HW 0

Ry Wiese

February 2019

1 Problem 1

1.

$$\frac{\partial}{\partial w}(||Xw - y||^2) =$$

$$\frac{\partial}{\partial w}((Xw - y) \cdot (Xw - y)) =$$

$$\frac{\partial}{\partial w}((Xw - y)^T(Xw - y)) =$$

$$\frac{\partial}{\partial w}(((Xw)^T - y^T)(Xw - y)) =$$

$$\frac{\partial}{\partial w}(((w^TX^T) - y^T)(Xw - y)) =$$

$$\frac{\partial}{\partial w}(w^TX^TXw - w^TX^Ty - y^TXw - y^Ty) =$$

$$\frac{\partial}{\partial w}(w^TX^TXw - w^TX^Ty - y^TXw).$$

Since y^TXw is a scalar value, $y^TXw = (y^TXw)^T = w^TXy$, and thus

$$\frac{\partial}{\partial w}(||Xw - y||^2) = \frac{\partial}{\partial w}(w^TX^TXw - 2(w^TX^Ty)) =$$

$$\frac{\partial}{\partial w}(w^TX^TXw) - 2\frac{\partial}{\partial w}(w^TX^Ty).$$

Looking at these components separately,

$$\frac{\partial}{\partial w}(w^TX^TXw) = \frac{\partial}{\partial w}[(w^TX^T)(Xw)] =$$

$$[\frac{\partial}{\partial w}(w^TX^T)][Xw] + [w^TX^T][\frac{\partial}{\partial w}(Xw)] =$$

$$[\frac{\partial}{\partial w}(Xw)]^T[Xw] + [w^TX^T][X] = X^TXw + w^TX^TX = 2X^TXw$$

and
$$\frac{\partial}{\partial w}(w^TX^Ty) = [\frac{\partial}{\partial w}(y^TXw)]^T = (y^TX)^T = X^Ty.$$

Putting these together,

$$\frac{\partial}{\partial w}(||Xw - y||^2) = \frac{\partial}{\partial w}(w^TX^TXw) - 2\frac{\partial}{\partial w}(w^TX^Ty) = 2X^TXw - 2X^Ty$$

To calculate the optimal w, let $2X^TXw - 2X^Ty = 0$.

$$2X^TXw = 2X^Ty$$
$$Xw = y$$
$$w = X^{-1}y$$

2.
$$\frac{\partial}{\partial w}(||Xw - y||^2 - \lambda||w||^2) = 2X^TXw - 2X^Ty - \frac{\partial}{\partial w}(\lambda||w||^2) =$$
$$2X^TXw - 2X^Ty - \lambda\frac{\partial}{\partial w}(w^Tw) = 2X^TXw - 2X^Ty - 2\lambda w.$$

Let $2X^TXw - 2X^Ty - 2\lambda w = 0$.

$$X^TXw - \lambda w = X^Ty$$
$$(X^TX - \lambda)w = X^Ty$$
$$w = (X^TX - \lambda)^{-1}X^Ty.$$

2 Problem 2

1. $Pr(H) \cdot Pr(H) \cdot Pr(T) \cdot Pr(T) \cdot Pr(H) = p \cdot p \cdot (1-p) \cdot (1-p) \cdot p = p^3 \cdot (1-p)^2$.
$ln(p^3 \cdot (1-p)^2)$.

2. a) Let E be the event that we get the outcome H,H,T,T,H.
$Pr((E)\&(p = \frac{1}{2})) = Pr(E|p = \frac{1}{2}) \cdot Pr(p = \frac{1}{2}) = (\frac{1}{2})^3 \cdot (1 - \frac{1}{2})^2 \cdot \frac{1}{2} = (\frac{1}{2})^6$.

2. b) $Pr((E)\&(p=\frac{2}{3}))=Pr(E|p=\frac{2}{3})\cdot Pr(p=\frac{2}{3})=(\frac{2}{3})^3\cdot(1-\frac{2}{3})^2\cdot\frac{1}{2}=(\frac{2}{3})^3\cdot(\frac{1}{3})^2\cdot\frac{1}{2}=\frac{2^2}{3^5}$.

3. Let $f(p)=ln(p^3\cdot(1-p)^2)$.
$f(p)=ln(p^3\cdot(1-p)^2)=ln(p^3(p^2-2p+1))=ln(p^5-2p^4+p^3)$.
$f'(p)=\frac{1}{p^5-2p^4+p^3}\cdot(5p^4-8p^3+3p^2)=\frac{p^2(5p^2-8p+3)}{p^3(p^2-2p+1)}=\frac{p^2(5p-3)(p-1)}{p^3(p-1)^2}=\frac{5p-3}{p(p-1)}$.
$(f'(p)=0)\iff(5p-3=0)\iff(p=\frac{3}{5})$.

3 Problem 3

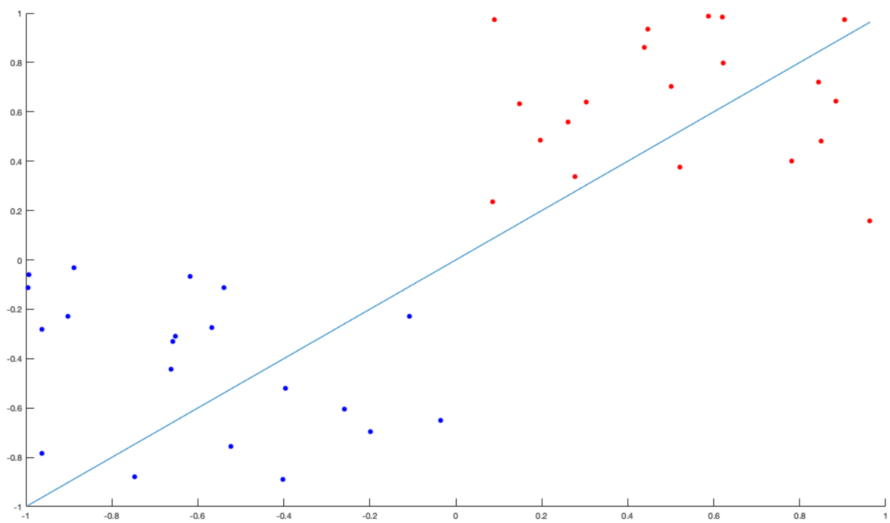

Figure 1: Initialization

The perceptron algorithm cannot converge on the data in data2.mat because the data points are not linearly separable. The algorithm continues to run until $(w\cdot x^t)*r^t\leq 0$ for all training samples. This is not possible for the red point in the bottom left or for the blue point in the top right of Figure 3, and thus the algorithm never terminates.

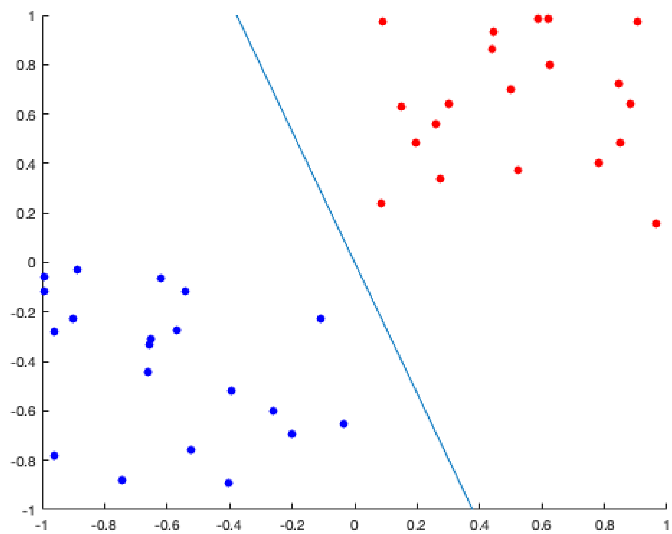

Figure 2: Convergence

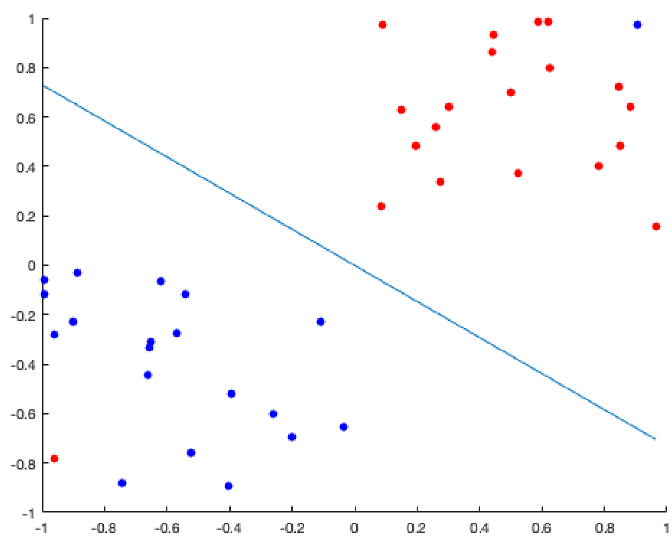

Figure 3: LP algorithm on data2.mat