National University of Singapore
School of Computing
CS2109S: Introduction to AI and Machine Learning
Semester 2, 2024/2025

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 key learning points from Lecture:

1. Visualising Regularisation
2. SVM
3. Bias & Variance

A Visualising Regularisation

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

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

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

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

Now, in this problem, we will investigate how regularisation 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 regularisers. The diamond contours represent the absolute error term while the elliptic contours represent the squared error term. The diamond (Figure 1) and circle (Figure 2) contours represent the regularisation penalty term when $\lambda = 5$. The total area of the regularisation contours represent the set of all feasible solutions for $\mathbf{w}$

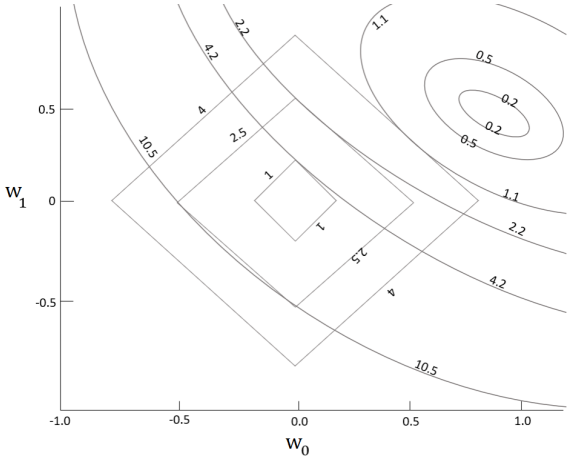

Figure 1: Contour plots for the linear regression problem with L1 regularisation

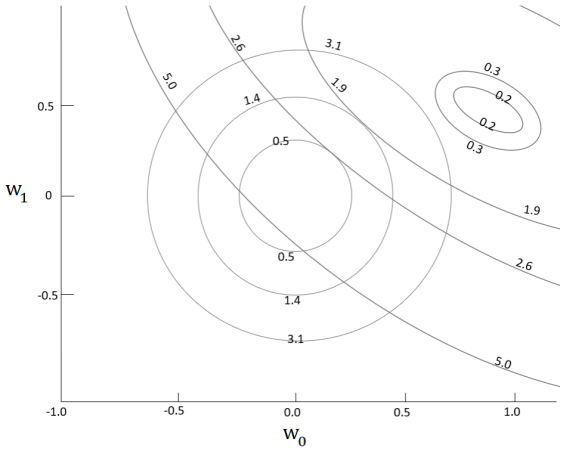

Figure 2: Contour plots for the linear regression problem with L2 regularisation

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 (regularisation 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.
   (a) No regularisation.
   (b) L1 regularisation with $\lambda = 5$.
   (c) L2 regularisation with $\lambda = 5$.
2. How does L2 Regularisation differ from L1 Regularisation in terms of what they do to the parameters?

B SVM

In class, we learned an SVM classification model that constructs a separating hyperplane between data points of two classes in an optimal way. The SVM classifier takes the following form

$$\text{SVM}_\alpha(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha^{(i)} k(x^{(i)}, x^{(j)})\right), \tag{1}$$

where the kernel function $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ for this tutorial is simply the dot product, $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}$ for a linear SVM classifier. In addition, the numbers $\alpha^{(1)}, \alpha^{(2)}, ..., \alpha^{(N)}$ are the parameters of the model. This formulation is called the ***dual*** formulation. In this tutorial, we investigate basic concepts for the SVM and the ***primal*** formulation of the SVM.

Note: In this tutorial, we mathematically only discuss zero-offset hyperplanes. However, some of the trained hyperplanes below will have a non-zero offset as well.

1. Suppose we have an SVM that has already been trained with the following datapoints.

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

   As usual, the data points are denoted by $x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix}$. The normal vector of the hyperplane, i.e., the vector that is normal to the hyperplane, is denoted by w. The model was trained with the result that the normal vector is $w = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ and the offset is 0.

   (a) Which of the points are found on the SVM margins?
   Hint: The distance between a point u and the zero-offset hyperplane can be written as $\frac{|w^T u|}{||w||}$. For a point that is "on the SVM margins", there is no other point that has a shorter distance to the hyperplane.

   (b) Suppose we introduce another point, $x^{(6)}$, with features $[-5, 1]^T$ and label -1, then retrain the SVM. Will the learned model change?

   (c) Let's remove data point $x^{(2)}$ and retrain. What will happen to the model? **Hint:** The offset of the hyperplane can also change during retraining.

   (d) How would the results differ when we remove $x^{(3)}$ instead of $x^{(2)}$ and retrain the model?

2. Now that we have seen examples of how to determine which points are on the margins, let us consider the general case. For SVMs, the optimal decision boundary is the one with the "fattest margin". Here, we only consider zero-offset hyperplanes. Assume that the data is linearly separable.

(a) Write down the expression for the smallest distance of all points to a hyperplane defined by some w.

(b) Write down the expression for maximising the smallest distance of all points to a hyperplane defined by some w.

(c) Does this expression satisfy the correct classification constraints of the SVM, i.e., do the points lie on the correct side of the hyperplane.

(d) (Optional) Using (a), show that the SVM optimisation problem can be stated as follows: $\min_w \frac{1}{2}||w||^2$ subject to the constraints that $\forall i \; y^{(i)}(w^T x^{(i)}) \geq 1$.

Hint: Because w is perpendicular to the decision boundary, we can freely scale w by any constant factor without affecting the underlying boundary.

C 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 regularisation. In this problem, we will investigate how loss varies with the number of training samples under different conditions. Consider a dataset where we know that the "correct" hypothesis is a 2nd-degree polynomial.

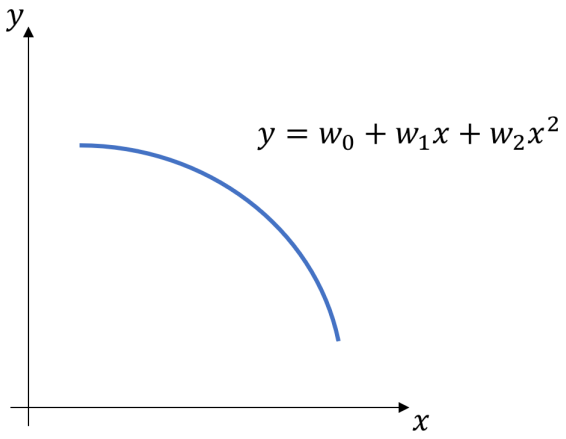

Figure 3: True hypothesis

1. 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. $H_w(x) = w_0 + w_1x$
   2. $H_w(x) = w_0 + w_1x + w_2x^2 + \cdots + w_{10}x^{10}$

   The training and test errors obtained were plotted for each model, and the resulting graphs are shown below as placeholder models A and B:

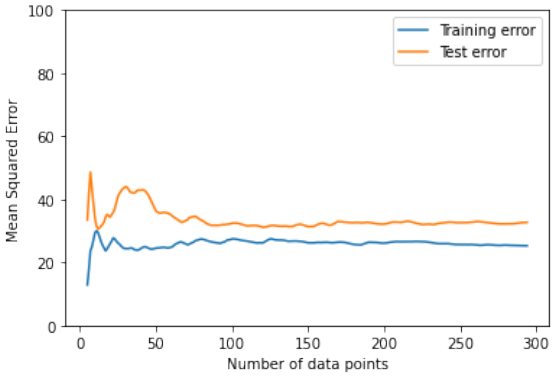

Figure 4: Learning curves for Model A

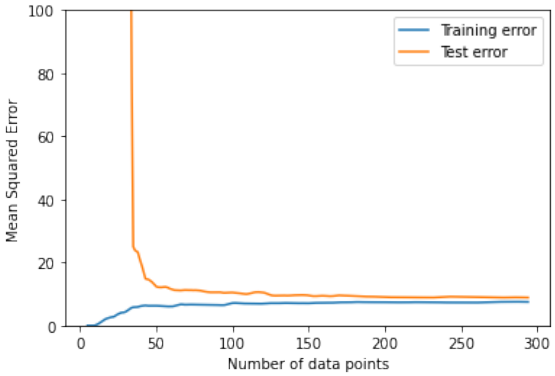

Figure 5: Learning curves for Model B

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

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

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

(d) (Optional) The models above are un-regularised. How might regularisation affect the graphs for each of them?

(e) (Optional) Can you think of a model trained on the dataset that has both a high bias and high variance?