NUS | Computing

National University of Singapore

Eric Han
eric_han@nus.edu.sg
https://eric-han.com

Computer Science

T06 – 17 Oct 2024

Week 9

CS2109s TG35,36

Content I

Section 1: **Visualising Regularisation**

Figure plots the loss (regularization / error) respectively; objective is to find the smallest.

- L1: $J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{n} |w_i|\right]$
- L2: $J(w) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{n} {w_i}^2\right]$

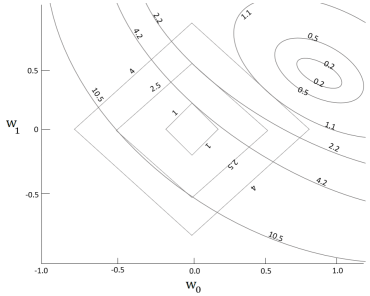

Figure 1: LR with L1 Reg.

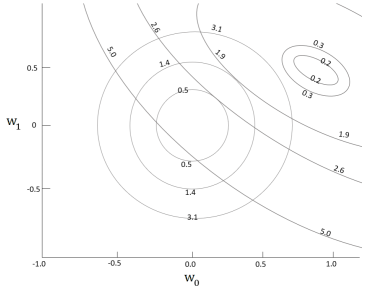

Figure 2: LR with L2 Reg.

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 $\lambda = 5$.
   d. [@] Why does L1 often cause values to go to zero?
2. How does L2 Regularisation differ from L1 Regularisation in terms of what they do to the parameters?

Recap

- How to read this graph?
  - Versus 1D error graphs?
- What is the difference between L1 and L2 reg.
- Ridge vs Lasso

Answer

1. Find the point (w_0, w_1) with the smallest Cost
   a. $(0.9, 0.5)$, Cost: approx 0 (no MSE and no regularization penalty).
   b. $(0.0, 0.5)$, Cost: 4.7 = 2.2 (MSE) + 2.5 (L1 penalty).
   c. $(0.2, 0.25)$, Cost: 3.1 = 2.6 (MSE) + 0.5 (L2 penalty).

Answer

1. Find the point (w_0, w_1) with the smallest Cost
   a. $(0.9, 0.5)$, Cost: approx 0 (no MSE and no regularization penalty).
   b. $(0.0, 0.5)$, Cost: 4.7 = 2.2 (MSE) + 2.5 (L1 penalty).
   c. $(0.2, 0.25)$, Cost: 3.1 = 2.6 (MSE) + 0.5 (L2 penalty).
   d. Why does L1 often cause values to go to zero?:
      - L1 has absolute values, which means it has a discontinuity at 0, which means that any optimization that cross 0 to be zeroed out. Effectively feature selection.
      - L2 heavily penalizes larger parameters, prefering smaller values.
2. L2 Regularisation vs L1 Regularisation on parameters
   - » L2 penalizes larger parameters (Where did we see this prior?)

Answer

1. Find the point (w_0, w_1) with the smallest Cost
   a. $(0.9, 0.5)$, Cost: approx 0 (no MSE and no regularization penalty).
   b. $(0.0, 0.5)$, Cost: 4.7 = 2.2 (MSE) + 2.5 (L1 penalty).
   c. $(0.2, 0.25)$, Cost: 3.1 = 2.6 (MSE) + 0.5 (L2 penalty).
   d. Why does L1 often cause values to go to zero?:
      - L1 has absolute values, which means it has a discontinuity at 0, which means that any optimization that cross 0 to be zeroed out. Effectively feature selection.
      - L2 heavily penalizes larger parameters, prefering smaller values.
2. L2 Regularisation vs L1 Regularisation on parameters
   - L2 penalizes larger parameters (Where did we see this prior?) MSE vs MAE
   - L1 may just set certain parameters to zero, ie. feature selection.

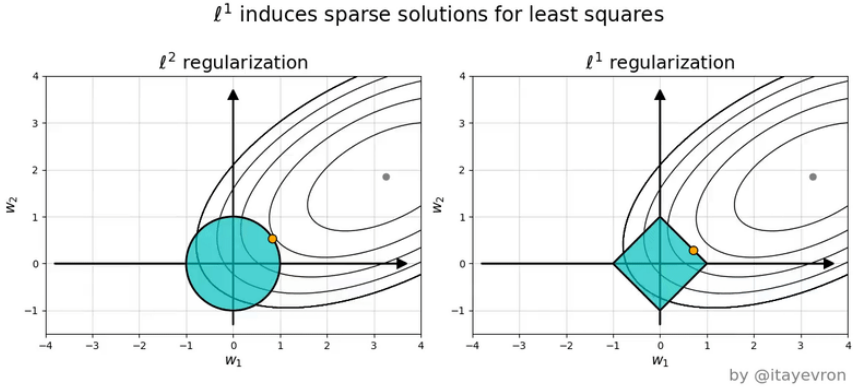

Figure 3: https://eric-han.com/teaching/AY2425S1/CS2109s/T06.week-9_regularization-and-validation_ridge-lasso.gif

Section 2: **SVM and Hinge Loss**

Question 1 [G]

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

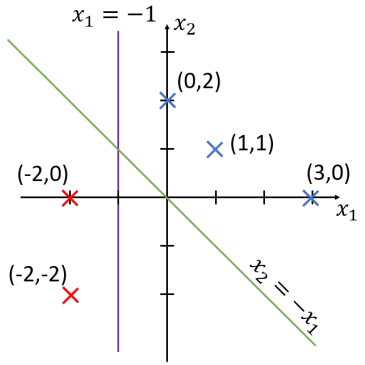

Figure 4: SVM example.

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

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?

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. [@] Solve $\max_\alpha \sum_{i=1}^n \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha^{(i)} \alpha^{(j)} \bar{y}^{(i)} \bar{y}^{(j)} \left(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \right)$ 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 $\mathbf{w}$ found in part **(b)** and the one calculated here?

Recap

- What is SVM, what is the intuition behind SVM?
- What are support vectors vs non-support vectors?
  - How to identify?

Answer 1a - Green

From equation given, we change the form to SVM's form ($\mathbf{w} \cdot \mathbf{x} + b = 0$)

$$x_2 = -x_1 \implies x_1 + x_2 = 0 \implies [1,1]^T \cdot [x_1, x_2] + 0 = 0$$

We know that $\mathbf{w}$ is constraint by margin M, where $\frac{2}{|\mathbf{w}|} = M$ (given), from geometry we can calculate the margin by taking the distance between $(0, 2)$ and $(-2, 0)$.

$$M = \sqrt{(-2-0)^2 + (0-2)^2} = 2\sqrt{2} \implies |\mathbf{w}| = \frac{2}{2\sqrt{2}} = \frac{1}{\sqrt{2}}$$

We need to scale the equation $[1,1]^T \cdot \mathbf{x} = 0$ such that $|\mathbf{w}| = \frac{1}{\sqrt{2}}$, we can scale by c and solve for c (you can also scale to unit vector then scale it up):

$$[c,c]^T \cdot \mathbf{x} = 0 \times c \implies |[c,c]| = \sqrt{c^2 + c^2} = \sqrt{2}c = \frac{1}{\sqrt{2}} \implies c = \frac{1}{2}$$

Hence, $\mathbf{w} = [\frac{1}{2}, \frac{1}{2}]^T$ and $b = 0$.

From graph, the -ve gutter will run through $(-2, 0)$ and $+$ve gutter will run through $(0, 2), (1, 1)$, so they may be support vectors.

Hence, the (obviously) non-support: $(-2, -2)$ and $(3, 0)$, so $\alpha^{(1)} = \alpha^{(5)} = 0$.

So, using the constraint $\mathbf{w} = \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)}$ (given):

$$\mathbf{w} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = -\alpha^{(2)} \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \alpha^{(3)} \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \alpha^{(4)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies \begin{bmatrix} 2\alpha^{(2)} + \alpha^{(4)} \\ 2\alpha^{(3)} + \alpha^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

Then we also use the constraint $\sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} = 0$ (given):

$$-\alpha^{(2)} + \alpha^{(3)} + \alpha^{(4)} = 0$$

With the 3 equations, we can solve our linear equations:

$$\begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha^{(2)} \\ \alpha^{(3)} \\ \alpha^{(4)} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix} \implies \left[\begin{array}{ccc|c} 2 & 0 & 1 & 0.5 \\ 0 & 2 & 1 & 0.5 \\ -1 & 1 & 1 & 0 \end{array}\right]$$

Hence, solving it gives $\alpha^{(2)} = \alpha^{(3)} = \frac{1}{4}$ and $\alpha^{(1)} = \alpha^{(4)} = \alpha^{(5)} = 0$.

From constraint $\alpha^{(2)} \geq 0, \quad \alpha^{(3)} \geq 0, \quad \alpha^{(4)} \geq 0$ (given), we see that the solution is valid.

Therefore, assuming hard margin SVM, the green line is a valid solution.

Answer 1a - Purple

$$\mathbf{w}^T \cdot [x_1, x_2] + b = 0 \implies \mathbf{w} = [c, 0]^T, b = c, c \in \mathbb{R}, \frac{2}{|\mathbf{w}|} = 2 \implies \mathbf{w} = [1, 0]^T$$

Identify the (obviously) non-support: $(1, 1)$ and $(3, 0)$, so $\alpha^{(4)} = \alpha^{(5)} = 0$,
So, using the $\mathbf{w} = \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)}, \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} = 0$ (in lectures) and constraints:

$$\mathbf{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = -\alpha^{(1)} \begin{bmatrix} -2 \\ -2 \end{bmatrix} - \alpha^{(2)} \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \alpha^{(3)} \begin{bmatrix} 0 \\ 2 \end{bmatrix} \implies \begin{bmatrix} \alpha^{(1)} + \alpha^{(2)} \\ \alpha^{(1)} + \alpha^{(3)} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix}$$

$$-\alpha^{(1)} - \alpha^{(2)} + \alpha^{(3)} = 0$$

$$\alpha^{(1)} \geq 0, \quad \alpha^{(2)} \geq 0, \quad \alpha^{(3)} \geq 0$$

Hence, $\alpha^{(3)} = 1/2, \alpha^{(1)} = -1/2, \alpha^{(2)} = 1 \implies$ Contradiction $\alpha^{(1)} \geq 0$

Answer 1b

Assuming $\mathbf{w} = (k, k)$ *and* $b = 0$. *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* $k = 0.5$.

Answer 1c

By observation: Green is better.

Intuition: Both lines completely separate without mislabels, Green has max margin:

- Green: $2 \times \sqrt{2}$
- Purple: 2×1 (And not a converged SVM line)

Note

- Margin can be calculated by geometry
- Or by $\frac{2}{|w|}$
- Do not use loss argument here

Answer 1d

From (1a), it is possible to partially solve, such that $\alpha^{(2)} = \alpha^{(3)} = k$ (instead of solving for $k = 1/4$) and that $\alpha^{(1)} = \alpha^{(4)} = \alpha^{(5)} = 0$:

$$\begin{aligned}
&\max_{\alpha} \sum_{i=1}^{n} \alpha^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha^{(i)} \alpha^{(j)} \bar{y}^{(i)} \bar{y}^{(j)} \left(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \right) \\
&= \max_{\alpha} \alpha^{(2)} + \alpha^{(3)} - \frac{1}{2} [\alpha^{(2)} \alpha^{(2)} \bar{y}^{(2)} \bar{y}^{(2)} \left(\mathbf{x}^{(2)} \cdot \mathbf{x}^{(2)} \right) + \alpha^{(2)} \alpha^{(3)} \bar{y}^{(2)} \bar{y}^{(3)} \left(\mathbf{x}^{(2)} \cdot \mathbf{x}^{(3)} \right) \\
&\qquad + \alpha^{(3)} \alpha^{(2)} \bar{y}^{(3)} \bar{y}^{(2)} \left(\mathbf{x}^{(3)} \cdot \mathbf{x}^{(2)} \right) + \alpha^{(3)} \alpha^{(3)} \bar{y}^{(3)} \bar{y}^{(3)} \left(\mathbf{x}^{(3)} \cdot \mathbf{x}^{(3)} \right)] \\
&= \max_{k} k + k - \frac{1}{2} [(k)(k)(-1)(-1) \begin{bmatrix} -2 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 0 \end{bmatrix} + (k)(k)(-1)(1) \begin{bmatrix} -2 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\
&\qquad + (k)(k)(1)(-1) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 0 \end{bmatrix} + (k)(k)(1)(1) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \end{bmatrix}] \\
&= \max_{k} 2k - \frac{1}{2} [4k^2 + 0 + 0 + 4k^2] \\
&= \max_{k} 2k - 4k^2
\end{aligned}$$

Differentiating and setting to 0, we get $k = \frac{1}{4}$.

$$\begin{aligned}
\mathbf{w} &= \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)} \\
&= \frac{1}{4}(-1) \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \frac{1}{4}(1) \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\
&= \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}
\end{aligned}$$

Hinge loss was used as part of the derivation, not suprising to reach same answer as 2a.

Section 3: **Bias & Variance**

Two model hypotheses:

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

With the 2 training/test error learning curves:

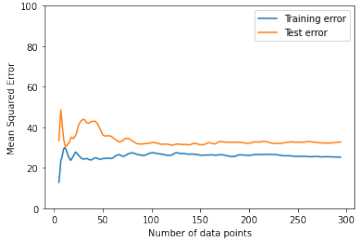

Figure 5: Model A.

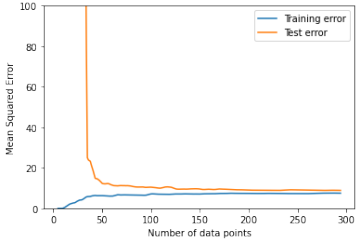

Figure 6: Model B.

a. Which graph indicates model with a higher bias?
   - » How does bias seem to vary with the number of samples?

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

c. Which hypotheses (1,2) belong to the model (A,B)? Why?

d. [@] How might regularization affect the graphs for each of them?

Recap

- What is bias?
- What is variance?
- What is the relation with model complexity?

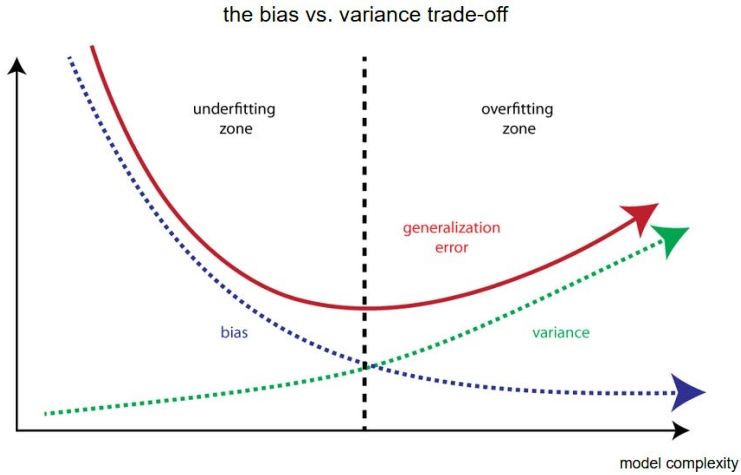

Figure 7: Bias-Variance Intuition.

Answer

a. Model A. 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.

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

c. 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. Regularizaion would greatly benefit the more complex model by combating overfitting ie. for L1 feature selection of the polynomial terms for model 2.

Section 4: **Gaussian Kernel**

Proof that the Gaussian Kernel has Infinite Dimensional Features

a. [@] How does it relate to RBF? How can we invent new kernels with this property?

Recap

- Kernels have special powers (from previous tutorials)

Answer

We simply to $K(x, x') = \exp(-(x - x')^2)$, then use Taylor series of e^x, at $x = 0$.

$$\begin{aligned}
\exp(-(x - x')^2) &= \exp(-x^2) \times \exp(-x'^2) \times \exp(2xx') \\
&= \exp(-x^2)\exp(-x'^2)\left[1 + 2xx' + \frac{2^2 x'^2 x^2}{2!} + \cdots\right] \\
&= \exp(-x^2)\exp(-x'^2)\sum_{k=0}^{\infty}\frac{2^k x'^k x^k}{k!} \\
&= \sum_{k=0}^{\infty}\left[\sqrt{\frac{2^k}{k!}}\exp(-x^2)x^k \times \sqrt{\frac{2^k}{k!}}\exp(-x'^2)x'^k\right]
\end{aligned}$$

- Formed by taking an infinite sum (dot product) over polynomial kernels
- Map the current vector into an infinite dim. space and compute the distance.
- Though this is an infinite dimensional space, each variable is highly constrained, so the solution space is not exactly totally unbounded.

To help you further your understanding, not compulsory; Work for Snack/EXP!

Tasks

1. Implement SVM from scratch (numpy) to solve green line, no boilerplate code given.
   - Hint: Search for tutorial
   - Don't need to implement Optimizers, just use optimizers within numpy/scipy.

1. [@] and Bonus declaration is to be done here; You should show bonus to Eric.
2. Attempted tutorial should come with proof (sketches, workings etc...)
3. Random checks may be conducted.
4. Guest student should come and inform me.

Figure 8: Buddy Attendance: https://forms.gle/q5Secb3dHshmXNXd7

References I