National University of Singapore
School of Computing
CS2109S: Introduction to AI and Machine Learning
Semester II, 2022/2023

Tutorial 6
SVMs and Regularisation

These questions will be discussed during the tutorial session. Please be prepared to answer them.

Summary of Key Concepts

In this tutorial, we will discuss and explore the following learning points from Lecture:

1. Visualising Regularisation
2. Hinge Loss in Support Vector Machines (SVM)
3. Bias and Variance

Problems

L1-norm vs. L2-norm.

In lecture, we discussed using regularization on linear regression using the L2-norm. This is also called Ridge Regression, and the cost function is:

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

It is also possible to do regularization using the L1-norm. This is called Lasso Regression, and the resulting cost function is:

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

Now, in this problem, we will investigate how regularization will work with these 2 norms in a 2D space. 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 $\lambda = 5$.

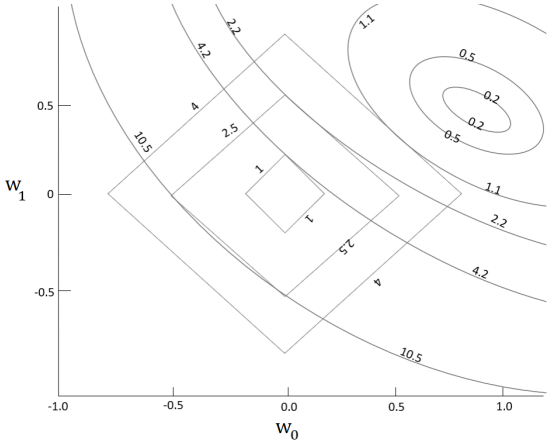

Figure 1: Contour plots for the linear regression problem with L1 regularisation

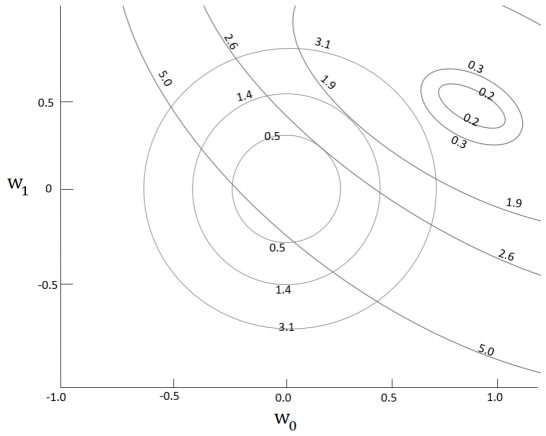

Figure 2: Contour plots for the linear regression problem with L2 regularisation

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 $2.5 + 10.5 = 13$.

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

1. No regularisation.
2. L1 regularisation with $\lambda = 5$.
3. L2 regularisation with $\lambda = 5$.

Additionally, what do you think makes L2 Regularisation different from L1 Regularisation in terms of what they do to the parameters?

Understanding SVM and Hinge Loss.

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

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} \bar{y}^{(i)} \bar{y}^{(j)} K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$$

subject to $\forall i\ \alpha^{(i)} \geq 0$ and $\sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} = 0$, where $K(\mathbf{x}, \mathbf{l}) = \mathbf{x} \cdot \mathbf{l}$ for a linear SVM classifier.

2. Consider the data points given in tabular and graphical form (Figure 3) below:

i	$x_1^{(i)}$	$x_2^{(i)}$	$\bar{y}^{(i)}$
1	-2	-2	-1
2	-2	0	-1
3	0	2	1
4	1	1	1
5	3	0	1

The red and blue points correspond to points with $\hat{y} = -1$ and $\hat{y} = 1$ respectively. The decision boundary for a linear model on this data would be the function $h(x_1, x_2) = \sum_{i=1}^{5} \alpha^{(i)} \bar{y}^{(i)} \left(x_1 x_1^{(i)} + x_2 x_2^{(i)} \right) + b$.

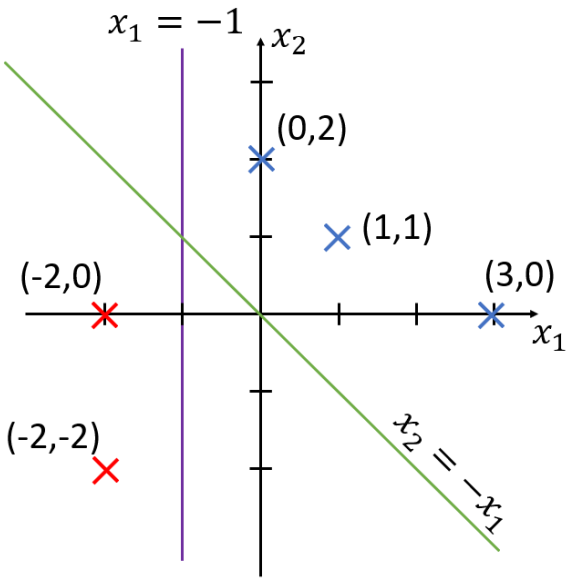

Figure 3: SVM example.

(a) 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 $\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \alpha^{(4)}, \alpha^{(5)}, b$ for the 2 lines? *Hint: To simplify the system of equations, α for non-support vectors is 0.*

Recall that when we were deriving the SVM optimization problem we came across the loss function

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{1} \max\{0, 1 - \bar{y}^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)\}$$

The term $\max\{0, 1 - \bar{y}^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)\}$ is called the hinge loss.

Lets now compute the total losses associated with purple line using the above loss function with $C = 1$. Assuming $\boldsymbol{w} = (k, 0)$ and $b = k$, the hinge loss is:

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

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.

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

(c) Which line is a better solution to the SVM?

(d) (Optional) Solve $\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} \bar{y}^{(i)} \bar{y}^{(j)} \left(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \right)$ using the possible values of $\alpha^{(i)}$ found in part **(a)** for the green line. Using the equation $\mathbf{w} =$

$\sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)}$, what can you conclude about the value of w found in part **(b)** and the one calculated here?

Understanding Bias & Variance.

In lecture, we discussed how the loss varies with the degree of a polynomial (that is used as the hypothesis), and with λ, when there is regularization. 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.

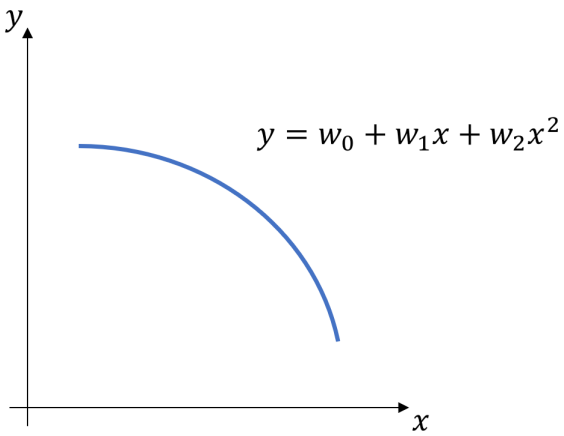

Figure 4: Sample data points.

3. Two different models were trained on the dataset many times, gradually increasing the number of samples used in training. In each iteration, the training error and test error were recorded. The model hypotheses are as below:

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

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

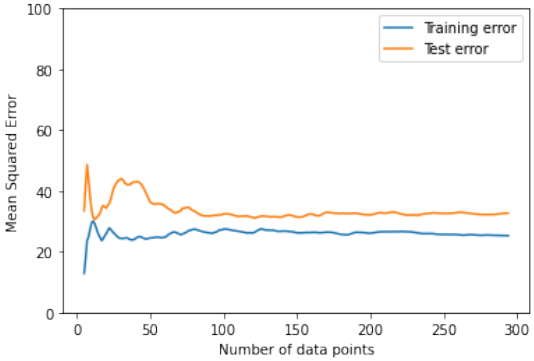

Figure 5: Learning curves for Model X.

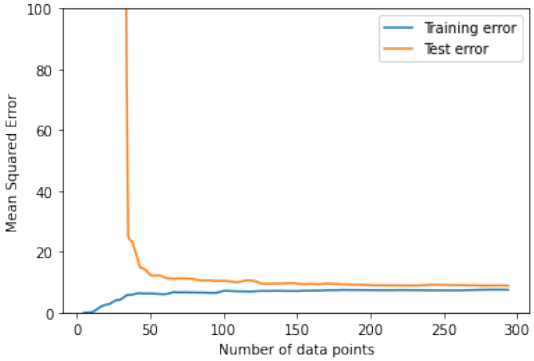

Figure 6: Learning curves for Model Y.

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

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

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

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