**Problem 1.1**

$$\begin{aligned}||Xw-y||^2 &= (Xw-y)^T(Xw-y)\\ &= (w^TX^T-y^T)(Xw-y)\\ &= w^TX^TXw - y^TwX \ - w^TX^Ty + y^Ty\end{aligned}$$

Let $f = ||Xw-y||^2$. Then, $\frac{\partial f}{\partial w} = 2wX^TX - 2X^Ty$

Set $\frac{\partial f}{\partial w} = 0$: $2wX^TX - 2X^Ty = 0$

$$\begin{aligned}&\Rightarrow 2X^T(wX-y) = 0\\ &\Rightarrow X^TXw - x^Ty = 0\\ &\Rightarrow X^TXw = x^Ty\\ &\Rightarrow w = \tfrac{x^Ty}{X^TX}\end{aligned}$$

**Problem 1.2**

$$\begin{aligned}||Xw-y||^2 + \lambda||w||^2 &= (Xw-y)^T(Xw-y) + \lambda w^Tw\\ &= (w^TX^T-y^T)(Xw-y) + \lambda w^Tw\\ &= w^TX^TXw - w^TX^Ty - y^TXw + y^Ty + \lambda w^Tw\end{aligned}$$

Let $f = ||Xw-y||^2 + \lambda||w||^2$. Then, $\frac{\partial f}{\partial w} = 2wX^TX - 2X^Ty + 2\lambda w$

Set $\frac{\partial f}{\partial w} = 0$: $2wX^TX - 2X^Ty + 2\lambda w = 0$

$$\begin{aligned}&\Rightarrow wX^TX - X^Ty + \lambda w = 0\\ &\Rightarrow wX^TX + \lambda w = X^Ty\\ &\Rightarrow w(X^TX + \lambda I) = X^Ty\\ &\Rightarrow w = \tfrac{X^Ty}{(X^TX+\lambda I)}\end{aligned}$$

**Problem 2.1**

$Pr(H) = p,\ Pr(T) = 1-p$ (given)

Probability of observing sequence (H,H,T,T,H) in five tosses:

$$\begin{aligned}Pr(H,H,T,T,H) &= p*p*(1-p)*(1-p)*p\\ &= p^3(1-p)^2 \qquad \text{(factored)}\\ &= p^5 - 2p^4 + p^3 \qquad \text{(expanded)}\end{aligned}$$

Natural logarithm of probability above:

$$\begin{aligned}\ln[Pr(H,H,T,T,H)] &= \ln[p^3(1-p)^2] \qquad \text{(using factored form)}\\ &= 3\ln p + 2\ln(1-p)\end{aligned}$$

## Problem 2.2

**a.** Probability that chosen coin was fair coin: $Pr(fair) = 1/2$

$Pr(H) = 1/2,\ Pr(T) = 1/2$ (for fair coin)

Joint probability that outcome was fair coin with sequence (H,H,T,T,H):

$Pr(fair) \bullet Pr(H) \bullet Pr(H) \bullet Pr(T) \bullet Pr(T) \bullet Pr(H) = (1/2)^6$

$= 1/64$

**b.** Probability that chosen coin was biased coin: $Pr(biased) = 1/2$

$Pr(H) = 2/3,\ Pr(T) = 1/3$ (for biased coin)

Joint probability that outcome was biased coin with sequence (H,H,T,T,H):

$Pr(biased) \bullet Pr(H) \bullet Pr(H) \bullet Pr(T) \bullet Pr(T) \bullet Pr(H) = (1/2) \bullet (2/3)^3 \bullet (1/3)^2$

$= 4/243$

## Problem 2.3

Maximize $\ln[Pr(H,H,T,T,H)] = 3\ln p + 2\ln(1-p)$, i.e. set the derivative of the function with respect to p to 0 to find the critical point:

$\frac{d}{dp}(3\ln p + 2\ln(1-p)) = 0 \Rightarrow \frac{d}{dp}(3\ln p) + \frac{d}{dp}[2\ln(1-p)] = 0$

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

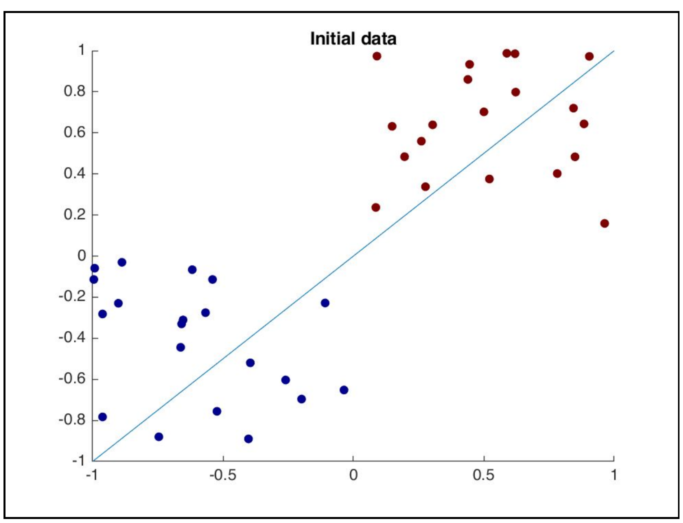
$\Rightarrow \frac{3}{p} = \frac{2}{1-p}$

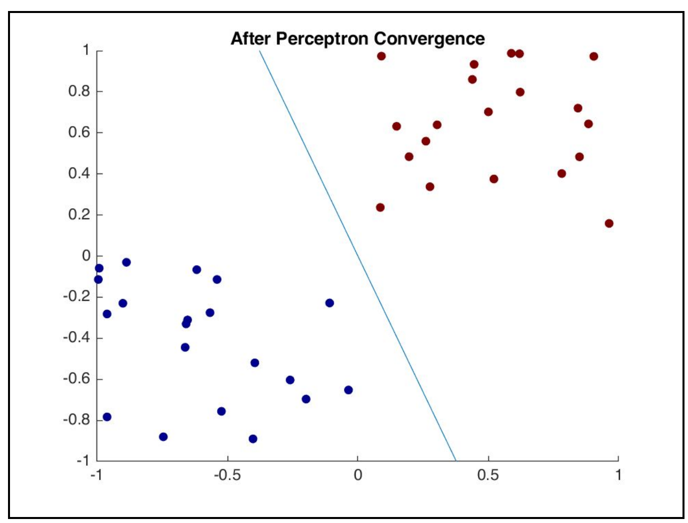
$\Rightarrow 3 - 3p = 2p$

$\Rightarrow 3 = 5p$

$\Rightarrow p = 3/5$

## Problem 3.1




Using the data in 'data1.mat', the perceptron algorithm takes 3 iterations to converge.

**Problem 3.2**

With the linearly non-separable data in 'data2.mat' and w = [1;-1], the perceptron algorithm **cannot** converge. This is because the perceptron is a *linear classifier*, so it only works on linearly separable data. Specifically, the positive class must be able to be separated from the negative class by a hyperplane. Learning will only fail with this algorithm in the linearly non-separable case.

Upon using a new algorithm (a "soft" linear classifier) to slightly tolerate errors, the following linear classifier is produced: