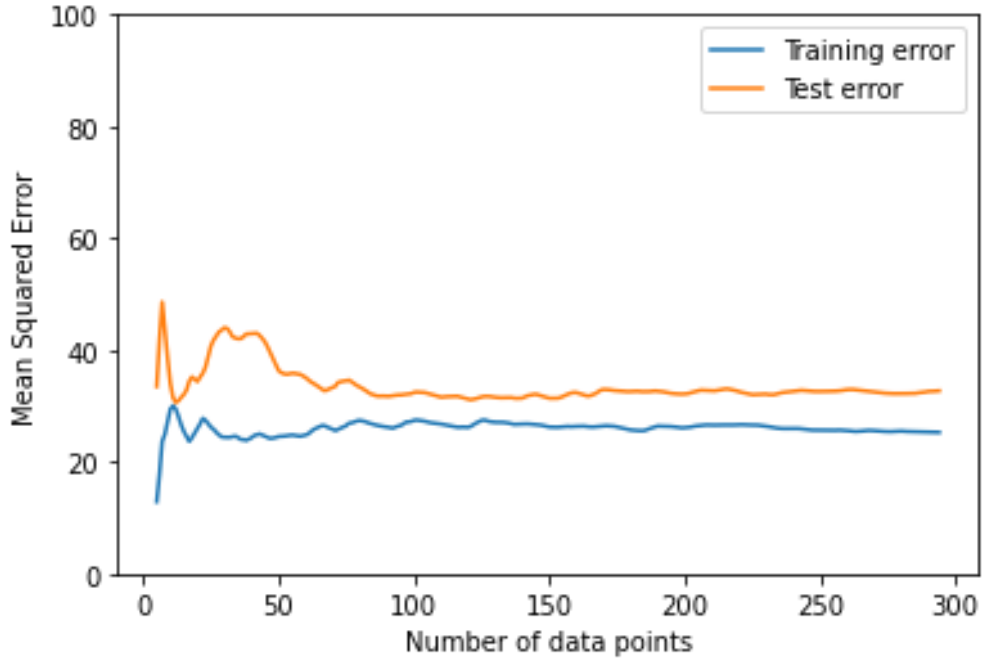
In this problem, we will investigate how loss varies with the number of training samples under different conditions. Consider a dataset like the following set of points, where we know that the ``correct'' hypothesis is a 2nd-degree polynomial.



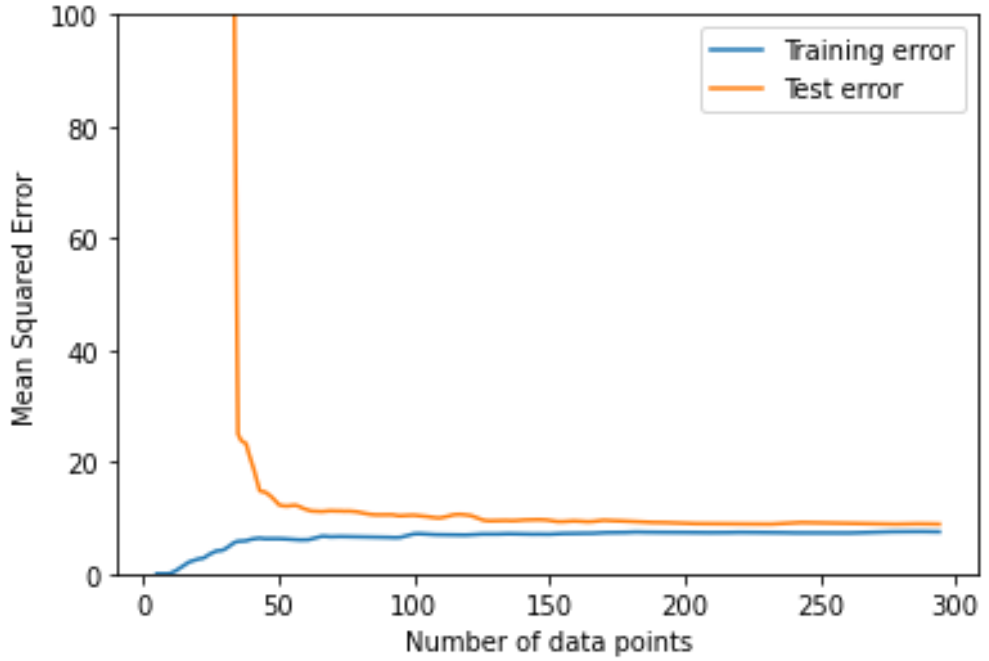
Two different models were trained on the dataset many times. In each iteration, the training error and test error were recorded. The model hypotheses are as below:

- $H_w(x) = w_0 + w_1 x$
- $H_w(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_{10} x^{10}$

The training and test errors obtained were plotted for each model, and the resulting graphs are as below:



Model X



Model Y

Question 3

Between the two graphs, which one indicates a model with a higher bias? How does bias seem to vary with the number of samples?

Model X. Relatively higher error, even as samples increase, indicates inability to capture the true relationship sufficiently, hinting high bias. Bias does not (generally) improve with increase in number of samples.

Question 3

Which graph indicates a model with a higher variance? How does variance seem to vary with the number of samples?

Model Y. Lower error, but initially higher difference between the 2 error indicates high variance. Getting more data points is likely to help variance.

Question 3

Which model do you think each graph belongs to. Explain your reasoning.

Model X (high bias) is the linear model, because: linear model can't capture quadratic relationship, has high bias. Model Y (high var) is the high degree polynomial, because: overfits the points so initially high difference in errors. As number of samples increases, the "degree of overfitting" reduces approaching a roughly quadratic curve.

Note: The models above are un-regularized. How might regularization affect the graphs for each of them?

Optional

In past lectures and Problem Set 3, you were introduced to he concept of Feature Scaling: the process of taking a dataset with features having different scales, and scaling them so that they all have roughly the same magnitudes.

SVMs, on the related hand, are *very* sensitive to feature scaling. Is this harmful to SVM performance? Explain your choice on why or why not.

Feature scaling involves taking large or small values and bringing them into the range of $[0, 1]$. This means dilating and transforming the points in a dataset, bringing them to similar ranges.

- SVMs use the features' dimensions to find the best hyper-plane for the dataset. It does this by maximising the margins between points of opposing classes along **all** dimensions.
- If the scale of the data is haphazard and all over the place, finding the hyper-plane will be very difficult because certain wide-range dimensions/features will have very wide margins, thereby dominating the distance measurement. On the other hand, other narrow-range dimensions/features will have tighter margins.

By scaling them all to similar ranges, all dimensions have equal influence on the distance metric used.