

IN4320 Machine Learning Assignment 1

Mingxi Li(4735366)

February 24, 2018

1

1.a

The loss function is

$$L(m_+) = \sum_{i=1}^{N^-} \frac{1}{N^-} \|x_i^- - 1\|^2 + \sum_{i=1}^{N^+} \frac{1}{N^+} \|x_i^+ - r_+\|^2 + \lambda \|1 - r_+\|_1 \quad (1)$$

Since we have no observation for - class, we assume the first term is a constant term C:

$$L(m_+) = r_+^2 + 1 + \lambda |1 - r_+| + C \quad (2)$$

Figure 1 shows the loss function as a function of r_+ for all $\lambda \in \{0, 1, 2, 3\}$.

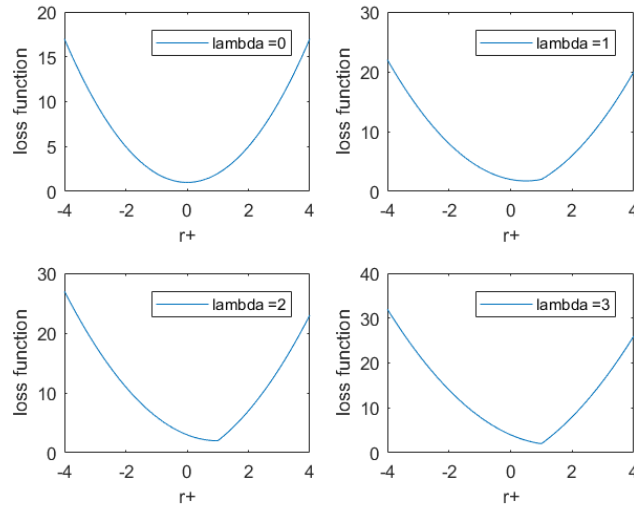


Figure 1: The loss function as a function of r_+ for all $\lambda \in \{0, 1, 2, 3\}$.

1.b

Derive for every of the four functions the minimizer and their minimum values. Also determine all points where the derivative equals 0.

1. The minimizers and their minimum values

- (a) $\lambda = 0, \min = 1, \text{minimizer} = (0, 1 + C)$
- (b) $\lambda = 1, \min = \frac{7}{4}, \text{minimizer} = (\frac{1}{2}, \frac{7}{4} + C)$
- (c) $\lambda = 2, \min = 2, \text{minimizer} = (1, 2 + C)$
- (d) $\lambda = 3, \min = 2, \text{minimizer} = (1, 2 + C)$

2. The points where the derivative equals 0

- (a) $\lambda = 0 : (0, 1 + C)$
- (b) $\lambda = 1 : (\frac{1}{2}, \frac{7}{4} + C)$
- (c) $\lambda = 2 : (1, 2 + C)$
- (d) $\lambda = 3 : \text{The derivative can not equal 0.}$

2

The regularizer in Equation (1) tries to penalize the L1 distance between r_+ and r_- . If λ gets larger and larger, the regularization term accounts for a dominant position eventually. If λ is large enough, we can only get minimum value of loss function when $r_+ = r_-$, which means the L1 distance becomes 0. In this case, the model can not classify anything, which is underfitting.

3

3.a

It consists of two parts. The regularized loss function is a set of concentric circles. They have the same center point with different size. The regularized term is a square. Vertices of the square are on the coordinate axes.

3.b

If λ is large enough, we can only get the minimum value when $r_+ = r_-$. The loss function is $L(r) = 2r^2 - 2r + 6$ where $r = r_+ = r_-$. Finally, we get the minimum value when $r_+ = r_- = \frac{1}{2}$.

4

4.a

Gradient decent algorithm is implemented as optimization algorithm. The main idea is if we take a step towards the minus direction at every turn, we'll at the bottom eventually. The main challenges are how to calculate gradients of regularization term and choose learning rate. To solve the former, we can use `sign()` function to get the gradient of L1. As for the latter, we need to search a proper learning rate so that we can get the minimal value eventually within a short time. I set learning rate as a function of the number of executing iterations, that is $\alpha = \frac{1}{k^{0.8}}$. In several rounds at the beginning, learning rate is big so we take a long step. Finally when we reach the bottom, the value of α becomes very small so that we won't miss the minimum.

4.b

Figure 2 shows representor images for $\lambda = 0$.

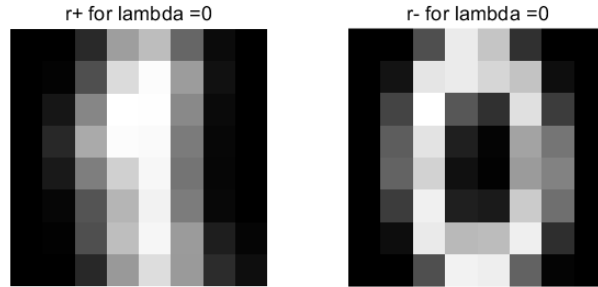


Figure 2: Represantor images for $\lambda = 0$

When λ gets around 300, the images don't change any more. Figure 3 shows represantor images for $\lambda = 300$.

We can find that when λ is large enough, images of r_- and r_+ are identical.

4.c

I repeat the experiment 100 times. Each time, I select one single training example per class randomly. Then calculate apparent error and true error. Finally, get 100 apparent errors and 100 true errors. I use the mean apparent and true error and draw regularization curves as shown in Figure 4.

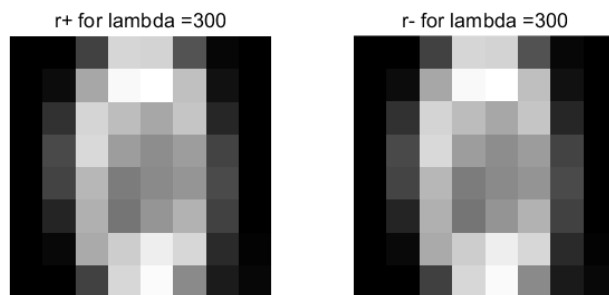


Figure 3: Representer images for $\lambda = 300$

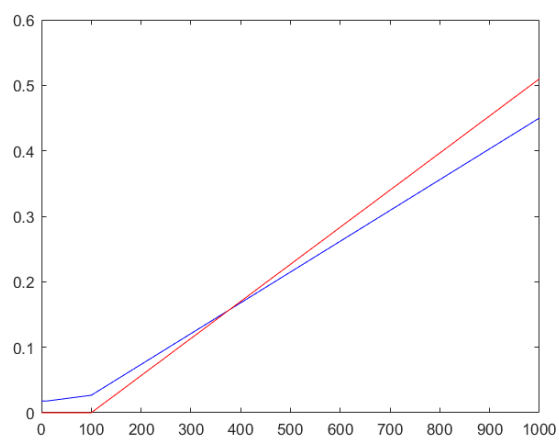


Figure 4: The apparent and true error against $\lambda \in \{0, \frac{1}{10}, 1, 10, 100, 1000\}$