CM146, Winter 2019
Problem Set 2: Perceptron and regression
Due Feb 12, 2019

Jingjing Nie

1 Problem 1

(a) Problem 1a OR

Solution:

Possible θ solutions:

$$\theta = \{1, 1, 1\}$$

Another possible solution:

$$\theta = \{1, 0.7, 0.7\}$$

(b) Problem 1b XOR

Solution:

A valid perceptron does not exist, because in this XOR situation, the data is not linearly separable.

2 Problem 2

(a) Problem 2a

Solution:

$$\begin{aligned}
\frac{\partial h_\theta(x)}{\partial \theta_j} &= \frac{\partial \sigma(\theta^T x)}{\partial \theta_j} = \frac{\partial(\frac{1}{1+e^{-a}})}{\partial \theta_j^T x} \times \frac{\partial \theta_j^T x}{\partial \theta_j} \\
&= \frac{e^{-a}}{(1+e^{-a})^2} \times x_{n,j} \\
&= \frac{1}{(1+e^{-a})} \times \frac{e^{-a}}{(1+e^{-a})} \times x_{n,j} \\
&= \frac{1}{(1+e^{-a})} \times \left(1 - \frac{1}{(1+e^{-a})}\right) \times x_{n,j} \\
&= \sigma(\theta^T x)(1-\sigma(\theta^T x))x_{n,j}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J}{\partial \theta_j} &= \frac{\partial J}{\partial h_\theta(x)} \times \frac{\partial h_\theta(x)}{\partial \theta_j} \\
&= -\sum_{n=1}^{N} \left(\frac{y_n}{\sigma(\theta^T x)} \times \sigma(\theta^T x)(1-\sigma(\theta^T x))x_{n,j}\right) - \\
&\left(\frac{1-y_n}{1-\sigma(\theta^T x)} \times \sigma(\theta^T x)(1-\sigma(\theta^T x))x_{n,j}\right) \\
&= -\sum_{n=1}^{N} \left(y_n(1-\sigma(\theta^T x))x_{n,j} - (1-y_n)\sigma(\theta^T x)x_{n,j}\right) \\
&= -\sum_{n=1}^{N} (y_n - \sigma(\theta^T x))x_{n,j} \\
&= \sum_{n=1}^{N} (h_\theta(x) - y_n)x_{n,j}
\end{aligned}$$

(b) Problem 2b

Solution:

$$\begin{aligned}
\frac{\partial^2 J}{\partial\theta_j \partial\theta_k} &= \frac{\partial \frac{\partial J}{\partial\theta_j}}{\partial\theta_k} \\
&= \frac{\partial(-\sum_{n=1}^{N}(y_n - h_\theta(x))x_{n,j})}{\partial\theta_k} \\
&= \sum_{n=1}^{N} \frac{\partial h_\theta(x) x_{n,j}}{\partial\theta_k} \\
&= \sum_{n=1}^{N} \sigma(\theta^T x)(1 - \sigma(\theta^T x)) x_{n,k} x_{n,j} \\
&= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n)) x_{n,k} x_{n,j}
\end{aligned}$$

Therefore, the Hessian can be written as:

$$H = \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n)) x_n x_n^T$$

(c) Problem 2c

Solution:

From the answer found above in 2b,

$$H = \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))x_n x_n^T$$

Therefore, for any real vectors z,

$$\begin{aligned} z^T H z &= z^T [\sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))x_n x_n^T] z \\ &= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n)) z^T x_n x_n^T z \\ &= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))(z^T x_n)(z^T x_n)^T \\ &= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))||z^T x_n||^2 \end{aligned}$$

$||z^T x_n||^2 \geq 0$, and $h_\theta(x_n)$ and $(1 - h_\theta(x_n))$ have values that are in the range 0 to 1 (i.e., both are greater than or equal to 0). Thus,

$$\sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))||z^T x_n||^2 \geq 0$$

$$z^T H z \geq 0$$

It can be concluded that J is a convex function as $H \geq 0$.

3 Problem 3

(a) Problem 3a

Solution:

$$\frac{\partial J}{\partial \theta_0} = \sum_{n=1}^{N} 2w_n(\theta_0 + \theta_1 x_{n,1} - y_n)$$

$$\frac{\partial J}{\partial \theta_1} = \sum_{n=1}^{N} 2w_n x_{n,1}(\theta_0 + \theta_1 x_{n,1} - y_n)$$

(b) Problem 3b

Solution:

$$Let:$$

$$\overline{x} = \frac{\sum_{n=1}^{N} w_n x_{n,1}}{\sum_{n=1}^{N} w_n}$$

$$\overline{y} = \frac{\sum_{n=1}^{N} w_n y_n}{\sum_{n=1}^{N} w_n}$$

$$\overline{xy} = \frac{\sum_{n=1}^{N} w_n x_{n,1} y_n}{\sum_{n=1}^{N} w_n}$$

$$\overline{x^2} = \frac{\sum_{n=1}^{N} w_n x_{n,1}^2}{\sum_{n=1}^{N} w_n}$$

$$\begin{aligned}
\frac{\partial J}{\partial \theta_0} &= 0 \\
\sum_{n=1}^{N} 2w_n(\theta_0 + \theta_1 x_{n,1} - y_n) &= 0 \\
\sum_{n=1}^{N} w_n(\theta_0 + \theta_1 x_{n,1} - y_n) &= 0 \\
\sum_{n=1}^{N} w_n\theta_0 + \sum_{n=1}^{N} w_n\theta_1 x_{n,1} - \sum_{n=1}^{N} w_n y_n &= 0 \\
\theta_0 \sum_{n=1}^{N} w_n + \theta_1 \sum_{n=1}^{N} w_n x_{n,1} - \sum_{n=1}^{N} w_n y_n &= 0
\end{aligned}$$

Thus, by dividing the form by $\sum_{n=1}^{N} w_n$,

$$\begin{aligned}
\theta_0 + \theta_1\overline{x} - \overline{y} = 0 \\
\theta_0 = \overline{y} - \theta_1\overline{x}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J}{\partial \theta_1} &= 0 \\
\sum_{n=1}^{N} 2w_n x_{n,1}(\theta_0 + \theta_1 x_{n,1} - y_n) &= 0 \\
\sum_{n=1}^{N} w_n x_{n,1}(\theta_0 + \theta_1 x_{n,1} - y_n) &= 0 \\
\sum_{n=1}^{N} w_n x_{n,1}\theta_0 + \sum_{n=1}^{N} w_n x_{n,1}\theta_1 x_{n,1} - \sum_{n=1}^{N} w_n x_{n,1} y_n &= 0 \\
\theta_0 \sum_{n=1}^{N} w_n x_{n,1} + \theta_1 \sum_{n=1}^{N} w_n x_{n,1}^2 - \sum_{n=1}^{N} w_n x_{n,1} y_n &= 0
\end{aligned}$$

Thus, by dividing the form by $\sum_{n=1}^{N} w_n$,

$$\begin{aligned} \theta_0\overline{x} + \theta_1\overline{x^2} - \overline{xy} &= 0 \\ \theta_0 &= \frac{\overline{xy} - \theta_1\overline{x^2}}{\overline{x}} \end{aligned}$$

As a result, from the equations derived from the two partial expressions, we obtain:

$$\begin{aligned} \theta_0 &= \overline{y} - \theta_1\overline{x} \\ \theta_0 &= \frac{\overline{xy} - \theta_1\overline{x^2}}{\overline{x}} \end{aligned}$$

Therefore, θ_0 and θ_1 can be found as:

$$\begin{aligned} \theta_1 &= \frac{\bar{x}\bar{y} - \overline{xy}}{\overline{x}^2 - \overline{x^2}} \\ \theta_0 &= \overline{y} - \frac{\bar{x}\bar{y} - \overline{xy}}{\overline{x}^2 - \overline{x^2}}\overline{x} \end{aligned}$$

(c) Problem 3c

Solution:

The sizes of the matrices are: $X : (N \times 2)$, $y : (N \times 1)$, $\theta : (2 \times 1)$. Therefore, in order for the result J to be found correctly, the dimension for W should be $(N \times N)$.

$$\sum_{n=1}^{N} (\theta_0 + \theta_1 x_{n,1} - y_n) = X\theta - y$$

W is a $(N \times N)$ with all zeros except for its diagonal. Its N diagonal numbers are $w_1, w_2, w_3, ..., w_n$.

$$\begin{aligned} J(\theta) &= (X\theta - y)^T W (X\theta - y) \\ &= (X\theta - y)^T \sqrt{W}\sqrt{W}(X\theta - y) \\ &= (X\theta - y)^T \sqrt{W}^T \sqrt{W}(X\theta - y) \\ &= (\sqrt{W}(X\theta - y))^T \sqrt{W}(X\theta - y) \end{aligned}$$

Since $\sqrt{W}(X\theta - y)$ is a vector form with the nth row as $\sqrt{w_n}(\theta_0 + \theta_1 x_{n,1} y_n)$

$J(\theta) = (\sqrt{w_n}(\theta_0 + \theta_1 x_{n,1} y_n))^2$

$J(\theta) = w_n(\theta_0 + \theta_1 x_{n,1} y_n)$, $J(\theta) = \sum_{n=1}^{N} w_n(\theta_0 + \theta_1 x_{n,1} - y_n)$

So $J(\theta) = \sum_{n=1}^{N} w_n(\theta_0 + \theta_1 x_{n,1} - y_n)$, and thus $J(\theta)$ can be written in the form above.

4 Problem 4

(a) Problem 4a

Solution:

According to the plots generated, for the linear regression model, the training data has followed a negative slope. Nevertheless, the plots, especially the one for the test data, only follow a pretty weak linear relation. Therefore, linear regression is not efficient in predicting the data.

(b)

(c)

(d) Problem 4d

Solution:

For a series of eta values:

```
eta = 0.0001
number of iterations: 10000
cost: 4.0863970367957645
coef [ 2.27044798 -2.46064834]

eta = 0.0010
number of iterations: 7020
cost: 3.9125764057919463
coef [ 2.4464068 -2.816353 ]

eta = 0.0100
number of iterations: 764
cost: 3.912576405791487
coef: [ 2.44640703 -2.81635346]

eta = 0.0407
number of iterations: 10000
cost: 2.710916520014198e+39
coef [-9.40470931e+18 -4.65229095e+18]
```

Based on the results, it can be seen that if the learning rate is too small, the coefficients will be close to the closed-form value, but it will take a long time or a lot of steps for the algorithm to converge since the step size is small. If the learning rate is too large, the coefficients and costs are much larger than the previous three cases, so the algorithm no longer converge for this learning rate.

(e) Problem 4e

Solution:

Closed Form
cost: 3.9125764057914645
coef [2.44640709 -2.81635359]

The coefficients and costs of the closed-form are close to the ones found using GD. The time taken for the algorithms to converge is much shorter for this closed-form one compared to the ones by GD. This is due to the fact that this method does not need loops.

(f) Problem 4f

Solution:

Learning rate
number of iterations: 1678
cost: 3.912576405792011
coef [2.44640678 -2.81635296]

So it takes 1678 iterations for the algorithm to converge.

(g)

(h) Problem 4h

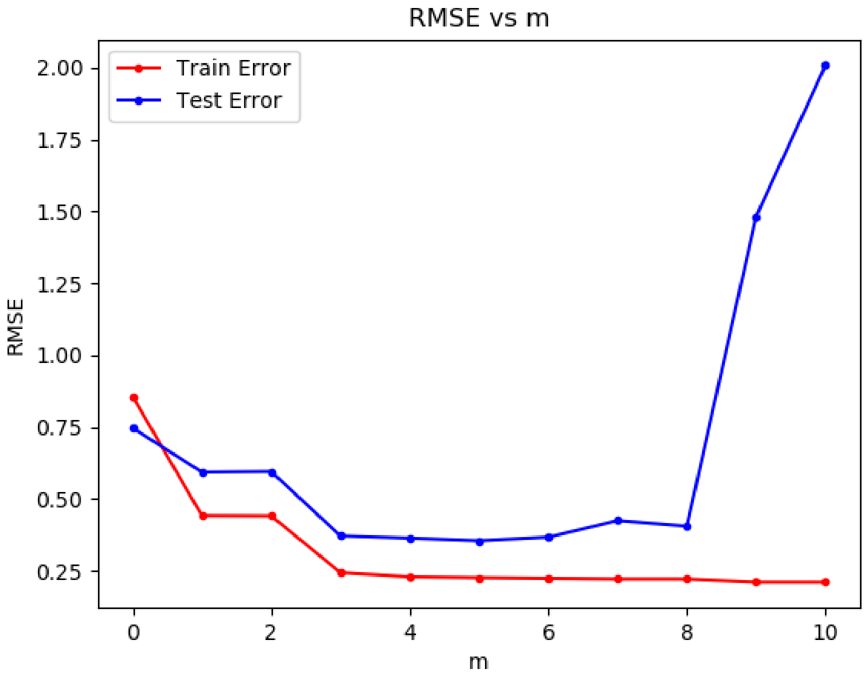
Solution:

The reason that we might prefer RMSE as a metric over J(θ) is that RMSE can eliminate the effects of the number of training instances, so it can hep us better examine the results between two trials with different training set sizes.
Also, the square root used in RMSE, by offering the standard deviation

measures of the data, can make the comparisons easier to interpret for us since the effects of squares have been eliminated.

(i) Problem 4f

Solution:



According to the plot, when the polynomial degree is 4 or 5, the data have the best fits, as both the train error and the test error are at their lowest levels. If m is too small, the train error will be too big, which is the effect of underfitting. And if m is too big, then even though the training error decreases, the test errors will be too bog due to the effect of overfitting.