Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 1
Student Name: Siddharth Agrawal
Roll Number: 150716
Date: September 10, 2017

QUESTION 1

Part 1

$$\text{Let, } \mathbf{z} = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \mathbf{R} \text{ denotes the red prototype point}$$

$$\mathbf{G} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \mathbf{G} \text{ denotes the green prototype point}$$

$$\begin{aligned}
\therefore \mathbf{z} - \mathbf{R} &= \begin{bmatrix} x \\ y-1 \end{bmatrix} \\
U(\mathbf{z} - \mathbf{R}) &= \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} x \\ y-1 \end{bmatrix} \\
&= \begin{bmatrix} 3x \\ y-1 \end{bmatrix} \\
\therefore d(\mathbf{z}, \mathbf{R}) &= \langle \mathbf{z} - \mathbf{R}, U(\mathbf{z} - \mathbf{R}) \rangle \\
&= \begin{bmatrix} x & y-1 \end{bmatrix} \times \begin{bmatrix} 3x \\ y-1 \end{bmatrix} \\
&= 3x^2 + (y-1)^2
\end{aligned}$$

$$\begin{aligned}
\therefore \mathbf{z} - \mathbf{G} &= \begin{bmatrix} x-1 \\ y \end{bmatrix} \\
U(\mathbf{z} - \mathbf{R}) &= \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} x-1 \\ y \end{bmatrix} \\
&= \begin{bmatrix} 3(x-1) \\ y \end{bmatrix} \\
\therefore d(\mathbf{z}, \mathbf{G}) &= \langle \mathbf{z} - \mathbf{R}, U(\mathbf{z} - \mathbf{R}) \rangle \\
&= \begin{bmatrix} x-1 & y \end{bmatrix} \times \begin{bmatrix} 3(x-1) \\ y \end{bmatrix} \\
&= 3(x-1)^2 + y^2
\end{aligned}$$

Now, for $\mathbf{z}$ lying on the decision boundary -

$$d(\mathbf{z}, \mathbf{R}) = d(\mathbf{z}, \mathbf{G})$$
$$3x^2 + (y-1)^2 = 3(x-1)^2 + y^2$$
$$3x - y - 1 = 0$$

This is the mathematical expression of the required decision boundary

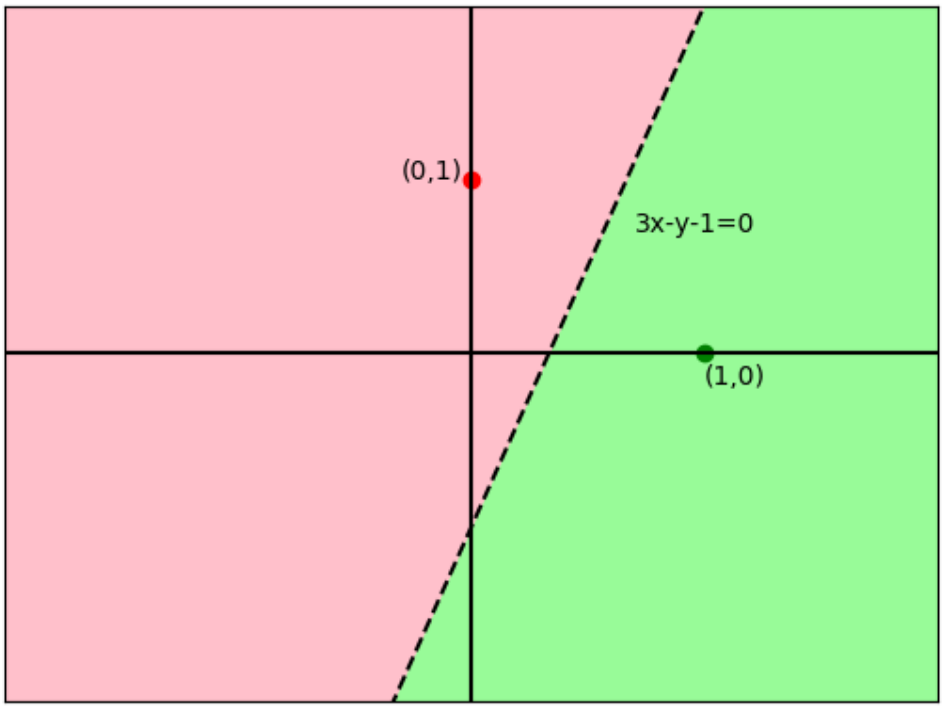

Figure 1: The decision boundary has the expression $3x - y - 1 = 0$

Part 2

$$\text{Let, } \mathbf{z} = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$ **R** denotes the red prototype point

$$\mathbf{G} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$ **G** denotes the green prototype point

$$\begin{aligned}
\therefore \mathbf{z} - \mathbf{R} &= \begin{bmatrix} x \\ y-1 \end{bmatrix} \\
U(\mathbf{z} - \mathbf{R}) &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} x \\ y-1 \end{bmatrix} \\
&= \begin{bmatrix} x \\ 0 \end{bmatrix} \\
\therefore d(\mathbf{z}, \mathbf{R}) &= \langle \mathbf{z} - \mathbf{R}, U(\mathbf{z} - \mathbf{R}) \rangle \\
&= \begin{bmatrix} x & y-1 \end{bmatrix} \times \begin{bmatrix} x \\ 0 \end{bmatrix} \\
&= x^2
\end{aligned}$$

$$\begin{aligned}
\therefore \mathbf{z} - \mathbf{G} &= \begin{bmatrix} x-1 \\ y \end{bmatrix} \\
U(\mathbf{z} - \mathbf{R}) &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} x-1 \\ y \end{bmatrix} \\
&= \begin{bmatrix} x-1 \\ 0 \end{bmatrix} \\
\therefore d(\mathbf{z}, \mathbf{G}) &= \langle \mathbf{z} - \mathbf{R}, U(\mathbf{z} - \mathbf{R}) \rangle \\
&= \begin{bmatrix} x-1 & y \end{bmatrix} \times \begin{bmatrix} x-1 \\ 0 \end{bmatrix} \\
&= (x-1)^2
\end{aligned}$$

Now, for $\mathbf{z}$ lying on the decision boundary -

$$
\begin{aligned}
d(\mathbf{z}, \mathbf{R}) &= d(\mathbf{z}, \mathbf{G}) \\
x^2 &= (x-1)^2 \\
2x - 1 &= 0 \qquad \text{This is the mathematical expression of the required decision boundary}
\end{aligned}
$$

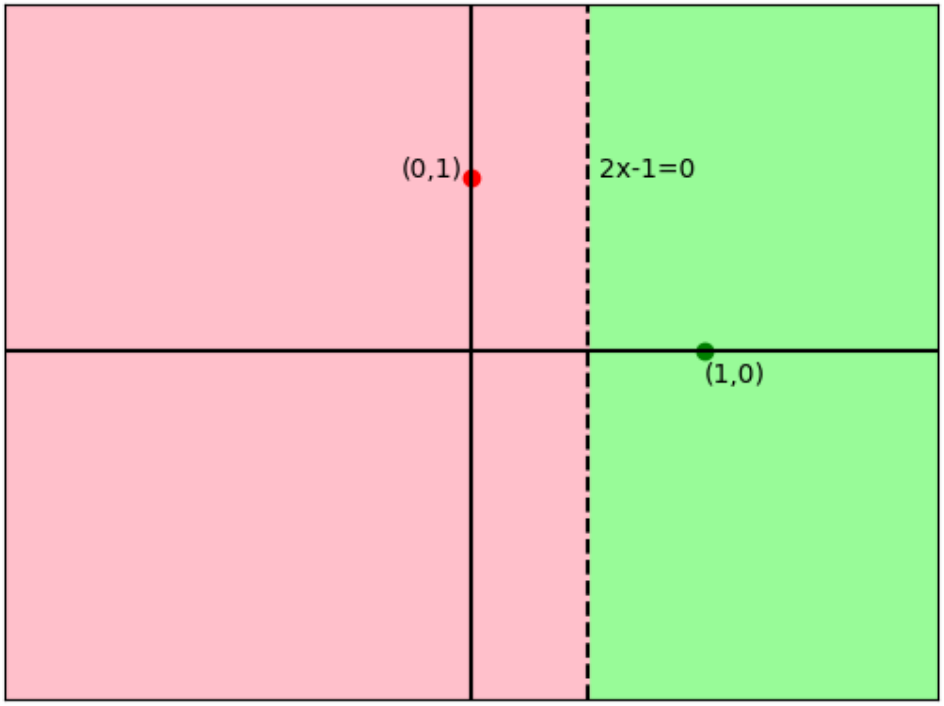

Figure 2: The decision boundary has the expression $2x - 1 = 0$

Prior Distribution

$$\mathbb{P}[\mathbf{w}] = \begin{cases} k & \text{if } ||\mathbf{w}||_2 \leq r, \text{ k is a suitable constant} \\ 0 & \text{otherwise} \end{cases}$$

$$\log \mathbb{P}[\mathbf{w}] = \begin{cases} \log k & \text{if } ||\mathbf{w}||_2 \leq r \\ -\infty & \text{otherwise} \end{cases}$$

Likelihood Distribution

$$\begin{aligned}
\mathbb{P}\left[y|\mathbf{x}^i, \mathbf{w}\right] &= \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) \\
\mathbb{P}[\mathbf{y}|\mathbf{X}, \mathbf{w}] &= \prod_{i=1}^{n} \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right) \\
\log \mathbb{P}[\mathbf{y}|\mathbf{X}, \mathbf{w}] &= \sum_{i=1}^{n} \left(-\log \sqrt{2\pi\sigma^2} - \frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right) \\
&= C - \sum_{i=1}^{n} \frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2} \qquad \triangleright\ C \text{ is a constant}
\end{aligned}$$

Check whether $\hat{\mathbf{w}}_{cls}$ is the MAP estimate of the above defined distribution

$$\begin{aligned}
\hat{\mathbf{w}}_{MAP} &= \underset{||\mathbf{w}|| \in \mathbb{R}^d}{\arg\max} \{\log \mathbb{P}[\mathbf{y}|\mathbf{X}, \mathbf{w}] + \log \mathbb{P}[\mathbf{w}]\} \\
&= \underset{||\mathbf{w}||_2 \leq r}{\arg\max} \left\{ C' - \sum_{i=1}^{n} \frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2} \right\} \qquad \triangleright\ C' = C + \log k, \text{ a constant} \\
&= \underset{||\mathbf{w}||_2 \leq r}{\arg\min} \left\{ \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 \right\} \\
&= \hat{\mathbf{w}}_{cls}
\end{aligned}$$

Prior Distribution

$$\begin{aligned}
\mathbb{P}\left[\mathbf{w}_j\right] &= \mathcal{N}(0, \frac{\sigma^2}{\alpha_j}) \\
\mathbb{P}\left[\mathbf{w}\right] &= \prod_{j=1}^{d} \mathcal{N}(0, \frac{\sigma^2}{\alpha_j}) \\
&= \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{\alpha_j}}} \exp(-\frac{\alpha_j(\mathbf{w}_j)^2}{2\sigma^2}) \\
\log \mathbb{P}\left[\mathbf{w}\right] &= \sum_{j=1}^{d} \left(-\log \sqrt{2\pi\frac{\sigma^2}{\alpha_j}} - \frac{\alpha_j(\mathbf{w}_j)^2}{2\sigma^2} \right)
\end{aligned}$$

Likelihood Distribution

$$\begin{aligned}
\mathbb{P}\left[y|\mathbf{x}^i, \mathbf{w}\right] &= \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) \\
\mathbb{P}\left[\mathbf{y}|\mathbf{X}, \mathbf{w}\right] &= \prod_{i=1}^{n} \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2} \right) \\
\log \mathbb{P}\left[\mathbf{y}|\mathbf{X}, \mathbf{w}\right] &= \sum_{i=1}^{n} \left(-\log \sqrt{2\pi\sigma^2} - \frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2} \right)
\end{aligned}$$

Check whether $\hat{\mathbf{w}}_{fr}$ is the MAP estimate of the above defined distribution

$$
\begin{aligned}
\hat{\mathbf{w}}_{MAP} &= \arg\max_{\mathbf{w}\in\mathbb{R}^d} \{\log \mathbb{P}\left[\mathbf{y}|\mathbf{X},\mathbf{w}\right] + \log \mathbb{P}\left[\mathbf{w}\right]\} \\
&= \arg\max_{\mathbf{w}\in\mathbb{R}^d} \left\{ C - \sum_{i=1}^{n} \frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2} - \sum_{j=1}^{d} \frac{\alpha_j (\mathbf{w}_j)^2}{2\sigma^2} \right\} \qquad \triangleright\ C \text{ is a constant} \\
&= \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \sum_{i=1}^{n} \frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2} + \sum_{j=1}^{d} \frac{\alpha_j (\mathbf{w}_j)^2}{2\sigma^2} \right\} \\
&= \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \sum_{j=1}^{d} \alpha_j (\mathbf{w}_j)^2 \right\} \\
&= \hat{\mathbf{w}}_{fr}
\end{aligned}
$$

Closed form expression for $\hat{\mathbf{w}}_{fr}$

$$\frac{\partial}{\partial \mathbf{w}} \left(\sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \sum_{j=1}^{d} \alpha_j (\mathbf{w}_j)^2 \right) = 0$$

$$\sum_{i=1}^{n} 2(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle) \frac{\partial}{\partial \mathbf{w}} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle) + \sum_{j=1}^{d} \frac{\partial}{\partial \mathbf{w}} (\alpha_j (\mathbf{w}_j)^2) = 0$$

$$\sum_{i=1}^{n} 2(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)(-\mathbf{x}^i) + \sum_{j=1}^{d} 2\alpha_j \mathbf{w}_j = 0$$

$$\sum_{i=1}^{n} (2(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)(-\mathbf{x}^i)) + 2A\mathbf{w} = 0$$

$$\sum_{i=1}^{n} ((\mathbf{x}^i)(\mathbf{x}^i)^\top + A) \cdot \mathbf{w} = \sum_{i=1}^{n} y_i \mathbf{x}^i$$

$$\therefore \hat{\mathbf{w}}_{MAP} = (\mathbf{X}\mathbf{X}^\top + A)^{-1}\mathbf{X}\mathbf{y}$$

where

$$A = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_d \end{bmatrix}$$

From the above expression we get -

$$\hat{\mathbf{w}}_{MAP} = (\mathbf{X}\mathbf{X}^\top + A)^{-1}\mathbf{X}\mathbf{y}$$
$$\therefore \hat{\mathbf{w}}_{fr} = (\mathbf{X}\mathbf{X}^\top + A)^{-1}\mathbf{X}\mathbf{y}$$

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION 4

Assignment Number: 1
Student Name: Siddharth Agrawal
Roll Number: 150716
Date: September 10, 2017

Claim 1: If $\left\{\mathbf{W}^0, \{\xi_i^0\}\right\}$ is an optimum of $(P1)$ then $\mathbf{W}^0$ must be an optimum of $(P2)$

From the constraints in $(P1)$, we have

$$\begin{aligned}
\left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle &\geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i && \forall i, \forall k \neq y^i \\
\xi_i &\geq 1 + \left(\left\langle \mathbf{w}^k, \mathbf{x}_i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \right) && \forall i, \forall k \neq y^i \\
&\geq 1 + \max_{k \neq y^i} \left\langle \mathbf{w}^k, \mathbf{x}_i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle && \forall i \\
\text{But, } \xi_i &\geq 0 && \forall i \\
\therefore \xi_i &\geq \left[1 + \max_{k \neq y^i} \left\langle \mathbf{w}^k, \mathbf{x}_i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \right]_+ && \forall i
\end{aligned}$$

Since ξ_i^0 is optimum, hence -

$$\xi_i^0 = \left[1 + \max_{k \neq y^i} \left\langle \mathbf{w}^k, \mathbf{x}_i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \right]_+$$

Now, we have

$$\begin{aligned}
\boldsymbol{\eta}^i &= \left\langle \mathbf{W}, \mathbf{x}^i \right\rangle \\
&= \left[\langle \mathbf{w}_1, \mathbf{x}^i \rangle, \langle \mathbf{w}_2, \mathbf{x}^i \rangle, ..., \langle \mathbf{w}_k, \mathbf{x}^i \rangle \right] \\
&= \left[\boldsymbol{\eta}_1^i, \boldsymbol{\eta}_2^i, ..., \boldsymbol{\eta}_k^i \right]
\end{aligned}$$

$$\begin{aligned}
\xi_i^0 &= \left[1 + \max_{k \neq y^i} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_{y^i}^i \right]_+ \\
&= \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)
\end{aligned}$$

Hence, we see that the optimum value of the slack variable comes out to be equal to the Cramer-Singer Loss function. Now substitute this to the equation $(P1)$-

$$\begin{aligned}
\widehat{\mathbf{W}} &= \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\| \mathbf{w}^k \right\|_2^2 + \sum_{i=1}^{n} \xi_i^0 \\
&= \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\| \mathbf{w}^k \right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)
\end{aligned}$$

Hence $\widehat{\mathbf{W}}$ is the optimum solution of equation $(P2)$

Claim 2: If $\mathbf{W}^1$ is an optimum of $(P2)$ then $\exists \xi_i^1$ s.t. $\left\{\mathbf{W}^1, \{\xi_i^1\}\right\}$ is an optimum of $(P1)$

Let:

$$f(\mathbf{W}) = \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$$
$$h(\mathbf{W}, \xi_i) = \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i^0$$

Suppose that our claim is false. Assume $\{\mathbf{W}^0, \xi^0\}$ is an optimum of $(P1)$.
Suppose that $\forall i$, some $\xi^1 = \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$ where $\boldsymbol{\eta}^i = \left\langle \mathbf{W}^1, \mathbf{x}^i \right\rangle$
.
Then,

$$\begin{aligned} f(\mathbf{W}^1) &< f(\mathbf{W}^0) \\ f(\mathbf{W}^1) &< h(\mathbf{W}^0, \xi_i^0) && \text{Using claim 1} \\ h(\mathbf{W}^1, \xi_i^1) &< h(\mathbf{W}^0, \xi_i^0) \end{aligned}$$

This gives us a contradiction since we assumed that $\{\mathbf{W}^0, \xi_i^0\}$ is an optimum of $(P1)$ but here actually $\{\mathbf{W}^1, \xi^1\}$ gives us a lower value.
Hence our claim is correct.

Both claim 1 and claim 2 signify equivalence between the two formulations.

To show that $\mathbf{g} \in \partial f(\mathbf{w})$, we need to prove -

$$f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$$
$$\sum_{i=1}^{n} \left[1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right]_+ \geq \sum_{i=1}^{n} \left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ + \left\langle \sum_{i}^{n} \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle$$
$$\sum_{i=1}^{n} \left[1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right]_+ \geq \sum_{i=1}^{n} \left(\left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ + \left\langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle \right)$$

Now, to prove the above, we will make a stronger claim that -

$$\forall i \in [n],\ \left[1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right]_+ \geq \left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ + \left\langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle$$

Proof for the above claim:
Case 1: $1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle > 0$

$$\begin{aligned}
\left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ &= 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle + y^i \langle \mathbf{w}', \mathbf{x}^i \rangle - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle \\
&= \left(1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right) + y_i \langle \mathbf{w}' - \mathbf{w}, x^i \rangle \\
\left[1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right]_+ &\geq \left(1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right) \\
&= \left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ + \langle -y^i \mathbf{x}^i, \mathbf{w}' - \mathbf{w} \rangle \\
&= \left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ + \langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \rangle
\end{aligned}$$

Case 2: $1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$
By definition, $\mathbf{h}^i = 0$ and $\left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ = 0$
Therefore, RHS=0
Also, by definition, $\left[1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right]_+ \geq 0$
Therefore, LHS≥0

$$\therefore \text{LHS} \geq \text{RHS}$$

Thus we see that our claim is correct.

Now, sum our claim over i to obtain

$$\sum_{i=1}^{n} \left[1 - y^i \langle \mathbf{w}', \mathbf{x}^i \rangle\right]_+ \geq \sum_{i=1}^{n} \left(\left[1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ + \left\langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle \right)$$

Hence, $\mathbf{g} \in \partial f(\mathbf{w})$

Part 1

The plot between accuracy and k for the k-NN algorithm (with Euclidean metric) is -

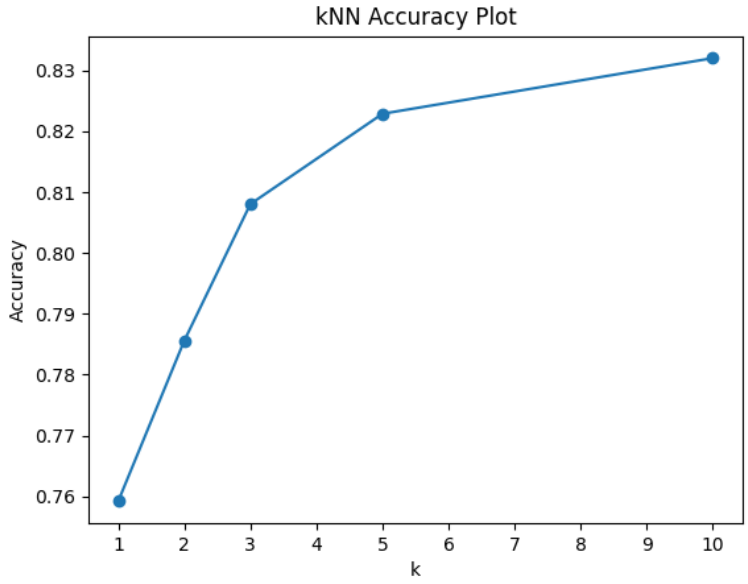

Figure 3: Note that the value of accuracy increases with k

We observe that by increasing k, our accuracy increases. So why is it so?? Actually, for very small values of k, the decision boundary is very 'strict', i.e., each point is separated by the boundaries. Now for any outlier, the boundaries change very drastically and we may say that the partition is too *'jumpy'*. This in turn causes *'overfitting'*, and hence the low accuracy.

But for large values of k (not shown here), the partitioning becomes too *'smooth'* and hence causes *'underfitting'*

In the next part we find the optimum value of k, which maximizes the accuracy.

Part 2

For the validation part, I have used the k-fold validation technique.
Number of folds = 6
Number of training data points considered - 60000(all)
Optimum value of k obtained = 13

Part 3

For the Mahalanobis metric, I have used the following specifications -
Number of training data points considered = 10000
maxiter value considered = 10000

Accuracy obtained by replacing the Euclidean metric with the Mahalanobis metric and using the optimum value of k in *test.py* = 83.835%