Jared Yu
Data Mining
Module 7 Assignment

1. Question 1
   a. Show that the distance from the hyperplane $g(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + w_0 = 0$ to the point $\mathbf{x}$ is $|g(\mathbf{x})|/\|\mathbf{w}\|$ by minimizing $\left\|\mathbf{x} - \mathbf{x_q}\right\|^2$ subject to the constraint $g(\mathbf{x_q}) = 0$.

<u>Ans:</u> References: [1.1], [1.2]

To solve this, I will use Lagrange multipliers. We are asked to minimize $\left\|\mathbf{x} - \mathbf{x_q}\right\|^2$ subject to the constraint $g(\mathbf{x_q}) = 0$. Then the function for the Lagrange multipliers is in the form of,

$$\begin{aligned}\mathcal{L}(\mathbf{x}, \lambda) &= f(\mathbf{x}) - \lambda(\mathbf{w}^\top\mathbf{x} + w_0)\\ &= \left(\mathbf{x} - \mathbf{x_q}\right)^\top\left(\mathbf{x} - \mathbf{x_q}\right) - \lambda\mathbf{w}^\top\mathbf{x} - \lambda w_o\\ &= \mathbf{x}^\top\mathbf{x} - 2\mathbf{x}^\top\mathbf{x_q} + \mathbf{x_q}^\top\mathbf{x_q} - \lambda\mathbf{w}^\top\mathbf{x} - \lambda w_o\end{aligned}$$

where

$$f(\mathbf{x}) = \left\|\mathbf{x} - \mathbf{x_q}\right\|^2.$$

Then we must minimize this function w.r.t. each of the variables,

$$\mathcal{L}_\mathbf{x} = 0, \mathcal{L}_\lambda = 0.$$

$$\mathcal{L}_\mathbf{x} = \frac{d\mathcal{L}}{d\mathbf{x}} = 2\mathbf{x} - 2\mathbf{x_q} + 0 - \lambda\mathbf{w} + 0 = 2\mathbf{x} - 2\mathbf{x_q} - \lambda\mathbf{w} = 0$$

$$\rightarrow \mathbf{x} = \mathbf{x_q} + \frac{1}{2}\lambda\mathbf{w}$$

$$\mathcal{L}_\lambda = \frac{d\mathcal{L}}{d\lambda} = -\mathbf{w}^\top\mathbf{x} - w_o = 0$$

$$\rightarrow -\mathbf{w}^\top\left(\mathbf{x_q} + \frac{1}{2}\lambda\mathbf{w}\right) - w_o = 0$$

$$\rightarrow -\mathbf{w}^\top\mathbf{x_q} - \frac{1}{2}\lambda\mathbf{w}^\top\mathbf{w} - w_0 = 0$$

$$\rightarrow \frac{1}{2}\lambda\mathbf{w}^\top\mathbf{w} = -\mathbf{w}^\top\mathbf{x_q} - w_0$$

$$\rightarrow \lambda = -2\frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}$$

$$\Rightarrow\ \mathbf{x} = \mathbf{x_q} + \frac{1}{2}\left(-2\frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}\right)\mathbf{w}$$

$$\rightarrow\ \mathbf{x} = \mathbf{x_q} - \frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}$$

$$\Rightarrow \left\|\mathbf{x} - \mathbf{x_q}\right\|^2 = \left\|\left(\mathbf{x_q} - \frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}\right) - \mathbf{x_q}\right\|^2 = \left\|\frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}\right\|^2 = \left(\frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}\right)^2\|\mathbf{w}\|^2 = \frac{g(\mathbf{x})^2}{\|\mathbf{w}\|^2}$$

Therefore, the distance after taking the square root can be seen to as follows,

$$\Rightarrow \left\|\mathbf{x} - \mathbf{x_q}\right\| = \frac{|g(\mathbf{x})|}{\|\mathbf{w}\|}\ \blacksquare$$

   b. Show that the projection of $\mathbf{x}$ onto the hyperplane is given by

$$\mathbf{x_p} = \mathbf{x} - \frac{g(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}.$$

<u>Ans:</u>

To prove this, we will first indicate what $\mathbf{w}$ is. The textbook states that if $\mathbf{x}_1$ and $\mathbf{x}_2$ are both on the decision surface, then

$$\mathbf{w}'\mathbf{x}_1 + w_0 = \mathbf{w}'\mathbf{x}_2 + w_0$$

or

$$\mathbf{w}'(\mathbf{x}_1 - \mathbf{x}_2) = 0.$$

This indicates that the constant vector $\mathbf{w}$ is actually normal or perpendicular to the hyperplane.

Then, using the result from part a), we have that the distance between some arbitrary vector $\mathbf{x}$ and the hyperplane can be found with $\frac{|g(\mathbf{x})|}{||\mathbf{w}||}$. What we want to do then is to multiply this minimum distance by $\frac{\mathbf{w}}{||\mathbf{w}||}$, which is the unit vector form of $\mathbf{w}$. Furthermore, let $\mathbf{x_p}$ represent the projection of $\mathbf{x}$ onto the hyperplane. This leads us to the following formula,

$$\mathbf{x_p} = \mathbf{x} - \frac{g(\mathbf{x})}{||\mathbf{w}||}\frac{\mathbf{w}}{||\mathbf{w}||} = \mathbf{x} - \frac{g(\mathbf{x})}{||\mathbf{w}||^2}\mathbf{w}. \blacksquare$$

2. Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be n q-dimensional samples and Q be any nonsingular positive definite $q \times q$ matrix. Show that the vector $\mathbf{x}$ that minimizes

$$\sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{x})^\top Q^{-1}(\mathbf{x}_k - \mathbf{x})$$

Is the sample mean, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$.

<u>Ans:</u> References: [2.1], [2.2]

Let the function $f(\mathbf{x})$ be defined as follows

$$f(\mathbf{x}) = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{x})^\top Q^{-1}(\mathbf{x}_k - \mathbf{x}).$$

To try and find the vector $\mathbf{x}$ that minimizes, we must first take the gradient w.r.t. $\mathbf{x}$. To begin, we can try to simplify $f(\mathbf{x})$.

$$f(\mathbf{x}) = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{x})^\top Q^{-1}(\mathbf{x}_k - \mathbf{x}) = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{x})^\top (Q^{-1}\mathbf{x}_k - Q^{-1}\mathbf{x})$$

$$= \sum_{k=1}^{n} \mathbf{x}_k^\top Q^{-1}\mathbf{x}_k - \mathbf{x}_k^\top Q^{-1}\mathbf{x} - \mathbf{x}^\top Q^{-1}\mathbf{x}_k + \mathbf{x}^\top Q^{-1}\mathbf{x}$$

Next, we can find the derivative of this function by utilizing the derivative of an inverse matrix w.r.t. a vector.

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} \sum_{k=1}^{n} \mathbf{x}_k^\top Q^{-1}\mathbf{x}_k - \mathbf{x}_k^\top Q^{-1}\mathbf{x} - \mathbf{x}^\top Q^{-1}\mathbf{x}_k + \mathbf{x}^\top Q^{-1}\mathbf{x}$$

$$= \sum_{k=1}^{n} \frac{\partial(\mathbf{x}_k^\top Q^{-1}\mathbf{x}_k)}{\partial \mathbf{x}} - \frac{\partial(\mathbf{x}_k^\top Q^{-1}\mathbf{x})}{\partial \mathbf{x}} - \frac{\partial(\mathbf{x}^\top Q^{-1}\mathbf{x}_k)}{\partial \mathbf{x}} + \frac{\partial(\mathbf{x}^\top Q^{-1}\mathbf{x})}{\partial \mathbf{x}}$$

$$\Rightarrow \sum_{k=1}^{n} -(\mathbf{x}_k^\top Q^{-1})^\top - Q^{-1}\mathbf{x}_k + [Q^{-1} + (Q^{-1})^\top]\mathbf{x} \overset{\text{set to}}{\triangleq} 0$$

$$\rightarrow \sum_{k=1}^{n} -[(Q^{-1})^\top + Q^{-1}]\mathbf{x}_k + [Q^{-1} + (Q^{-1})^\top]\mathbf{x} = 0$$

$$\rightarrow n[Q^{-1} + (Q^{-1})^\top]\mathbf{x} = [Q^{-1} + (Q^{-1})^\top]\sum_{k=1}^{n} \mathbf{x}_k$$

$$\rightarrow n[Q^{-1} + (Q^{-1})^\top]^{-1}\,[Q^{-1} + (Q^{-1})^\top]\mathbf{x} = [Q^{-1} + (Q^{-1})^\top]^{-1}[Q^{-1} + (Q^{-1})^\top]\sum_{k=1}^{n} \mathbf{x}_k$$

$$\rightarrow n\mathbf{x} = \sum_{k=1}^{n} \mathbf{x}_k$$

$$\rightarrow \mathbf{x}^* = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

Then to show that it is indeed the minimum, the second derivative must also be examined.

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} = \frac{\partial}{\partial \mathbf{x}}\left\{\sum_{k=1}^{n} -[(Q^{-1})^\top + Q^{-1}]\mathbf{x}_k + [Q^{-1} + (Q^{-1})^\top]\mathbf{x}\right\}$$

$$= \sum_{k=1}^{n} -\frac{\partial}{\partial \mathbf{x}}\{[(Q^{-1})^\top + Q^{-1}]\mathbf{x}_k\} + \frac{\partial}{\partial \mathbf{x}}\{[Q^{-1} + (Q^{-1})^\top]\mathbf{x}\}$$

$$= \sum_{k=1}^{n} [Q^{-1} + (Q^{-1})^\top] = n[Q^{-1} + (Q^{-1})^\top]$$

Then, since Q^{-1} is nonsingular positive definite, then $n[Q^{-1} + (Q^{-1})^\top]$ is positive definite. Therefore, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$ can be said to be the point that minimizes $\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{x})^\top\, Q^{-1}(\mathbf{x}_k - \mathbf{x})$. ■

3. Consider a linear classifier with discriminant functions $g_i(\mathbf{x}) = \mathbf{w}_i^\top\mathbf{x} + w_{i0},\ i = 1, \cdots, c$. Show that the decision regions are convex by showing that if $\mathbf{x}_1 \in \mathcal{R}_i$ and $\mathbf{x}_2 \in \mathcal{R}_i$ then $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in \mathcal{R}_i$ if $0 \le \lambda \le 1$.

Ans: References: [3.1], [3.2], [3.3], [3.4], [3.5]

Let us define $\hat{\mathbf{x}} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$, where $0 \le \lambda \le 1$, as the convex combination of vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. Furthermore, the set of vectors within $\mathcal{R}_i$ is convex if it contains all possible convex combinations of vectors. If this can be shown to be the case, then that implies that all decision regions $\mathcal{R}_i$, for $i = 1, \cdots, c$ are also convex.

Based on the linearity of the classifier, $g_i(\mathbf{x})$, we can also write

$$\begin{aligned} g_i(\hat{\mathbf{x}}) &= \mathbf{w}_i^\top(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) + w_{i0} \\ &= \lambda\mathbf{w}_i^\top\mathbf{x}_1 + (1-\lambda)\mathbf{w}_i^\top\mathbf{x}_2 + w_{i0} - \lambda w_{i0} + \lambda w_{i0} \\ &= \lambda\mathbf{w}_i^\top\mathbf{x}_1 + (1-\lambda)\mathbf{w}_i^\top\mathbf{x}_2 + (1-\lambda)w_{i0} + \lambda w_{i0} \\ &= \lambda(\mathbf{w}_i^\top\mathbf{x}_1 + w_{i0}) + (1-\lambda)(\mathbf{w}_i^\top\mathbf{x}_2 + w_{i0}) \\ &= \lambda g_i(\mathbf{x}_1) + (1-\lambda)g_i(\mathbf{x}_2). \end{aligned}$$

Now, since $\mathbf{x}_1 \in \mathcal{R}_i$ and $\mathbf{x}_2 \in \mathcal{R}_i$, and the weights λ and $(1-\lambda)$ are positive, then the following also holds,

$$\Rightarrow \lambda g_i(\mathbf{x}_1) > \lambda g_j(\mathbf{x}_1)\ \forall i \neq j$$
$$\Rightarrow (1-\lambda) g_i(\mathbf{x}_2) > (1-\lambda) g_j(\mathbf{x}_2)\ \forall i \neq j.$$
From this it follows that,
$$\Rightarrow\ \lambda g_i(\mathbf{x}_1) + (1-\lambda) g_i(\mathbf{x}_2) > \lambda g_j(\mathbf{x}_1) + (1-\lambda) g_j(\mathbf{x}_2)\ \forall i \neq j.$$
Therefore, it can be concluded that,
$$\Rightarrow g_i(\hat{\mathbf{x}}) > g_j(\hat{\mathbf{x}})\ \forall i \neq j.$$
This shows then that the decision regions $\mathcal{R}_i$, $i = 1, \cdots, c$ are convex. ■

4. In the gradient descent algorithm, $\mathbf{a}_{k+1}$ is obtained from $\mathbf{a}_k$ by
$$\mathbf{a}_{k+1} = \mathbf{a}_k - \rho_k \nabla \mathbf{J}(\mathbf{a}_k),$$
where ρ_k is a positive scale factor that sets the step size. Consider the criterion function
$$J_q(\mathbf{a}) = \sum_{y \in \mathcal{Y}} (\mathbf{a}^\top \mathbf{y} - b)^2$$
where $\mathcal{Y}(\mathbf{a})$ is the set of samples for which $\mathbf{a}^\top \mathbf{y} \leq b$. Suppose that $\mathbf{y}_1$ is the only sample in $\mathcal{Y}(\mathbf{a}_k)$. Show that $\nabla \mathbf{J}_q(\mathbf{a}_k) = 2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\mathbf{y}_1$ and that the matrix of second partial derivatives is given by $D = 2\mathbf{y}_1\mathbf{y}_1^\top$. Use this to show that when the optimal ρ_k is used in the gradient descent algorithm,
$$\mathbf{a}_{k+1} = \mathbf{a}_k + \frac{b - \mathbf{a}^\top \mathbf{y}_1}{\|\mathbf{y}_1\|^2}\mathbf{y}_1.$$

<u>Ans:</u> Reference: [4.1, 2.2]
The first step is to show that $\nabla \mathbf{J}_q(\mathbf{a}_k) = 2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\mathbf{y}_1$. In the case where $\mathcal{Y}(\mathbf{a}_k)$ only contains $\mathbf{y}_1$, then $J_q(\mathbf{a}) = (\mathbf{a}^\top \mathbf{y}_1 - b)^2$. Finding the derivative of this w.r.t. $\mathbf{a}$, we find that,
$$\frac{\partial}{\partial \mathbf{a}_k} J_q(\mathbf{a}_k) = \frac{\partial}{\partial \mathbf{a}_k}(\mathbf{a}_k^\top \mathbf{y}_1 - b)^2 = 2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\frac{\partial}{\partial \mathbf{a}_k}(\mathbf{a}_k^\top \mathbf{y}_1 - b) = 2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\mathbf{y}_1.$$
To find the matrix of second partial derivatives, we can take the partial derivative again to see that,
$$\frac{\partial^2}{\partial \mathbf{a}_k^\top \partial \mathbf{a}_k} J_q(\mathbf{a}_k) = \frac{\partial}{\partial \mathbf{a}_k^\top} 2\mathbf{y}_1(\mathbf{a}_k^\top \mathbf{y}_1 - b) = 2\frac{\partial}{\partial \mathbf{a}_k^\top}(\mathbf{y}_1 \mathbf{a}_k^\top \mathbf{y}_1 - b\mathbf{y}_1)$$
$$= 2\mathbf{y}_1 \frac{\partial}{\partial \mathbf{a}_k^\top}(\mathbf{a}_k^\top \mathbf{y}_1) = 2\mathbf{y}_1\mathbf{y}_1^\top = D$$
To find $\mathbf{a}_{k+1} = \mathbf{a}_k - \rho_k \nabla \mathbf{J}(\mathbf{a}_k)$, we can use the formula for ρ_k from the textbook.
$$\rho_k = \frac{\left\|\nabla \mathbf{J}_q(\mathbf{a}_k)\right\|^2}{\nabla \mathbf{J}_q(\mathbf{a}_k)^\top D \nabla \mathbf{J}_q(\mathbf{a}_k)} = \frac{\|2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\mathbf{y}_1\|^2}{[2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\mathbf{y}_1]^\top [2\mathbf{y}_1\mathbf{y}_1^\top][2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\mathbf{y}_1]}$$
$$= \frac{4(\mathbf{a}_k^\top \mathbf{y}_1 - b)^2 \mathbf{y}_1^\top \mathbf{y}_1}{8(\mathbf{a}_k^\top \mathbf{y}_1 - b)^2 \mathbf{y}_1^\top \mathbf{y}_1 \mathbf{y}_1^\top \mathbf{y}_1} = \frac{1}{2\mathbf{y}_1^\top \mathbf{y}_1} = \frac{1}{2\|\mathbf{y}_1\|^2}$$
Then, going back to the update formula we have the following,
$$\mathbf{a}_{k+1} = \mathbf{a}_k - \rho_k \nabla \mathbf{J}(\mathbf{a}_k)$$
$$= \mathbf{a}_k - \frac{\nabla \mathbf{J}(\mathbf{a}_k)}{2\|\mathbf{y}_1\|^2}$$
$$= \mathbf{a}_k - \frac{2(\mathbf{a}_k^\top \mathbf{y}_1 - b)\mathbf{y}_1}{2\|\mathbf{y}_1\|^2}$$

$$= \mathbf{a}_k + \frac{b - \mathbf{a}^\top \mathbf{y}_1}{\|\mathbf{y}_1\|^2} \mathbf{y}_1 \blacksquare$$

5. Show that the partial derivatives of the functions $y_i = \exp(a_i)/\sum_j \exp(a_j)$ used in multiple class logistic discrimination are given by
$$\frac{\partial y_i}{\partial a_j} = y_i(\delta_{ij} - y_j)$$
where $\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & \text{otherwise.} \end{cases}$

<u>Ans:</u> References: [5.1], [5.2]

To solve $\frac{\partial y_i}{\partial a_j}$, we must look at two cases. We must look for when $j = i$ and when $j \neq i$. This will yield a piecewise equation shown below.

$$\frac{\partial y_i}{\partial a_j} = \begin{cases} \dfrac{\exp(a_i)}{\sum_j \exp(a_j)} \dfrac{\sum_j \exp(a_j) - \exp(a_j)}{\sum_j \exp(a_j)} & i = j \\ -\dfrac{\exp(a_i)\exp(a_j)}{\left[\sum_j \exp(a_j)\right]^2} & i \neq j \end{cases} \tag{5.1}$$

Solving $\frac{\partial y_i}{\partial a_j}$ requires the use of the quotient rule, where $f'(x) = \frac{g'(x)h(x) - h'(x)g(x)}{[h(x)]^2}$ when $f(x) = \frac{g(x)}{h(x)}$. In this case, $g(x)$ can be thought of as $\exp(a_i)$ and $h(x)$ can be thought of as $\sum_j \exp(a_j)$. With $\sum_j \exp(a_j)$, the derivative w.r.t. a_k for some arbitrary k is always $\exp(a_k)$. However, looking at $\exp(a_i)$, the derivative w.r.t. a_k for some arbitrary k is only $\exp(a_k)$ when $i = k$.

To prove equation (5.1), we can first look at the case of $i = j$. Solving for $\frac{\partial y_i}{\partial a_j}$ we get

$$\frac{\partial}{\partial a_j}\left(\frac{\exp(a_i)}{\sum_j \exp(a_j)}\right) = \frac{\exp(a_i)\sum_j \exp(a_j) - \exp(a_j)\exp(a_i)}{\left[\sum_j \exp(a_j)\right]^2}$$
$$= \frac{\exp(a_i)}{\sum_j \exp(a_j)} \frac{\sum_j \exp(a_j) - \exp(a_j)}{\sum_j \exp(a_j)} = y_i(1 - y_j) = y_i(1 - y_i)$$

Then in the case of $i \neq j$ we have the following.

$$\frac{\partial}{\partial a_j}\left(\frac{\exp(a_i)}{\sum_j \exp(a_j)}\right) = \frac{0 - \exp(a_j)\exp(a_i)}{\left[\sum_j \exp(a_j)\right]^2}$$
$$= -\frac{\exp(a_j)}{\sum_j \exp(a_j)} \frac{\exp(a_i)}{\sum_j \exp(a_j)} = -y_j y_i$$

Therefore, equation (5.1) leads to the following,

$$\frac{\partial y_i}{\partial a_j} = \begin{cases} y_i(1 - y_i) & i = j \\ -y_i y_j & i \neq j. \end{cases} \tag{5.2}$$

Next, we must define the Kronecker delta to be the following,

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & \text{otherwise.} \end{cases}$$

After combining the Kronecker delta into the equation (5.2) we can get the following,

$$\frac{\partial y_i}{\partial a_j} = y_i(\delta_{ij} - y_j). \blacksquare$$

Reference:

[1.1] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

[1.2] https://math.stackexchange.com/questions/1210545/distance-from-a-point-to-a-hyperplane

[1.3] https://piazza.com/class/kc0jkwru805u1?cid=143

[2.1] https://piazza.com/class/kc0jkwru805u1?cid=147

[2.2] Duda, R. O. (2000). R. O. Duda's P. E. Hart's D. G. Stork's Pattern Classification (Pattern Classification (2nd Edition) [Hardcover])(2000) (2 edition). Wiley-Interscience.

[3.1] https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/07_multiclass.pdf

[3.2] https://math.stackexchange.com/questions/404143/what-is-convex-combination-of-two-points

[3.3] https://en.wikipedia.org/wiki/Convex_combination

[3.4] https://mathworld.wolfram.com/ConvexCombination.html

[3.5] https://piazza.com/class/kc0jkwru805u1?cid=144

[4.1] https://piazza.com/class/kc0jkwru805u1?cid=148

[5.1] https://www.ics.uci.edu/~pjsadows/notes.pdf

[5.2] https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative/