CIS5200: Machine Learning	Spring 2025
Homework 1	
Release Date: January 30, 2025	*Due Date: February 14, 2025*

Name: Haoze Wu
PennKey: haozewu
Collaborators: [List any collaborators here, AI and/or Human]

Note: This document is a read-only file. To create an editable version click on Menu in the top left corner of your screen and choose the Copy Project option.

Problem 1: Margin Perceptron

1.1 Proof:
To prove

$$\omega_\star^\top \omega_{t+1} \geq \omega_\star^\top \omega_t + \gamma \tag{1}$$

we can start by expanding the left-hand side following the Margin Perceptron algorithm:

$$\begin{aligned} \omega_\star^\top \omega_{t+1} &= \omega_\star^\top (\omega_t + y_i x_i) \\ &= \omega_\star^\top \omega_t + y_i \omega_\star^\top x_i \end{aligned} \tag{2}$$

Since all data is linearly separated by the hyperplane defined by $\omega_\star$, we have $y_i(\omega_\star^\top x_i) > 0$. And the lable $|y_i| = 1$. So, we have

$$\begin{aligned} \omega_\star^\top \omega_{t+1} &= \omega_\star^\top \omega_t + y_i \omega_\star^\top x_i \\ &= \omega_\star^\top \omega_t + \omega_\star^\top x_i \end{aligned} \tag{3}$$

Given by the definition of the margin γ, where

$$\gamma = \min_{i \in \{1,\ldots,m\}} |\omega_\star^\top x_i| \tag{4}$$

we than have

$$\begin{aligned} \omega_\star^\top \omega_{t+1} &= \omega_\star^\top \omega_t + \omega_\star^\top x_i \\ &\geq \omega_\star^\top \omega_t + \gamma \end{aligned} \tag{5}$$

Proved.

1.2 Proof:
To prove

$$||\omega_{t+1}||_2^2 \leq ||\omega_t||_2^2 + 3 \tag{6}$$

we may start with the left-hand side:

$$\begin{aligned} ||\omega_{t+1}||_2^2 &= ||\omega_t + y_i x_i||_2^2 \\ &= ||\omega_t||_2^2 + 2 y_i \omega_t^\top x_i + ||y_i x_i||_2^2 \end{aligned} \tag{7}$$

For the term $||y_i x_i||_2^2$, since $|y_i| = 1$, we have

$$||y_i x_i||_2^2 = ||x_i||_2^2 \tag{8}$$

For the term $2y_i\omega_t^\top x_i$, we may consider the rule of update in the Margin Perceptron algorithm. If $y_i \neq \text{sign}(\omega_t^\top x_i)$, we have

$$2y_i\omega_t^\top x_i \leq 0 \tag{9}$$

Thus, since all data sample have been processed under the Feature Scaling procedure, we have $||x_i||_2^2 \leq 1$, and thus,

$$\begin{aligned} ||\omega_{t+1}||_2^2 &= ||\omega_t||_2^2 + 2y_i\omega_t^\top x_i + ||y_i x_i||_2^2 \\ &\leq ||\omega_t||_2^2 + 1 \\ &\leq ||\omega_t||_2^2 + 3 \end{aligned} \tag{10}$$

If the update is due to $|\omega_t^\top x_i| < 1$, we have

$$|2y_i\omega_t^\top x_i| \leq 2 \tag{11}$$

Thus, we have

$$\begin{aligned} ||\omega_{t+1}||_2^2 &= ||\omega_t||_2^2 + 2y_i\omega_t^\top x_i + ||y_i x_i||_2^2 \\ &\leq ||\omega_t||_2^2 + |2y_i\omega_t^\top x_i| + ||y_i x_i||_2^2 \\ &\leq ||\omega_t||_2^2 + 2 + 1 \\ &\leq ||\omega_t||_2^2 + 3 \end{aligned} \tag{12}$$

Proved.

1.3 Proof:
From the Growth Lemma proved in 1.1, we have:

$$\begin{aligned} \omega_\star^\top \omega_{T+1} &\geq \omega_\star^\top \omega_T + \gamma \\ \omega_\star^\top \omega_{T+1} &\geq \omega_\star^\top \omega_{T-1} + 2\gamma \\ &\dots \\ \omega_\star^\top \omega_{T+1} &\geq \omega_\star^\top \omega_1 + \gamma T \end{aligned} \tag{13}$$

Since the initilization of ω_1 is $\mathbf{0}$, we have

$$\omega_\star^\top \omega_{T+1} \geq \gamma T \tag{14}$$

Also notice that

$$\begin{aligned} \omega_\star^\top \omega_{T+1} &\leq |\omega_\star^\top \omega_{T+1}| \\ &\leq ||\omega_\star||_2 ||\omega_{T+1}||_2 \\ &= ||\omega_{T+1}||_2 \end{aligned} \tag{15}$$

Thus, we have

$$||\omega_{T+1}||_2 \geq \omega_\star^\top \omega_{T+1} \geq \gamma T \tag{16}$$

From the Control Lemma proved in 1.2, we have:

$$
\begin{aligned}
||\omega_{T+1}||_2^2 &\leq ||\omega_T||_2^2 + 3 \\
&\leq ||\omega_{T-1}||_2^2 + 6 \\
\dots \\
&\leq ||\omega_1||_2^2 + 3T \\
&\leq 3T
\end{aligned}
\tag{17}
$$

] which is equivalent to

$$
||\omega_{T+1}||_2 \leq \sqrt{3T} \tag{18}
$$

Thus, combining the two inequalities, we have

$$
\gamma T \leq ||\omega_{T+1}||_2 \leq \sqrt{3T} \tag{19}
$$

Proved.

1.4 Proof:

From the conclusion in 1.3, we have:

$$
\begin{aligned}
\gamma T &\leq \sqrt{3T} \\
\gamma^2 T^2 &\leq 3T \\
\gamma^2 T &\leq 3 \\
T &\leq \frac{3}{\gamma^2}
\end{aligned}
\tag{20}
$$

Proved.

1.5 Proof:

Without losing generality, assume the Margin Perceptron algorithm ends after $T+1$ iterations, i.e., the output hyperplane is defined by ω_T. We need to prove that

$$
\min_i \frac{\omega_{T+1}^\top x_i}{||\omega_{T+1}||2} \geq \frac{\gamma}{3} \tag{21}
$$

From the conclusion in 1.3, we have

$$
\min_i \frac{\omega_{T+1}^\top x_i}{||\omega_{T+1}||2} \geq \min_i \frac{\omega_{T+1}^\top x_i}{\sqrt{3T}} = \frac{1}{\sqrt{3T}} \min_i \omega_{T+1}^\top x_i \tag{22}
$$

And from the conclusion from 1.4, we have:

$$
\frac{1}{\sqrt{3T}} \min_i \omega_{T+1}^\top x_i \geq \frac{\gamma}{3} \min_i \omega_{T+1}^\top x_i \tag{23}
$$

To prove the statement given, we need to prove

$$
\min_i \omega_{T+1}^\top x_i \geq 1 \tag{24}
$$

From the update rule of the Margin Perceptron algorithm, when the algorithm ends and output the final result ω_{T+1}, we must have

$$|\omega_{T+1}^{\top} x_i| \geq 1 \tag{25}$$

for all $i \in \{1, \dots, m\}$. Thus, we have

$$\min_i \omega_{T+1}^{\top} x_i \geq 1 \tag{26}$$

which is equivalent to

$$\min_i \frac{\omega_{T+1}^{\top} x_i}{||\omega_{T+1}||2} \geq \frac{\gamma}{3} \tag{27}$$

Proved.

1.6 Answer:
This Margin Perceptron algorithm is desirable to learn a predictor that has a large margin is because it treats those correctly classified samples with a margin less than 1 as misclassified samples, and then try to update the hyperplane to make the margin, i.e., the distance from the data point to the hyperplane, larger.

Problem 2: Bayes Optimal Classifier and Squared Loss

2.1 Proof:
We may first expand the expression of the expected squared loss:

$$\mathbb{E}_{y|x}[(h(x) - y)^2] = \mathbb{E}_{y|x}[(h(x)^2 - 2h(x)y + y^2)] \tag{28}$$

When the partial derivative of the expected squared loss with respect to $h(x)$ is zero, we have

$$\frac{\partial \mathbb{E}_{y|x}[(h(x) - y)^2]}{\partial h(x)} = 0 \tag{29}$$

which leads to

$$\begin{aligned} \frac{\partial \mathbb{E}_{y|x}[(h(x) - y)^2]}{\partial h(x)} &= \frac{\partial \mathbb{E}_{y|x}[(h(x)^2 - 2h(x)y + y^2)]}{\partial h(x)} \\ &= 2h(x) - 2\mathbb{E}_{y|x}[y] = 0 \end{aligned} \tag{30}$$

Thus, we have

$$h^{\star}(x) = \mathbb{E}_{y|x}[y] \tag{31}$$

For $\mathbb{E}_{y|x}[y]$, we have

$$\begin{aligned} \mathbb{E}_{y|x}[y] &= 1 \cdot P(y = 1|x) + (-1) \cdot P(y = -1|x) \\ &= P[y = 1|x] + (-1) \cdot (1 - P[y = 1|x]) \\ &= 2P[y = 1|x] - 1 \\ &= 2\eta(x) - 1 \end{aligned} \tag{32}$$

Proved.

2.2 Proof:
From Baye's Theorem, we have:

$$\begin{aligned} \eta(x) &= \Pr[y=1|x] \\ &= \frac{\Pr[x|y=1]\Pr[y=1]}{\Pr[x]} \end{aligned} \tag{33}$$

And for the term $\Pr[x]$, by using the law of total probability, we have:

$$\begin{aligned} \Pr[x] &= \Pr[x|y=1]\Pr[y=1] + \Pr[x|y=-1]\Pr[y=-1] \\ &= \frac{1}{2}\Pr[x|y=1] + \frac{1}{2}\Pr[x|y=-1] \end{aligned} \tag{34}$$

Thus, we have:

$$\begin{aligned} \eta(x) &= \frac{\Pr[x|y=1]\Pr[y=1]}{\Pr[x]} \\ &= \frac{\Pr[x|y=1]\Pr[y=1]}{\frac{1}{2}\Pr[x|y=1] + \frac{1}{2}\Pr[x|y=-1]} \\ &= \frac{\frac{1}{2}\mathcal{N}(\mu, I)}{\frac{1}{2}\mathcal{N}(\mu, \mathcal{I}) + \frac{1}{2}\mathcal{N}(-\mu, I)} \end{aligned} \tag{35}$$

Consider the Probability Density Function of the Guassian Distribution is

$$\mathcal{N}(\mu, I) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x-\mu)^2) \tag{36}$$

, and substitute the PDF into the equation, we have

$$\begin{aligned} \eta(x) &= \frac{\frac{1}{2}\mathcal{N}(\mu, I)}{\frac{1}{2}\mathcal{N}(\mu, \mathcal{I}) + \frac{1}{2}\mathcal{N}(-\mu, I)} \\ &= \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x-\mu)^2)}{\frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x-\mu)^2) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x+\mu)^2)} \\ &= \frac{\exp(-\frac{1}{2}(x-\mu)^2)}{\exp(-\frac{1}{2}(x-\mu)^2) + \exp(-\frac{1}{2}(x+\mu)^2)} \\ &= \frac{1}{1+\exp(-2\mu^\top x)} \end{aligned} \tag{37}$$

Proved.

2.3 Answer:
Combining the expression of $\eta(x)$ in 2.1 and 2.2, we have

$$h^\star(x) = 2\eta(x) - 1 = \frac{2}{1+\exp(-2\mu^\top x)} - 1 \tag{38}$$

and if $h^\star(x) = 0$, we have

$$\begin{aligned}\frac{2}{1+\exp(-2\mu^\top x)} - 1 &= 0 \\ \exp(-2\mu^\top x) &= 1 \\ -2\mu^\top x &= 0 \\ \mu^\top x &= 0\end{aligned} \tag{39}$$

, which means the decision boundary is defined by the hyperplane $\mu^\top x = 0$, indicating that $\omega = \mu$ and $b = 0$ for this case. When $\mu^\top x \geq 0$, this model would predict that $y = 1$, and when $\mu^\top x < 0$, this model would predict that $y = -1$.

Problem 3: k-NN Analysis

3.1 Proof:
Without losing generality, assume the difference between the kth coodinate of x and x' is ϵ, i.e., $x_k - x'_k = \epsilon$. From the triangle inequality, we have

$$\text{dist}(x, z) - \text{dist}(x', z) \leq \text{dist}(x, x') \tag{40}$$

For the distance between x and x', we have

$$\begin{aligned}\text{dist}(x, x') &= \sqrt{\sum_{i=1}^{d}(x_i - x'_i)^2} \\ &= \sqrt{\sum_{i=1}^{k-1}(x_i - x'_i)^2 + (x_k - x'_k)^2 + \sum_{i=k+1}^{d}(x_i - x'_i)^2} \\ &= \sqrt{(x_k - x'_k)^2} \\ &= \epsilon\end{aligned} \tag{41}$$

Thus, we have

$$\text{dist}(x, z) - \text{dist}(x', z) \leq \epsilon \tag{42}$$

Proved.

This conclusion also suggests that the k-NN classifier is quite robust to small perturbations in the test point data. When using suitable distance measurement, the k-NN classifier can still yield the correct prediction even if the test point data is slightly perturbed.

3.2 Proof:
From the conclusion in 3.1, we have

$$\text{dist}(y, x) - \text{dist}(y', x) \leq \epsilon \tag{43}$$

if there is a difference of ϵ in any one coordinate of y and y' out of the total d coordinates where x is the test point and y and y' are the nearest and second nearest traning points. If we perturb each coordinate of x by at most $\epsilon = \frac{\Delta}{2d}$, where $\Delta = \text{dist}(y, x) - \text{dist}(y', x)$, denoting the newly perturbed test point as x', we then have:

$$
\begin{aligned}
|\text{dist}(x', u) - \text{dist}(x, u)| \leq d \cdot \frac{\Delta}{2d} = \frac{\Delta}{2} \\
-\frac{\Delta}{2} \leq \text{dist}(x', u) - \text{dist}(x, u) \leq \frac{\Delta}{2} \\
\text{dist}(x, u) - \frac{\Delta}{2} \leq \text{dist}(x', u) \leq \text{dist}(x, u) + \frac{\Delta}{2}
\end{aligned}
\tag{44}
$$

for any training point u. Thus, for both the nearest neighbor y and second nearest neighbor y', we have:

$$
\begin{aligned}
\text{dist}(x', y) \leq \text{dist}(x, y) + \frac{\Delta}{2} \\
\text{dist}(x', y') \geq \text{dist}(x, y') - \frac{\Delta}{2}
\end{aligned}
\tag{45}
$$

Thus, we then consider the difference between the distance between the nearest neighbor y and the test point x' and the distance between the second nearest neighbor y' and the test point x':

$$
\begin{aligned}
\text{dist}(x', y') - \text{dist}(x', y) &\geq (\text{dist}(x, y') - \frac{\Delta}{2}) - (\text{dist}(x, y) + \frac{\Delta}{2}) \\
&= \text{dist}(x, y') - \text{dist}(x, y) - \Delta \\
&= \Delta - \Delta \\
&= 0
\end{aligned}
\tag{46}
$$

Since the difference in distances remains non-negative, the distance between the test point x' and the nearest neighbor y is still less than or equal to the distance between the test point x' and the second nearest neighbor y', which means the nearest training point and the prediction of this 1-NN classifier remains unchanged after the perturbation on the test point x.

Proved.