CSCI 5521 - INTRO TO MACHINE LEARNING

TANVEER SINGH VIRDI

ASSIGNMENT 0

SEPTEMBER 9, 2019

PROBLEM 1:

SOLUTION:

1. Feature Matrix : **X** $\in \mathbb{R}^{n \times m}$ (n >= m) ; Response vector : **y** $\in \mathbb{R}^{n \times 1}$; Vector of linear coefficients:

   w $\in \mathbb{R}^{m \times 1}$

   Standard linear regression can be formulated by solving the least square problem: $\min_{w} \|Xw - y\|^2$

   Euclidean norm is defined as : $\|x\| = \sqrt{(x^T x)}$

   Hence, $\|Xw - y\|^2 = \sqrt{((Xw - y)^T (Xw - y))}^2 = ((Xw - y)^T (Xw - y)) = (w^T X^T Xw - 2y^T Xw + y^T y)$

   Taking the gradient of this expression with respect to w we get:

   $$\nabla_w (w^T X^T Xw - 2y^T Xw + y^T y) = \nabla_w w^T X^T Xw - \nabla_w 2y^T Xw + \nabla_w y^T y$$

   Setting this gradient to zero vector, we get:

   $$\nabla_w w^T (X^T X) w - \nabla_w 2(y^T X) w + \nabla_w y^T y = 0$$

   Since, we know $\nabla_w w^T X w = 2Xw$ (if X symmetric) and $\nabla_w y^T w = y$, the above equation reduces to:

   $$\Rightarrow \quad 2X^T Xw - 2X^T y = 0 \qquad \Rightarrow X^T Xw = X^T y$$

   Multiplying both sides with $(X^T X)^{-1}$

   $$\boxed{w = (X^T X)^{-1} X^T y}$$

2. Feature Matrix : **X** $\in \mathbb{R}^{n \times m}$ (n >= m) ; Response vector : **y** $\in \mathbb{R}^{n \times 1}$; Vector of linear coefficients:

   w $\in \mathbb{R}^{m \times 1}$

   Objective function of the ridge regression is : $\min_{w} (\|Xw - y\|^2 + \lambda \|w\|^2) \qquad \lambda \geq 0$

   $\|Xw - y\|^2 + \lambda \|w\|^2 = ((Xw - y)^T (Xw - y)) + \lambda w^T w = (w^T X^T Xw - 2y^T Xw + y^T y) + (\lambda w^T w)$

   Taking the gradient of this expression with respect w we get:

   $$\nabla_w (w^T X^T Xw - 2y^T Xw + y^T y + \lambda w^T w) = \nabla_w w^T X^T Xw - \nabla_w 2y^T Xw + \nabla_w y^T y + \nabla_w \lambda w^T w$$

   Setting this gradient to zero vector, we get:

$$\nabla_w w^T X^T X w - \nabla_w 2y^T X w + \nabla_w y^T y + \nabla_w \lambda w^T w = 0$$

Since, we know $\nabla_w w^T X w = 2Xw$ (if X symmetric) and $\nabla_w y^T w = y$ and $\nabla_w w^T w = 2w$, we get:

$$\Rightarrow \quad 2X^T X w - 2X^T y + 2\lambda w = 0 \quad \Rightarrow X^T X w + \lambda w = X^T y$$

$$\Rightarrow \quad (X^T X + \lambda I) w = X^T y$$

Multiplying both sides with $(X^T X + \lambda I)^{-1}$

$$\boxed{w = (X^T X + \lambda I)^{-1} X^T y}$$

PROBLEM 2:

SOLUTION:

1. P(H) = p , P(T) = 1-p

Probability of observing the sequence H,H,T,T,H in 5 tosses P(E)= P(H)*P(H)*P(T)*P(T)*P(H)

= p*p*(1-p)*(1-p)*p

= $p^3(1-p)^2$

Natural log of this probability is = log(P(E)) = log($p^3.(1-p)^2$) = 3 log (p) + 2 log (1-p)

2.

(a) Probability of choosing the fair coin P(f) = 1/2 = 0.5

Probability of heads for fair coin P(H) = p = 1/2 = 0.5

Joint probability that a coin is fair(p = 1/2) and the outcome is H,H,T,T,H = $P(f \cap E)$

= P(f) * P(E)

= 0.5 * $p^3(1-p)^2$

= $0.5 * 0.5^3 * 0.5^2 = 0.5^6$ = 1/64 = 0.015625

(b) Probability of choosing the bias coin P(b) = 1/2 = 0.5

Probability of heads for bias coin P(H) = p = 2/3 = 0.67

Joint probability that a coin is bias(p = 2/3) and the outcome is H,H,T,T,H = $P(b \cap E)$

= P(b) * P(E)

= 0.5 * $p^3(1-p)^2$

= 0.5 * 0.67^3 * 0.33^2 = 0.016376

(c) Let the probability bias of observing heads P(H) = p

Probability of observing H,H,T,T,H is = p*p*(1-p)*(1-p)*p

= $p^3(1-p)^2$

To maximize the probability of observing H,H,T,T,H, we will maximize the log of the function $p^3(1-p)^2$ and set it to zero: $\frac{d}{dp}(\log(p^3.(1-p)^2)) = 0$

$$\Rightarrow \frac{d}{dp}(3\log p + 2\log\,(1-p)) = 0$$

$$\Rightarrow \frac{3}{p} - \frac{2}{(1-p)} = 0$$

$$\Rightarrow (5p - 3)/((p - 1)p) = 0 \Rightarrow p = 3/5$$

We differentiate the function again, $\frac{d}{dp}(\frac{3}{p} - \frac{2}{(1-p)}) = -\frac{3}{p^2} - \frac{2}{(p-1)^2}$

Substituting p = 3/5 in this expression, we get $\Rightarrow -\frac{3}{(\frac{3}{5})^2} - \frac{2}{(\frac{3}{5}-1)^2} = -20.8333$

Since the value of the expression on substituting p = 3/5 is negative, hence we can say that the probability of observing H,H,T,T,H will be maximum when p = 3/5. Thus the probability of observing H,H,T,T,H will be maximized when p = 3/5.

Corresponding probability of observing H,H,T,T,H when p = 3/5 is : $p^3(1-p)^2$

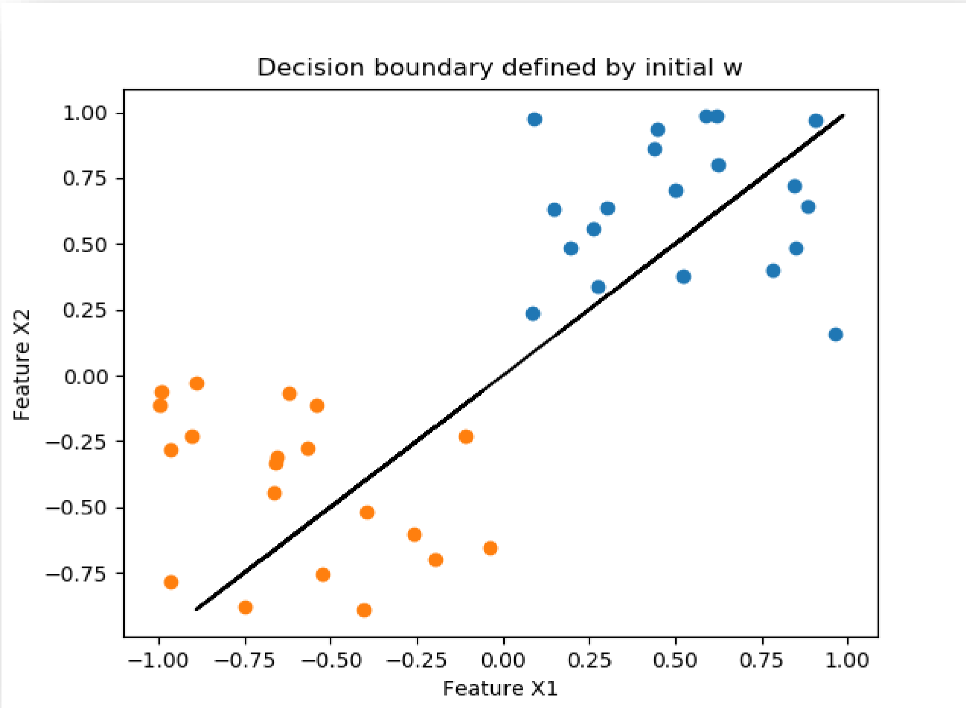
= $(3/5)^3*(2/5)^2$

= 0.03456

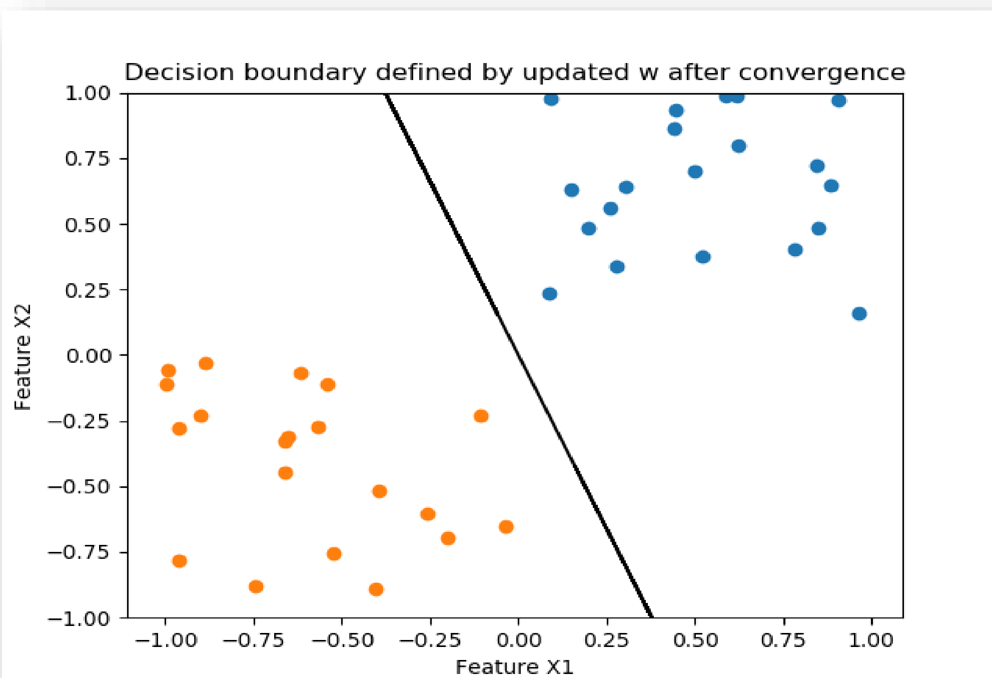
PROBLEM 3:

SOLUTION:

1. Plots :



Weight vector w converges after 3 iterations of the perceptron algorithm.



2. Initial weight vector w does not converge with perceptron algorithm because the data in *data2.mat* is not linearly separable.