CS2109s - Tutorial 6

Eric Han

Oct 18, 2023

Annoucements

Important admin

1. PS5 Task 3.6.1: some of the submissions got issues of non-convergence, if your graph looks weird / not converging:
    - Pick 2000 data
    - Normalise data

Question 1 [G]

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

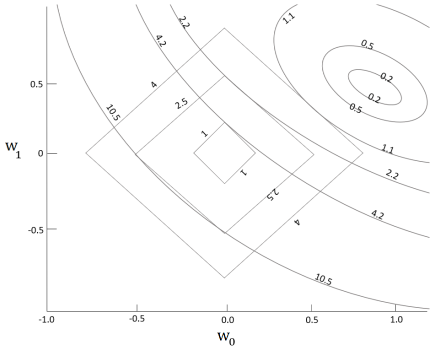


Figure 1: LR with L1 Reg.

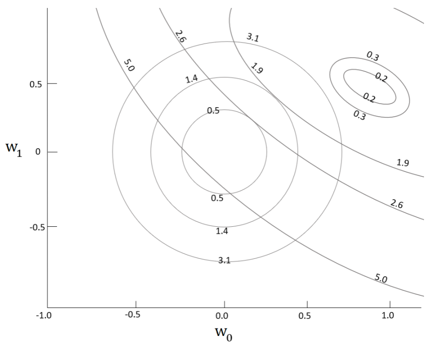


Figure 2: LR with L2 Reg.

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?

Recap

- How to read this graph (similar to the 1D error graphs but 2D)
- What is the difference between L1 and L2 reg.

Answer

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).

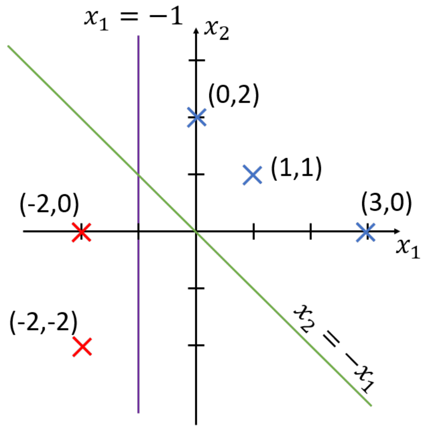
. . .

Note:

- 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.

Question 2 [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 $h(x_1, x_2) = \sum_{i=1}^{5} \alpha^{(i)} \bar{y}^{(i)} \left(x_1 x_1^{(i)} + x_2 x_2^{(i)} \right) + b$.



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

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?
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 2a - Green

$$(k, -k) \; \forall k \in \mathbb{R}$$

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

So, using boundary condition and constrains:

$$-\alpha^{(2)}(-2k+0)+\alpha^{(3)}(0-2k)+\alpha^{(4)}(k-k)+b=0, \quad b=0$$
$$-\alpha^{(2)}+\alpha^{(3)}+\alpha^{(4)}=0$$
$$\alpha^{(2)} \geq 0, \quad \alpha^{(3)} \geq 0, \quad \alpha^{(4)} \geq 0$$

Hence, $\alpha^{(2)} = \alpha^{(3)} = k_\alpha, \ \forall k_\alpha \in \mathbb{R}^+$ and $\alpha^{(1)} = \alpha^{(4)} = \alpha^{(5)} = 0$.

Pro-tip: k means a lot of different things here but it is used as an arbitary variable.

Answer 2a - Purple

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

Identify the (obviously) non-support: $(1, 1)$ and $(3, 0)$, so $\alpha^{(4)} = \alpha^{(5)} = 0$,

So, using the $\mathbf{w} = \sum_{i=1}^{n} \alpha^{(i)} \bar{y}^{(i)} \mathbf{x}^{(i)}$ (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 2b

Assuming $\mathbf{w} = (k, k)$ *and* $b = 0$. *Hinge loss is:*

$$\max(0, 1-4k) + \max(0, 1-2k) + \max(0, 1-2k) + \max(0, 1-2k) + \max(0, 1-3k)$$
$$= 3 \cdot \max(0, 1-2k) + \max(0, 1-3k) + \max(0, 1-4k)$$

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 2c

By observation: Green is better.

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

- Green: $2 \times \sqrt{2}$
- Purple: 2×1

Note

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

Answer 2d

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

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 resolve to the same.

Question 3 [G]

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:

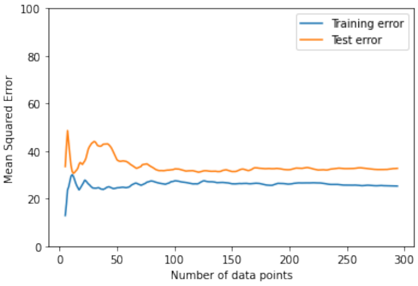


Figure 4: Model X.

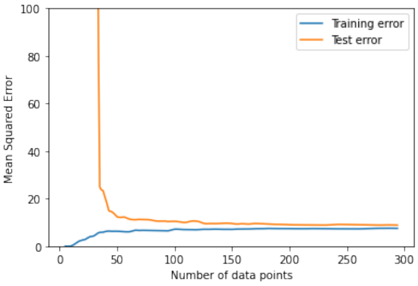


Figure 5: Model Y.

---

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 (X,Y)? 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?

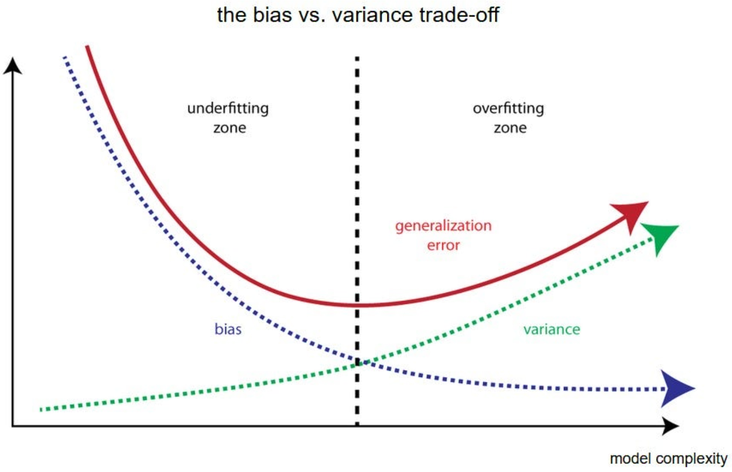


Figure 6: Bias-Variance Intuition.

Answer

a. 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.
b. Model Y. Lower error, but initially higher difference between the 2 error indicates high variance. Getting more data points is likely to help variance.
c. 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.
d. Regularizaion would greatly benefit the more complex model by combating overfitting ie. for L1 feature selection of the polynomial terms for model 2.

Question 4 [G]

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 $x = 0$.

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

- 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 dimentional space, each variable is highly constrainted, so the solution space is not exactly totally unbounded.

Bonus Qn

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

Tasks

1. Implement SVM from scratch (numpy) to solve Q2, no boilerplate code given.

Buddy Attendance Taking

Take Attendance for your buddy: https://forms.gle/Ckkq639TNwWEx3NT6

1. Random checks will be conducted - `python ../checks.py TG0`

Figure 7: Buddy Attendance