SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Homework #4

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *November 6th, 2019*

Homework Submission Instructions. Please write up your responses to the following problems clearly and concisely. We require you to write up your responses with A4 paper. You are allowed and encouraged to work together. You may discuss the homework to understand the problem and reach a solution in groups. However, *each student must write down the solution independently*. You must understand the solution well enough in order to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

- **Written Homeworks.** All calculation problems **MUST** be written on single-sided A4 paper. You should bring and hand in it before class on the day of the deadline. Submitting the scan or photo version on Sakai will **NOT** be accepted.
- **Coding Homeworks.** All coding assignments will be done in Jupyter Notebooks. We will provide a `.ipynb` template for each assignment. Your final submission will be a `.ipynb` file with your answers and explanations (you should know how to write in Markdown or LaTeX). Make sure that all packages you need are imported at the beginning of the program, and your `.ipynb` file should **work step-by-step without any error**.

Question 1

Show that maximization of the class separation criterion given by $m_2 - m_1 = \mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})$ with respect to $\mathbf{w}$, using a Lagrange multiplier to enforce the constraint $\mathbf{w^T w} = \mathbf{1}$, leads to the result that $\mathbf{w} \propto (\mathbf{m_2} - \mathbf{m_1})$.

Answer. From $m_2 - m_1 = \mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})$ we can construct the Lagrangian function

$$L = \mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1}) + \lambda(\mathbf{w}^T\mathbf{w} - 1).$$

Taking the gradient of L we obtain

$$\nabla L = \mathbf{m_2} - \mathbf{m_1} + 2\lambda\mathbf{w}$$

and setting this gradient to zero gives

$$\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m_2} - \mathbf{m_1})$$

form which it follows that $\mathbf{w} \propto \mathbf{m_2} - \mathbf{m_1}$.

Question 2

Show that the Fisher criterion

$$\mathrm{J}(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

can be written in the form

$$\mathrm{J}(\mathbf{w}) = \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_W w}}.$$

Hint.

$$y = \mathbf{w^T x}, \qquad m_k = \mathbf{w^T m_k}, \qquad s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

Answer. Starting with the numerator on the r.h.s. of $J(\mathbf{w})$,we can rewrite it as follows:

$$\begin{aligned}(m_2 - m_1)^2 &= (\mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1}))^2 \\ &= \mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})(\mathbf{m_2} - \mathbf{m_1})^\mathbf{T}\mathbf{w} \\ &= \mathbf{w^T S_B w}.\end{aligned}$$

Similarly, we can rewrite the denominator of the r.h.s. of $J(\mathbf{w})$:

$$\begin{aligned}s_1^2 + s_2^2 &= \sum_{n \in \mathcal{C}_1} (y_n - m_1)^2 + \sum_{k \in \mathcal{C}_2} (y_k - m_2)^2 \\ &= \sum_{n \in \mathcal{C}_1} (\mathbf{w^T}(\mathbf{x_n} - \mathbf{m_1}))^2 + \sum_{k \in \mathcal{C}_2} (\mathbf{w^T}(\mathbf{x_k} - \mathbf{m_2}))^2 \\ &= \sum_{n \in \mathcal{C}_1} \mathbf{w^T}(\mathbf{x_n} - \mathbf{m_1})(\mathbf{x_n} - \mathbf{m_1})^\mathbf{T}\mathbf{w} + \sum_{k \in \mathcal{C}_2} \mathbf{w^T}(\mathbf{x_k} - \mathbf{m_2})(\mathbf{x_k} - \mathbf{m_2})^\mathbf{T}\mathbf{w} \\ &= \mathbf{w^T S_w w}.\end{aligned}$$

Question 3

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-conditonal dendities $p(\phi|\mathcal{C}_k)$ where ϕ is the input feature vector. Suppose we are given a training data set $\{\phi_n, \mathbf{t}_n\}$ where $n = 1, ..., N$, and $\mathbf{t}_n$ is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $t_{nj} = I_{jk}$ if pattern n is from class $\mathcal{C}_k$. Assuming that the data points are drwn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_k = \frac{N_k}{N},$$

where N_k is the number of data points assigned to class $\mathcal{C}_k$.

Answer. The likelihood function is given by

$$p(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\{p(\phi_n|\mathcal{C}_k)\pi_k\}^{t_{nk}}$$

and taking the logarithm, we obtain

$$\ln p(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\{\ln p(\phi_n|\mathcal{C}_k) + \ln \pi_k\}.$$

In order to maximize the log likelihood with respect to π_k we need to perserve the constaint $\sum_k \pi_k = 1$. This can be done by introducing a Lagrange multiplier λ and maximizing

$$\ln p(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}) + \lambda(\sum_{k=1}^{K} \pi_k - 1).$$

Setting the derivative with respect to π_k equal to zero, we obtain

$$\sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda = 0.$$

Re-arranging then gives

$$-\pi_k \lambda = \sum_{n=1}^{N} t_{nk} = N_k.$$

Summing both sides over k we find that $\lambda = -N$, and using this to eliminate λ we obtain

$$\pi_k = \frac{N_k}{N}.$$

Question 4

Verify the relation

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \sigma(1-\sigma)$$

for the derivative of the logistic sigmoid function defined by

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

Answer. Differentiating $\sigma(a)$ we obtain

$$\begin{aligned}\frac{d\sigma}{da} &= \frac{e^{-a}}{(1=e^{-a})^2}\\ &= \sigma(a)\{\frac{e^{-a}}{1+e^{-a}}\}\\ &= \sigma(a)\{\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\}\\ &= \sigma(a)(1-\sigma(a)).\end{aligned}$$

Question 5

By making use of the result

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \sigma(1-\sigma)$$

for the derivative of the logistic sigmoid, show that the derivative of the error function for the logistic regression model is given by

$$\nabla \mathbb{E}(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n.$$

Hint. The error function for thr logistic regression model is given by

$$\mathbb{E}(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}.$$

Answer. We start by computing the derivative of $E(\mathbf{w})$ w.r.t. y_n

$$\begin{aligned}\frac{\partial E}{\partial y_n} &= \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \\ &= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} \\ &= \frac{t_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \\ &= \frac{y_n - t_n}{y_n(1-y_n)}.\end{aligned}$$

From the result

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \sigma(1-\sigma),$$

we see that

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)(1-\sigma(a_n)) = y_n(1-y_n).$$

Finally we have

$$\nabla a_n = \phi_n$$

where ∇ denotes the gradient with respect to $\mathbf{w}$. Combining these three equations using the chain rule, we obtain

$$\begin{aligned}\nabla E &= \sum_{n=1}^{N} \frac{\partial E}{\partial y_n}\frac{\partial y_n}{\partial a_n}\nabla a_n \\ &= \sum_{n=1}^{N} (y_n - t_n)\phi_n\end{aligned}$$

ae required.

Question 6

There are several possible ways in which to generalize the concept of linear discriminant functions from two classes to c classes. One possibility would be to use $(c-1)$ linear discriminant functions, such that $y_k(\mathbf{x}) > 0$ for inputs $\mathbf{x}$ in class C_k and $y_k(\mathbf{x}) < 0$ for inputs not in class C_k. By drawing a simple example in two dimensions for $c = 3$, show that this approach can lead to regions of x-space for which the classification is ambiguous. Another approach would be to use one discriminant function $y_{jk}(\mathbf{x})$ for each possible pair of classes C_j and C_k , such that $y_{jk}(\mathbf{x}) > 0$ for patterns in class C_j and $y_{jk}(\mathbf{x}) < 0$ for patterns in class C_k. For c classes, we would need $c(c-1)/2$ discriminant functions. Again, by drawing a specific example in two dimensions for $c = 3$, show that this approach can also lead to ambiguous regions.

Answer. .
Part 1: Since there are 3 classes, $c = 3$ and there are 2 discriminant functions $y_1(\mathbf{x})$ and $y2(\mathbf{x})$. For $\mathbf{x} \in C_1, y_1(\mathbf{x}) > 0$ and for $\mathbf{x} \in C_2, y_2(\mathbf{x}) > 0$.This leads to the following problem. How do we classify input patterns $\mathbf{x}$ which have the property that $y_1(\mathbf{x}) > 0$ and $y_2(\mathbf{x}) > 0$. Clearly they belong to both class C_1 and C_2. Figure 1 illustrates the problem. Making the discriminant lines parallel to each other does not resolve the problem since the intersection $y_1(\mathbf{x}) > 0$ and $y_2(\mathbf{x}) > 0$ is non-empty.Note that the intersection is a null set if and only if the two lines coincide which means $y_1(\mathbf{x}) = y_2(\mathbf{x})$.

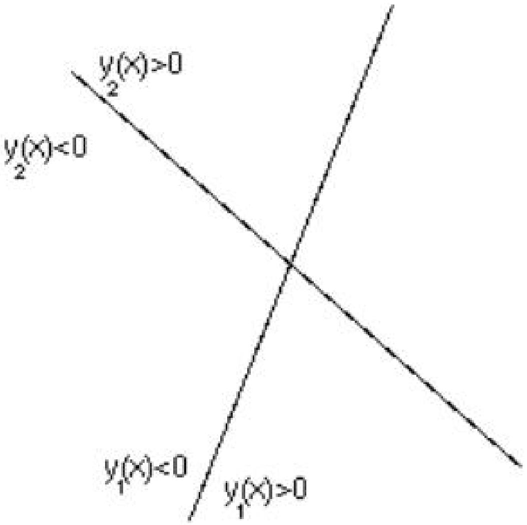

Figure 1: The two dividing linear discriminant boundaries clearly leave a region of space classified into two classes.

Part 2: Since there are 3 classes, $c(c-1)/2 = 3$ and there are three discriminant functions, $y_{12}(\mathbf{x}), y_{23}(\mathbf{x})$ and $y_{23}(x)$. The classification structure is as follows.

1. If $y_{12}(\mathbf{x}) > 0$ and $y_{13}(\mathbf{x}) > 0$, then $\mathbf{x} \in C_1$
2. If $y_{12}(\mathbf{x}) < 0$ and $y_{23}(\mathbf{x}) > 0$, then $\mathbf{x} \in C_2$
3. If $y_{13}(\mathbf{x}) < 0$ and $y_{23}(\mathbf{x}) < 0$, then $\mathbf{x} \in C_3$

This leads to the following problems as illustrated in Figure 2. The following regions are unclassified.

1. $y_{12}(\mathbf{x}) < 0$ and $y_{23}(\mathbf{x}) > 0$
2. $y_{12}(\mathbf{x}) > 0$ and $y_{23}(\mathbf{x}) > 0$ and $y_{13}(\mathbf{x}) < 0$

The intersections are null sets if and only if $y_{12}(\mathbf{x}) = y_{13}(\mathbf{x}) = y_{23}(\mathbf{x})$

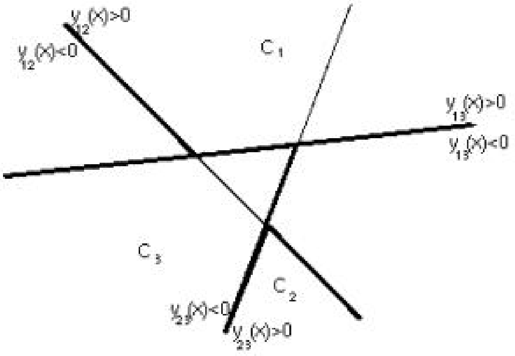

Figure 2: The three dividing linear discriminant boundaries clearly leave a region of space unclassified.

Question 7

Given a set of data points $\{\mathbf{x}^n\}$ we can define the convex hull to be the set of points $\mathbf{x}$ given by

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}^n \tag{1}$$

where $\alpha_n >= 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{\mathbf{z}^m\}$ and its corresponding convex hull. The two sets of points will be linearly separable if there exists a vector $\hat{\mathbf{w}}$ and a scalar w_0 such that $\hat{\mathbf{w}}^T\mathbf{x}^n + w_0 > 0$ for all $\mathbf{x}^n$, and $\hat{\mathbf{w}}^T\mathbf{z}^m + w_0 < 0$ for all $\mathbf{z}^m$. Show that, if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that, if they are linearly separable, their convex hulls do not intersect.

Answer. First, lets calculate the linear discriminants for the points belonging to the two convex hulls. For points in the convex hull of $\{\mathbf{x}^n\}$, the linear discriminant is:

$$y(\mathbf{x}) = \hat{\mathbf{w}}^T\mathbf{x}^n + w_0 \tag{2}$$

Substituting (1) in (2), we get

$$y(\mathbf{x}) = \hat{\mathbf{w}}^T(\sum_n \alpha_n \mathbf{x}^n) + w_0 \tag{3}$$

Since α_n is a scalar quantity, we can bring the summation in (3) outside resulting in

$$\begin{aligned} y(\mathbf{x}) &= \sum_n \alpha_n(\hat{\mathbf{w}}^T\mathbf{x}^n) + w_0 \\ &= \sum_n \alpha_n(\hat{\mathbf{w}}^T\mathbf{x}^n + w_0) \end{aligned} \tag{4}$$

where weve made us of the fact that $\sum_n \alpha_n = 1$. Similarly, we can develop the linear discriminant for the points belonging to the convex hull of $\{\mathbf{z}^m\}$:

$$y(\mathbf{z}) = \sum_m \beta_m(\hat{\mathbf{w}}^T\mathbf{z}^m + w_0) \tag{5}$$

where $\beta_m \geq 0$ and $\sum_m \beta_m = 1$.

Convex hulls intersect: If the convex hulls intersect, there must be at least one point in common between $\{\mathbf{x}\}$ and $\{\mathbf{z}\}$. Lets call that point $\mathbf{xz}$. Since $\mathbf{xz}$ belongs to both convex hulls, there must be a set of $\{\alpha_n\}$ and $\{\beta_m\}$ that give rise to $\mathbf{xz}$. The linear discriminant for $\mathbf{xz}$ can now be written in two separate but equivalent ways. From (4) and (5), we get

$$y(\mathbf{xz}) = \sum_n \alpha_n(\hat{\mathbf{w}}^T\mathbf{x}^n + w_0) = \sum_m \beta_m(\hat{\mathbf{w}}^T\mathbf{z}^m + w_0) \tag{6}$$

For linear separability, we must have

$$\begin{aligned} y(\mathbf{x}^n) &= \hat{\mathbf{w}}^T\mathbf{x}^n + w_0 > 0 \\ y(\mathbf{z}^m) &= \hat{\mathbf{w}}^T\mathbf{z}^m + w_0 < 0 \end{aligned} \tag{7}$$

From the non-negativity and simplex constraints on α and β, (6) and (7), we have a contradiction. The linear discriminant $y(\mathbf{xz})$ has to be simultaneously greater than and less than zero which is impossible.

Patterns are linearly separable: If the patterns are linearly separable, we know that

$$\begin{aligned} y(\mathbf{x}^n) &= \hat{\mathbf{w}}^T\mathbf{x}^n + w_0 > 0 \\ y(\mathbf{z}^m) &= \hat{\mathbf{w}}^T\mathbf{z}^m + w_0 < 0 \end{aligned} \tag{8}$$

Assume that there is a point $\mathbf{xz}$ lying in the intersection of the convex hulls. From (6) above

$$y(\mathbf{xz}) = \sum_n \alpha_n(\hat{\mathbf{w}}^T\mathbf{x}^n + w_0) = \sum_m \beta_m(\hat{\mathbf{w}}^T\mathbf{z}^m + w_0) \tag{9}$$

The equality in (9) is not possible given the fact from (8) that the patterns are linearly separable.

Program Question