3. a) $g(\tilde{w}) = \sum_{p=1}^{P} \log(1+e^{-y_p \tilde{x}_p^T \tilde{w}})$ $\quad \tilde{x}_p = \begin{bmatrix} 1 \\ \bar{x}_p \end{bmatrix}$ $\quad \tilde{w} = \begin{bmatrix} b \\ \bar{w} \end{bmatrix}$

$$\nabla g(\tilde{w}) = \sum_{p=1}^{P} \frac{e^{-y_p \tilde{x}_p^T \tilde{w}}}{1+e^{-y_p \tilde{x}_p^T \tilde{w}}} \cdot (-y_p \tilde{x}_p).$$

$$= -\sum_{p=1}^{P} \frac{1}{1+e^{y_p \tilde{x}_p^T \tilde{w}}} (y_p \tilde{x}_p).$$

$$= -\sum_{p=1}^{P} \sigma(-y_p \tilde{x}_p^T \tilde{w}) y_p \tilde{x}_p = \bar{0}_{(N+1)\times 1}$$

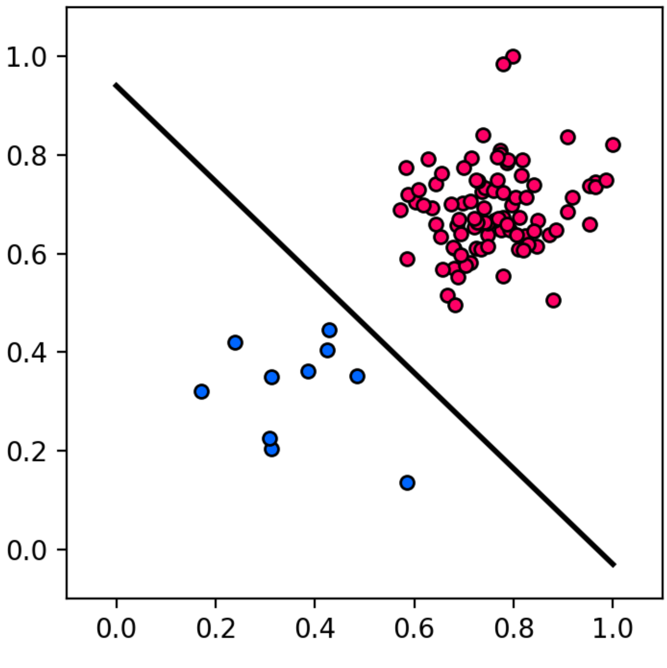
b) in python, we could replace "Σ" by matrix manipulation.

$$Y = [y_1, y_2, \cdots, y_P]_{1\times P} \qquad \tilde{X}' = \begin{bmatrix} 1, & 1, & \cdots, & 1 \\ \bar{x}_1, & \bar{x}_2, & \cdots, & \bar{x}_P \end{bmatrix}_{(N+1)\times P}.$$

$$\nabla g(\tilde{w}) = \cancel{-\sum_{p=1}^{P} \sigma(-y_p \tilde{x}_p^T \tilde{w}) \tilde{x}_p y_p} \; -\sigma(-Y^T \tilde{X}'^T \tilde{w}) \tilde{X}' Y^T = \bar{0}_{(N+1)\times 1}.$$

$$= \tilde{X} \bar{r}$$

$$\bar{r} = Y^T = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_P \end{bmatrix}$$



1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

5(a), Seperating hyperplane:

$$b + \bar{x}^T \bar{w} = 0.$$

$$cb + c\bar{x}^T \bar{w} = 0 \quad \text{doesn't change.}$$

if the point x_p is classified correctly, then $-y_p(b+\bar{x}_p^T\bar{w})<0$.

so, $-y_p(cb+c\bar{x}_p^T\bar{w}) < -y_p(b+\bar{x}_p^T\bar{w})$ $(c>1)$.

So: $\sum_{p=1}^{P} \log(1+e^{-y_p(b+\bar{x}_p^T\bar{w})}) > \sum_{p=1}^{P} \log(1+e^{-y_p(cb+c\bar{x}_p^T\bar{w})})$.

b) if the parameters grows infinitely large, ~~even though~~ the judge condition grows infinitely small. we couldn't find $\nabla g(b,\bar{w})=0$, so. we need to introduce $\lambda\|w\|^2$ to prevent parameter getting so large.

12. For softmax cost function

$$g(b,\bar{w}) = \sum_{p=1}^{P} \log(1+e^{-y_p(b+\bar{x}_p^T\bar{w})})$$

if. $y_p = 1$

$$g(b,\bar{w}) = \sum_{p=1}^{P} \log(1+e^{-(b+\bar{x}_p^T\bar{w})})$$

$y_p = -1$.

$$g(b,\bar{w}) = \sum_{p=1}^{P} \log(1+e^{(b+\bar{x}_p^T\bar{w})})$$

so it is equivalent.

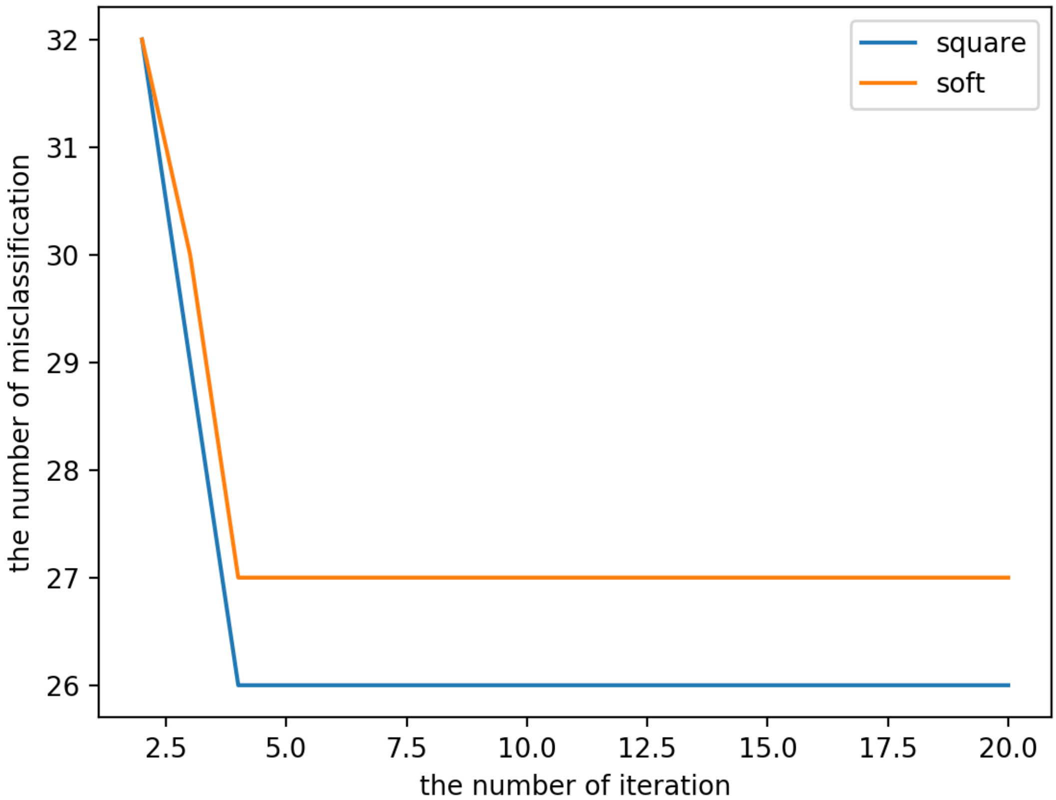
For (4.08)

if $y_p = 1$ $\bar{y}_p = 1$.

$$h(b,\bar{w}) = -\sum_{p=1}^{P} \log(\frac{1}{1+e^{-(b+\bar{x}_p^T\bar{w})}})$$
$$= \sum_{p=1}^{P} \log(1+e^{-(b+\bar{x}_p^T\bar{w})}).$$

$y_p = -1$ $\bar{y}_p = 0$

$$h(b,\bar{w}) = -\sum_{p=1}^{P} \log(1-\frac{1}{1+e^{-(b+\bar{x}_p^T\bar{w})}})$$
$$= \sum_{p=1}^{P} \log(\frac{e^{-(b+\bar{x}_p^T\bar{w})}}{1+e^{-(b+\bar{x}_p^T\bar{w})}})$$
$$= -\sum_{p=1}^{P} \log(\frac{1}{1+e^{(b+\bar{x}_p^T\bar{w})}}) = \sum_{p=1}^{P} \log(1+e^{(b+\bar{x}_p^T\bar{w})})$$



square
soft
32
31
30
29
28
27
26
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
the number of iteration
the number of misclassification