Solution 1.1

The objective is

$$\underset{w}{\text{minimize}} \|Xw - y\|^2 \tag{1}$$

where, $X \in \mathbb{R}^{n \times m}$, $(n \geq m)$ represents the feature matrix, $y \in \mathbb{R}^{n \times 1}$ represents the response vector and $w \in \mathbb{R}^{m \times 1}$ is the vector variable of the linear coefficients.

The cost function is

$$J(w) = \|Xw - y\|^2 = (Xw - y)^T(Xw - y) \tag{2}$$

By definition,

$$\begin{aligned} J(w) &= \sum_{i=1}^{n} \left[\sum_{j=1}^{m} (x_{ij} w_j) - y_i \right]^2 \\ \implies \frac{\partial J}{\partial w_k} &= \sum_{i=1}^{n} 2 \left[\sum_{j=1}^{m} (x_{ij} w_j) - y_i \right] x_{ik} \end{aligned} \tag{3}$$

Equating (3) to 0 for each $w_k, k = 1, 2, \ldots, m$, and in vector form, we get -

$$\begin{aligned} & X^T(Xw^* - y) = 0 \\ \implies & X^T X w^* = X^T y \\ \implies & w^* = (X^T X)^{-1} X^T y \end{aligned} \tag{4}$$

Solution 1.2

In case of Ridge Regression, the objective is

$$\underset{w}{\text{minimize}} \|Xw - y\|^2 + \lambda \|w\|^2 \tag{5}$$

where, $\lambda > 0$

The cost function is

$$J(w) = \|Xw - y\|^2 + \lambda \|w\|^2 = (Xw - y)^T(Xw - y) + \lambda w^T w \tag{6}$$

By definition,

$$\begin{aligned} J(w) &= \sum_{i=1}^{n} \left[\sum_{j=1}^{m} (x_{ij} w_j) - y_i \right]^2 + \sum_{j=1}^{m} \lambda w_j^2 \\ \implies \frac{\partial J}{\partial w_k} &= \sum_{i=1}^{n} 2 \left[\sum_{j=1}^{m} (x_{ij} w_j) - y_i \right] x_{ik} + 2\lambda w_k \end{aligned} \tag{7}$$

Equating (7) to 0 for each $w_k, k = 1, 2, \ldots, m$, and in vector form, we get -

$$\begin{aligned} & X^T(Xw^* - y) + \lambda w^* = 0 \\ \implies & (X^T X + \lambda I) w^* = X^T y \\ \implies & w^* = (X^T X + \lambda I)^{-1} X^T y \end{aligned} \tag{8}$$

Solution 2.1

$Pr(H) = p$ and $Pr(T) = 1 - p$

The probability of observing the sequence H, H, T, T, H in five tosses is, $P_{seq} = p \times p \times (1-p) \times (1-p) \times p = p^3(1-p)^2$. Therefore,

$$ln(P_{seq}) = 3ln(p) + 2ln(1 - p) \tag{9}$$

Solution 2.2

2.2(a)

Probability of choosing the fair coin($p = \frac{1}{2}$) is $\frac{1}{2}$ and in this case observing the sequence H, H, T, T, H in five tosses is $\frac{1}{2} \times p^3(1-p)^2 = \frac{1}{2} \times (\frac{1}{2})^3(1-(\frac{1}{2}))^2 = 0.015625$

2.2(b)

Probability of choosing the biased coin($p = \frac{2}{3}$) is $\frac{1}{2}$ and in this case observing the sequence H, H, T, T, H in five tosses is $\frac{1}{2} \times p^3(1-p)^2 = \frac{1}{2} \times (\frac{2}{3})^3(1-(\frac{2}{3}))^2 = 0.016461$

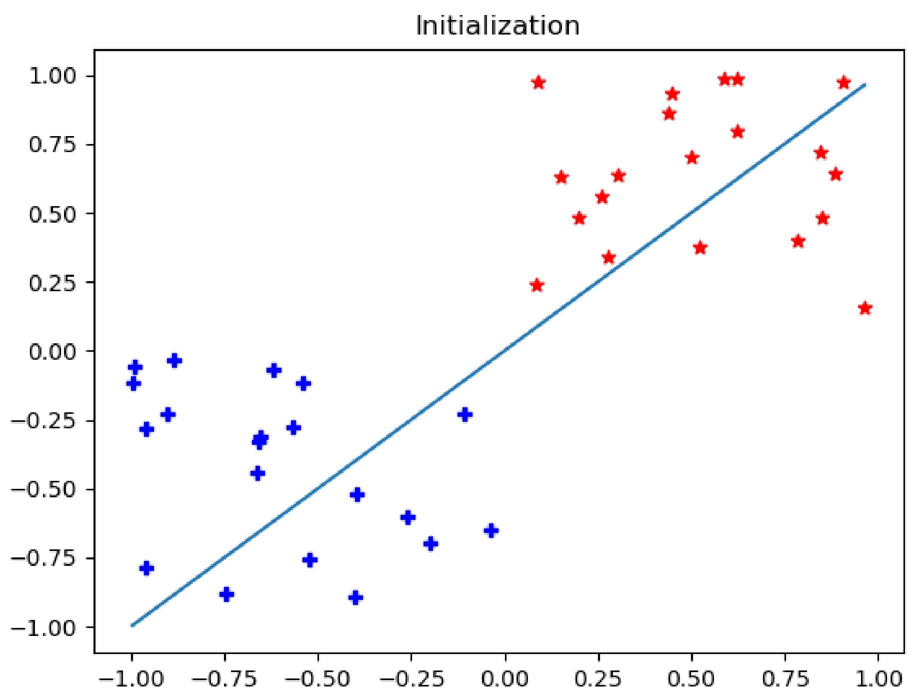
Solution 2.3

To find the bias p^* to maximize (9),

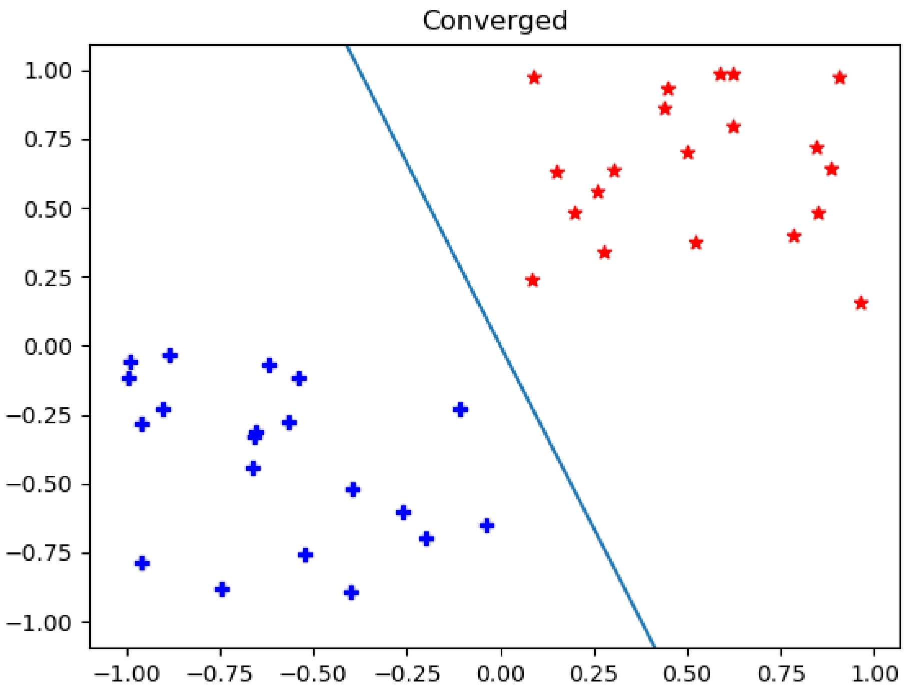
$$
\begin{aligned}
\frac{d(ln(P_{seq}))}{dp} &= 0 \\
\implies \frac{3}{p^*} - \frac{2}{1-p^*} &= 0 \\
\implies p^* &= \frac{3}{5}
\end{aligned}
\tag{10}
$$

and the corresponding probability is ${p^*}^3(1-p^*)^2 = 0.03456$

Solution 3.1



It takes 1 iteration to converge and the corresponding figure is shown below:



Solution 3.2

In this case, as the classes are not linearly separable, perceptron algorithm will not converge. The perceptron algorithm iterates till all the points are classified correctly by a linear decision boundary. But in case of linearly non-separable classes, there is no linear decision boundary which separates all the points correctly. After using a 'soft' linear classifier to tolerate error, we get the following plot: