CM146, Fall 2017
Problem Set 2: Perceptron and Regression
Due Nov 6, 2017

Author: Shivraj Gill

1 Problem 1

(a) Problem 1a

Solution: Solution to problem 1a

Table 1: AND

x_1	x_2	y
+1	+1	+1
-1	+1	-1
+1	-1	-1
-1	-1	-1

For the solution to be valid, $y = sign(\theta^T x)$ where $X = (1, x_1, x_2)$.
We see that one solution would be $\theta = (-1, 1, 1)$ since that would give us the desired y value for each $x_1 and x_2$ in the table.
This solution is not unique since $\theta = (-1, .6, .6)$ works as well.

(b) Problem 1b

Solution: Solution to problem 1b

Table 2: XOR

x_1	x_2	y
+1	+1	-1
-1	+1	+1
+1	-1	+1
-1	-1	-1

No perceptron exists that satisfies this function because the XOR function is not linearly separable.

2 Problem 2

(a) Problem 2a

Solution: Solution to problem 2a

$$\begin{aligned}\frac{\partial J}{\partial \theta_j} &= -\frac{\partial(\sum_{n=1}^{N}[y_n log h_\theta(x_n) + (1-y_n)log(1-h_\theta(x_n))])}{\partial \theta_j} \\ &= -\sum_{n=1}^{N}[y_n \frac{\partial(log h_\theta(x_n))}{\partial \theta_j} + (1-y_n)\frac{\partial(log(1-h_\theta(x_n)))}{\partial \theta_j}]\end{aligned} \tag{1}$$

To compute the partial derivatives above we need to know $\frac{\partial h_\theta(x)}{\partial \theta_j}$. Since $h(\theta) = \sigma(\theta^T x)$ and using the fact that $\sigma'(x) = \sigma(x)(1-\sigma(x))$, we can calculate $\frac{\partial h_\theta(x)}{\partial \theta_j}$

$$\begin{aligned}\frac{\partial h_\theta(x)}{\partial \theta_j} &= \frac{\partial \sigma(\theta^T x)}{\partial \theta_j} \\ &= \frac{\partial \sigma(\theta^T x)}{\partial \theta^T x}\frac{\theta^T x}{\theta_j} \\ &= \sigma(\theta^T x)(1-\sigma(\theta^T x))x_j \\ &= h_\theta(x)(1-h_\theta(x))x_j\end{aligned} \tag{2}$$

Now we can use (2) to compute $\frac{\partial log h_\theta(x_n)}{\partial \theta_j}$ and $\frac{\partial log(1-h_\theta(x_n))}{\partial \theta_j}$

$$\begin{aligned}\frac{\partial log h_\theta(x_n)}{\partial \theta_j} &= \frac{\partial log h_\theta(x_n)}{\partial h_\theta}\frac{\partial h_\theta(x_n)}{\partial \theta_j} \\ &= \frac{1}{h_\theta(x_n)}h_\theta(x_n)(1-h_\theta(x_n))x_j \\ &= (1-h_\theta(x_n))x_j\end{aligned} \tag{3}$$

Likewise,

$$\frac{\partial log(1-h_\theta(x_n))}{\partial \theta_j} = -h_\theta(x_n)x_j \tag{4}$$

Using the partials derived in (3) and (4) for (1), we get

$$\begin{aligned}\frac{\partial J}{\partial \theta_j} &= -\sum_{n=1}^{N}[y_n(1-h_\theta(x_n))x_j - (1-y_n)h_\theta(x_n)x_j] \\ &= \sum_{n=1}^{N}(h_\theta(x_n)-y_n)x_{n,j}\end{aligned}$$

(b) Problem 2b

Solution: Solution to problem 2b

For every j, k

$$\begin{aligned}\frac{\partial^2 J}{\partial \theta_j \theta_k} &= \frac{\partial}{\partial \theta_j}(\frac{\partial J}{\partial \theta_k}) \\ &= \frac{\partial}{\partial \theta_j}(\sum_{n=1}^{N}(h_\theta(x_n)-y_n)x_{n,k}) \\ &= \frac{\sum_{n=1}^{N}\partial(h_\theta(x_n))x_{n,k}}{\partial\theta_j} = \\ &= \sum_{n=1}^{N}h_\theta(x_n)(1-h_\theta(x_n))x_{n,j}x_{n,k}From(2) \\ &= \sum_{n=1}^{N}h_\theta(x_n)(1-h_\theta(x_n))(x_n x_n^T)_{j,k} \\ &= H_{j,k}\end{aligned} \tag{5}$$

And since this is true for all j,k we can say H = $\boldsymbol{\sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))x_n x_n^T}$

(c) Problem 2c

Solution: Solution to problem 2c

From part b, we know that H = $\boldsymbol{\sum_{n=1}^{N} h_\theta(x_n)(1-h_\theta(x_n))x_n x_n^T}$. H must be PSD for it to be convex. A function F is PSD if for all real vectors z,

$$\boldsymbol{z^T F z \geq 0} \tag{6}$$

Now let's check if H is PSD,

$$\begin{aligned}
\boldsymbol{z^T H z} &\boldsymbol{= z^T(\sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))x_n x_n^T)z} \\
&\boldsymbol{= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))z^T x_n x_n^T z_n} \\
&\boldsymbol{= \sum_{n=1}^{N} h_\theta(x_n)(1 - h_\theta(x_n))(x_n^T z_n)^2} \\
&\boldsymbol{\geq 0}
\end{aligned}$$

Since $\boldsymbol{0 \leq h_\theta(x_n) \leq 1}$ and $\boldsymbol{0 \leq (1 - h_\theta(x_n) \leq 1}$, and $\boldsymbol{(x_n^T z_n)^2 \geq 0.}$

Thus, H is convex since it PSD.

3 Problem 3

(a) Problem 3a

Solution: Solution to problem 3a

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

(b) Problem 3a

Solution: Solution to problem 3a

$$\begin{aligned}&\frac{J}{\theta_0} = 0 \\ \Rightarrow &\sum_{n=1}^{N} 2w_n(\theta_0 + \theta_1 x_{n,1} - y_n) = 0 \\ \Rightarrow &(\sum_{n=1}^{N} w_n)\theta_0 + (\sum_{n=1}^{N} w_n x_{n,1})\theta_1 = \sum_{n=1}^{N} w_n y_n \end{aligned} \tag{7}$$

$$\begin{aligned}&\frac{J}{\theta_1} = 0 \\ \Rightarrow &\sum_{n=1}^{N} 2w_n x_{n,1}(\theta_0 + \theta_1 x_{n,1} - y_n) = 0 \\ \Rightarrow &(\sum_{n=1}^{N} w_n x_{n,1})\theta_0 + (\sum_{n=1}^{N} w_n x_{n,1}^2)\theta_1 = \sum_{n=1}^{N} w_n y_n x_{n,1} \end{aligned} \tag{8}$$

We divide (7) and (8) by $\sum_{n=1}^{N} w_n$ since $w_n > 0$

Equation (7) becomes

$$\theta_0 + \frac{(\sum_{n=1}^{N} w_n x_{n,1})\theta_1}{\sum_{n=1}^{N} w_n} = \frac{\sum_{n=1}^{N} w_n y_n}{\sum_{n=1}^{N} w_n} \tag{9}$$

Equation (8) becomes

$$\frac{(\sum_{n=1}^{N} w_n x_{n,1})\theta_0}{\sum_{n=1}^{N} w_n} + \frac{(\sum_{n=1}^{N} w_n x_{n,1}^2)\theta_1}{\sum_{n=1}^{N} w_n} = \frac{\sum_{n=1}^{N} w_n y_n x_{n,1}}{\sum_{n=1}^{N} w_n} \tag{10}$$

Because of the $\sum_{n=1}^{N} w_n$ term in the denominator, we can define weighted averages as follows:

$$\begin{aligned}
\bar{x} &= \frac{(\sum_{n=1}^{N} w_n x_{n,1})\theta_0}{\sum_{n=1}^{N} w_n} \\
\overline{x^2} &= \frac{(\sum_{n=1}^{N} w_n x_{n,1}^2)\theta_1}{\sum_{n=1}^{N} w_n} \\
\bar{y} &= \frac{\sum_{n=1}^{N} w_n y_n}{\sum_{n=1}^{N} w_n} \\
\overline{xy} &= \frac{\sum_{n=1}^{N} w_n y_n x_{n,1}}{\sum_{n=1}^{N} w_n}
\end{aligned}$$

Substituting these terms in equations (9) and (10) gives us the following:

$$\theta_0 + \bar{x}\theta_1 = \bar{y} \tag{11}$$

$$\bar{x}\theta_0 + \overline{x^2}\theta_1 = \overline{xy} \tag{12}$$

Plugging in θ_0 from equation (11) into (12) gives us

$$\begin{aligned}
&\bar{x}(\bar{y} - \bar{x}\theta_1) + \overline{x^2}\theta_1 = \overline{xy} \\
\Rightarrow &\hat{\theta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}
\end{aligned}$$

Plugging $\hat{\theta}_1$ into (11) and solving for θ_0 gives us

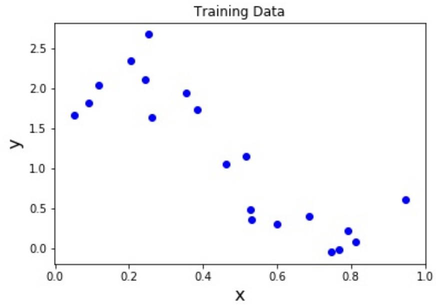
$$\begin{aligned}
&\theta_0 + \bar{x}(\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}) = \bar{y} \\
&(\overline{x^2} - \bar{x}^2)\theta_0 + \bar{x}(\overline{xy} - \bar{x}^2\bar{y}) = \bar{y}(\overline{x^2} - \bar{x}^2) \\
&\theta_0 = \bar{y} - \bar{x}\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \\
\Rightarrow &\theta_0 = \bar{y} - \bar{x}\hat{\theta}_1
\end{aligned}$$

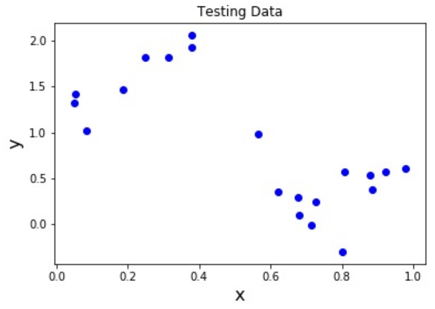
The solution for weighted regression is very similar to least squares just that we use weighted averages instead of unweighted averages.

4 Problem 4

(a) Problem 4a

Solution: Solution to problem 4a





We see that the data is not linearly separable, so fitting a linear model should give us poor results. Instead, a nonlinear function such as polynomial function should give us better results since the data looks like it follows a sine curve.

(b) Problem 4b

Solution: Solution to problem 4b Updated the function to support linear regression

(c) Problem 4c

Solution: Solution to problem 4c Updated the predict function

(d) Problem 4d

Solution: Solution to problem 4d

Table 3: Convergence for different values of Step-size

α	iterations	Final Value
.0407	384	3.91257640579
.01	1467	3.91257640579
.001	10000	3.91257640947
.0001	10000	5.49356558874

We see that for smaller values of α, gradient descent converges slower. When we use $\alpha = .0001$, our objective function has a relatively large value compared to the values of the objective function with smaller α. For $alpha = .0407$, gradient descent converges in 1/5th of the steps needed for $alpha = .01$. For $alpha = .001$, gradient descent reaches the max number of iterations, but basically had converged to the minimum of the objective function which is around 3.91257640579.

(e) Problem 4e

Solution: Solution to problem 4e

For the closed form solution, the objective function equals 3.91257640579 which is the same value as gradient descent when it converged for α $= .0407$ and $\alpha = .01$. The closed form algorithm runs a lot quicker than gradient descent since it requires less matrix operations because it is only needed to run once instead of over multiple iterations as in gradient descent.

(f) Problem 4f

Solution: Solution to problem 4f The algorithm reaches the max number of iterations to converge (10000 iterations) which makes sense since our step size is getting smaller and smaller each iteration so it converges slower and slower through each run of gradient descent.

(g) Problem 4g

Solution: Solution to problem 4g Updated the function to support polynomial regression

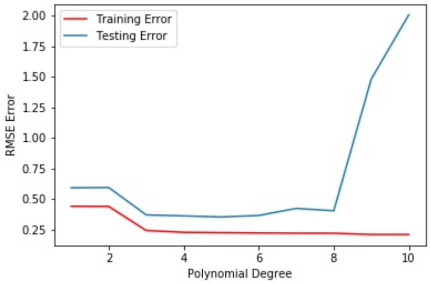
(h) Problem 4h

Solution: Solution to problem 4h

Because our cost function is the sum of the difference squared, the value of the cost function is an order higher than our output and hence could be a poor way to evaluate our model if our predicted and actual values are large. RMSE represents the standard deviation between the actual and predicted value, and hence is the same order as our outputs. This gives us a better picture of how close our predictions are to the actual values.

(i) Problem 4i

Solution: Solution to problem 4i



A polynomial of degree 5 best fits our data because we achieve the lowest RMSE error for the test data with this polynomial. We can see underfitting for polynomials of degree 1 and 2, since both the training error and test error and relatively high. Around m = 9, we see overfitting since the test RMSE is significantly large and the training RMSE is very small.