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 $\mathbf{w}$

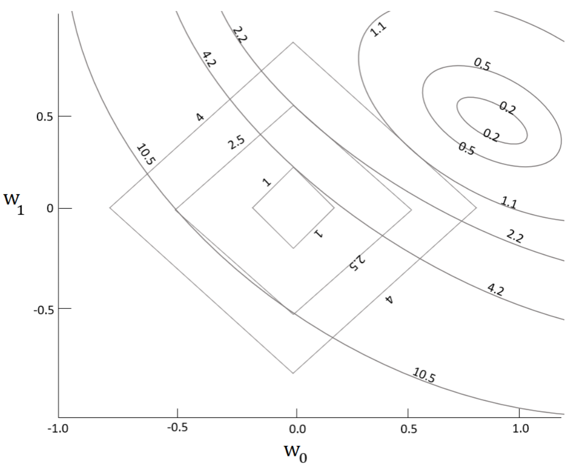

Figure 1: Contour plots for the linear regression problem with L1 regularisation

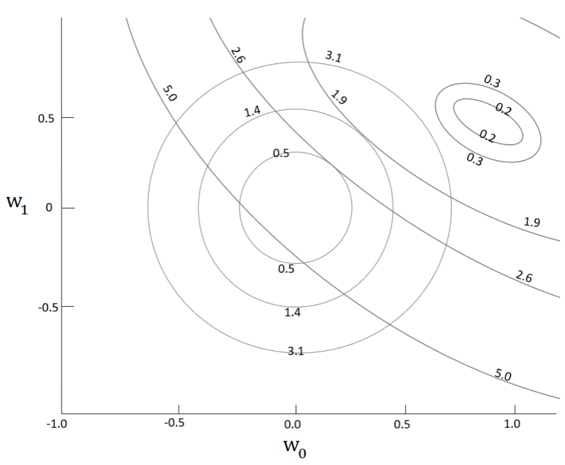

Figure 2: Contour plots for the linear regression problem with L2 regularisation

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.

   > **Solution:**
   >
   > $w_0 = 0.9$, $w_1 = 0.5$. Cost: approx 0 (no MSE and no regularisation penalty).

   (b) L1 regularisation with $\lambda = 5$.

   > **Solution:**
   >
   > $w_0 = 0.0$, $w_1 = 0.5$. Cost: 4.7 = 2.2 (MSE) + 2.5 (L1 penalty).

   (c) L2 regularisation with $\lambda = 5$.

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

> **Solution:**
>
> You have to check for different values of regularisation cost and squared error term to see which gives the smallest sum. Also note that the contours (as given in

the problem description,) already factor in $\lambda = 5$, so there is no need to multiply by λ when computing the sum.

2. How does L2 Regularisation differ from L1 Regularisation in terms of what they do to the parameters?

Solution:

In contrast to L1 regularisation, L2 regularisation heavily penalises larger parameters. However, there are diminishing returns for shrinking parameters that are are already small. So all parameters shrink toward but not exactly to zero.

The penalty grows at the same rate no matter how big or small the parameter for L1 regularisation. If a parameter's contribution to reducing the loss isn't strong enough to justify its penalty, the optimal solution sets it to exactly zero. Meanwhile, parameters that impact the loss significantly would be less penalised than L2 and reduced by less.

Observe that by setting some parameter values to zero, L1 regularisation implicitly selects features to be 'excluded' and behaves as a feature selection strategy.

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

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

where the kernel function $k(\mathbf{x}^{(i)}, \mathbf{x})$ for this tutorial is simply the dot product, $k(\mathbf{x}^{(i)}, \mathbf{x}) = \mathbf{x}^{(i)} \cdot \mathbf{x}$ 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. 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.

Solution:

The distances can be calculated by first using the dot product between w and each data point $x^{(i)}$, which is as follows

$$X = \begin{bmatrix} -2 & -2 \\ -2 & 0 \\ 0 & 2 \\ 1 & 1 \\ 3 & 0 \end{bmatrix}, Xw = \begin{bmatrix} -2 \\ -1 \\ 1 \\ 1 \\ 1.5 \end{bmatrix},$$

To obtain the distances normalise with respect to the length of w as

$$Distances = \frac{|Xw|}{||w||} = \sqrt{2} \begin{bmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 1.5 \end{bmatrix}.$$

The points that lie on the margins are as follows: $x^{(2)}$, $x^{(3)}$, $x^{(4)}$.

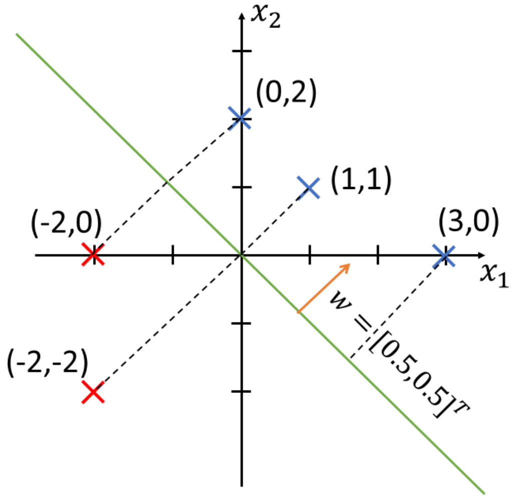

Figure: Plot of data points relative to decision boundary

(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?

Solution:

The distance of $x^{(6)}$ is $\frac{|w^T x^{(6)}|}{||w||} = 2\sqrt{2}$. Given that the distance of $x^{(6)}$ is greater than those found on the margin, it is not a support vector. Thus, it has no impact on the final model. This property of SVMs is what allows them to be sparse as only the parameters for the support vectors are needed when making predictions.

(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.

Solution:

Given that $x^{(2)}$ is a support vector found on the negative margin for the pre-computed SVM, removing it would mean that the model needs to use a different negative datapoint, $x^{(1)}$, to construct the margin. The result is a different model where the hyperplane is shifted in parallel to go through the midpoint between $x^{(1)}$ and $x^{(4)}$. The hyperplane has non-zero offset, which we have not discussed mathematically in class.

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

Solution:

The hyperplane will then go through the midpoint between $x^{(2)}$ and $x^{(4)}$, and perpendicular to the line segment between these two points.

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.

Solution:

The expression is

$$\min_i \frac{\left|w^T x^{(i)}\right|}{||w||}.$$

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

Solution:

The expression is

$$\max_w \left(\min_i \frac{\left|w^T x^{(i)}\right|}{||w||} \right).$$

Intuitively, we can read this formula in 2 parts:

1. Search across all data points to find the point closest to the current decision boundary defined by w.
2. Change the value of w in order to maximise this distance.

(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.

Solution:

No. All points could lie on one side of the hyperplane. The label $y^{(i)}$ does not appear in the expression. Only the distances appear. However, when we impose the correct constraints on the optimization then we can nevertheless use this min-max formulation.

(d) (Optional) Using (c), 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.

Solution:

For every w, there exists at least one point $x^{(i)}$ that has the closest distance, hence is on the margin. Let us scale the length of w by some factor so that

$\left|w^T x^{(i)}\right| = 1$. Now, we can guarantee that the value for the distance is $\frac{1}{||w||}$. The optimisation problem can now be simplified to $\max\limits_{w} \frac{1}{||w||}$.

In addition, we need to introduce constraints to ensure correct classification of the training points. First, we make sure that that the distances for the other points are scaled by the same factor as in the previous paragraph. We can do this by setting constraints that $\forall i \left|w^T x^{(i)}\right| \geq 1$. This ensures that points not on the margins have a distance $\geq \frac{1}{||w||}$ from the hyperplane.

For a correct hyperplane w, when $y^{(i)} = -1$, $w^T x^{(i)}$ is also -1 as the data point must be on the negative side of the decision boundary and vice versa when $y^{(i)} = 1$. Thus we rewrite the constraints $\forall i \left|w^T x^{(i)}\right| \geq 1$ to $\forall i\ y^{(i)}(w^T x^{(i)}) \geq 1$. These are the constraints we have to impose to achieve a correct hyperplane w.

Above, we simplified the unconstrained optimisation problem to $\max\limits_{w} \frac{1}{||w||}$. The inverse of $||w||$ is not so easy to optimize (for example, compute the partial derivative). Instead of the maximising the inverse, we can equivalently minimise the value of $||w||$. Additionally we can multiply by a scaling factor $\frac{1}{2}$ and square $||w||$ to make solving for w easier, resulting in the final form $\min \frac{1}{2}||w||^2$ subject to the constraint that $\forall i\ y^{(i)}(w^T x^{(i)}) \geq 1$.

While beyond the scope of CS2109s, solutions for the problem can easily be now found using methods in constraint optimisation, particularly Quadratic Solvers.

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.

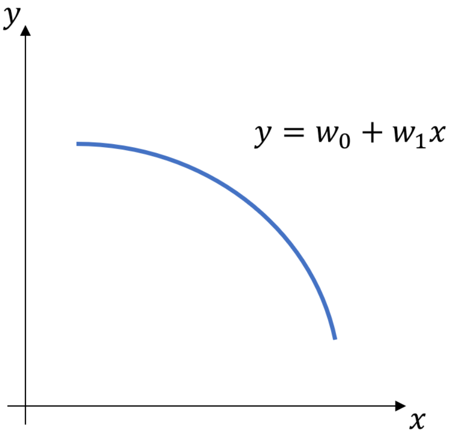

Figure 3: True hypothesis

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. $H_w(x) = w_0 + w_1 x$
   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 shown below as placeholder models A and B:

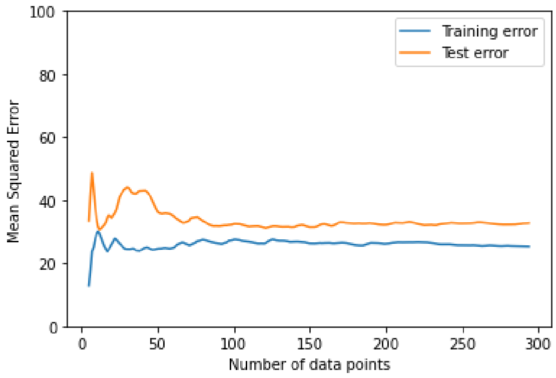

Figure 4: Learning curves for Model A

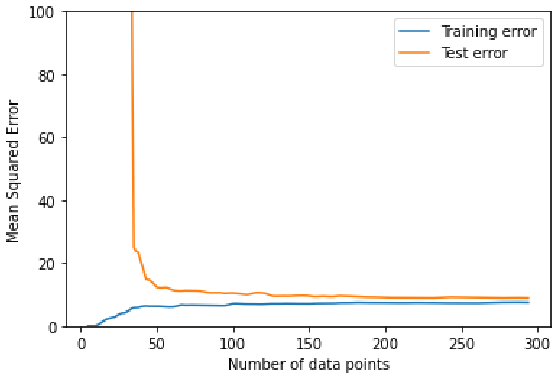

Figure 5: Learning curves for Model B

   (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?

Solution:

Model A. Relatively higher error, even as data points increase, indicates inability to capture the true relationship sufficiently, hinting high bias. In general, bias does not improve with an increased 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?

Solution:

Model B. Lower error, but an initial larger difference between the training and testing error indicates high variance. Obtaining more data points is likely to decrease variance.

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

Solution:

Model A (high bias) is the linear model, because: linear model can't capture quadratic relationship, has high bias. Model B (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.

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

Solution:

For both models, regularisation might help to reduce the testing error by preventing overfitting at the cost of increasing training error. Regularisation is more relevant for Model B, as it has a higher variance and is more likely to overfit to the quadratic curve.

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

Solution:

In theory, here is typically a trade off between bias and variance. Beyond this setting, there are real world examples that lead to high bias and high variance situations. These could potentially include

1. Weak dummy classifiers that are not trained to optimise a loss function. Some examples could include using the max value of a training dataset for predictions
2. A model trained with a significant amount of outliers from a random

distribution distinct from the training data.

3. Perturbing the optimal results of a classifier by a random amount each run

While these examples may be arbitrary, they convey that bias and variance is not completely a trade off and certain modelling choices or mistaken assumptions in data could lead to deprovements in both.