CS146 HW2

Terry Ye

November 10, 2018

1 Question 1

answered in CCLE

2 Question 2

answered in CCLE

3 Question 3

(a) Let $p^+ = min(\omega^T x_i + \theta)\forall(x_i, y_i)$ that $y_i = 1$. Let $p^- = max(\omega^T x_i + \theta)\forall(x_i, y_i)$ that $y_i = -1$. Because D is linearly separable, then $p^+ \geq 0$ and $p^- < 0$,then $\exists\epsilon \geq 0$ such that $p^+ - \epsilon \geq 0 > p^- - \epsilon$. So $(\omega^T x_i + \theta - \epsilon)$ also separate the data set. Let ϵ be the one that makes $(\omega^T x_i + \theta)$ equally distant from p^+ and p^-, by calculation we get $\epsilon = \frac{p^+ + p^-}{2}$. So the new separation line is

$$\omega^T x_i + \theta - \frac{p^+ + p^-}{2})$$

By definition, $min(\omega^T x_i + \theta) = p^+$ if y = 1 and $max(\omega^T x_i + \theta) = p^-$ if y = -1. So the new min if y = 1 is $p^+ - \epsilon = \frac{p^+ - p^-}{2}$, the new min if y = -1 is $p^- - \epsilon = \frac{p^- - p^+}{2}$, we get

$$y_i(\omega^T x_i + \theta - \frac{p^+ + p^-}{2}) \geq \frac{p^+ - p^-}{2}$$

.

Then let new $\theta = \frac{\theta - \epsilon}{\frac{p^+ - p^-}{2}}$, the new $\omega^T = \frac{\omega^T}{\frac{p^+ - p^-}{2}}$, we get $y_i(\omega^T x_i + \theta) \geq 1$, so this is the optimal solution for $y_i(\omega^T x_i + \theta) \geq 1 - \delta$ because $\delta = 0$.

(b) If $0 < \delta < 1$,we can apply similar proof as above that $y_i(\omega^T x_i) < 0$ when $y_i = -1$ and ¿ 0 when $y_i = 1$,so the data set is linearly separable. If

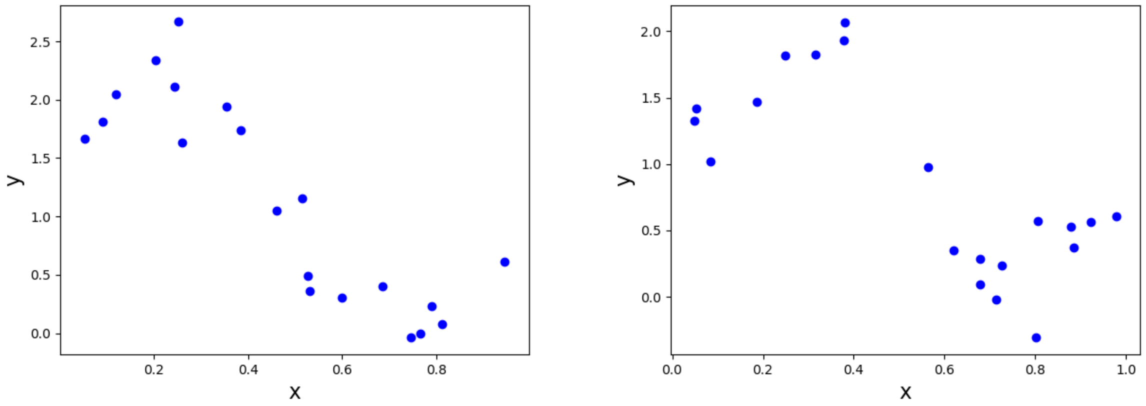
$\delta \geq 1$, then the separability is not sure.

(c) We can choose $\omega = 0, \theta = 0, \delta = 0$, then this will be true regardless of y value, so it satisfies the formula and this is optimal because 0 is the minimal value for δ. However, this is not a valid hyperplane to split, so this formula cannot be used.

(d) The data set is separable, so the $\delta = 0$. And plug in x_1, y_1, x_2, y_2,we get $\omega_1 + \omega_2 + \omega_3 + \theta \geq 1$ and $\omega_1 + \omega_2 + \omega_3 - \theta \leq -1$, so optimal solutions (ω, θ, δ) are when $\delta = 0$ and $\omega_1 + \omega_2 + \omega_3 \geq |1 - \theta|$.

4 Question4

(a)



The linear model will behave ok on training data because the data shows a trend of negative correlation but will behave badly on test data because test data does not show any linear correlation or relation.

(b)I have done this.
(c)I have done this.

(d)
The cost is 40.234 if the coefficent vector is zero vector.

step size	coefficient	iterations	final value of J	time
0.0001	[2.27044798 -2.46064834]	10000	4.086	0.305
0.001	[2.4464068 -2.816353]	7021	3.913	0.218
0.01	[2.44640703 -2.81635347]	765	3.913	0.00965
0.0407	[-9.40470931e+18 -4.65229095e+18]	10000	2.711e+39	0.301

Generally, when the step size increases, the model will converge faster and have a smaller error. But when the step size is too large like 0.0407, it might can't converge because it will diverge.

(e)

The closed-form solution coefficient is [2.44640709 -2.81635359]. The cost is 3.9125764057914636. The solution and the final cost is basically the same as using SD with step size 0.001 and 0.01. The time is 0.00399. In this case, the algorithm run faster than GD because the dimension is not too big due to the limited data size. So calculate inverse of matrix is more efficient than iterating.

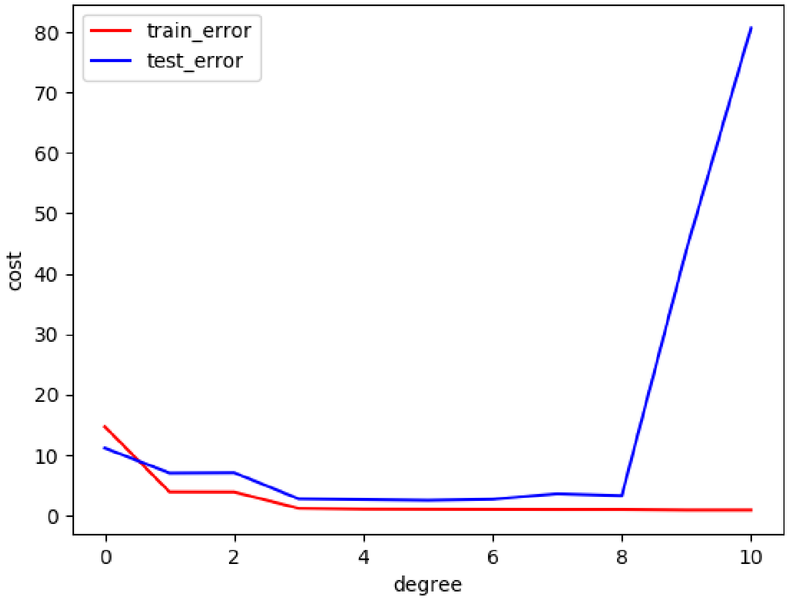
(f)

Using a function of k as learning rate, after 1719 iterations, the new coefficient is [2.44640672 -2.81635282],the cost is 3.9125764057922674 and the time is 0.0520. So the result is similar as using GD with fixed step size 0.01 and 0.001, but faster than using 0.001 and slower than using 0.01.

(g)

RMSE is better than J(w) because it reduces the effect of data size on the error by dividing it. The error function will now demonstrate the wellness of model better.

(h)



The best degree that fits the data is 5. There is evidence of overfitting, when m=10, the train error is really low while the test error is exceptionally high which is a pattern of overfitting.