Homework IV

12110644 周思呈

Question 1

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

$$
\begin{aligned}
L(\mathbf{w}) = \mathbf{w^T(m_2 - m_1)} - \lambda\mathbf{(w^T w - 1)} \\
\frac{\partial L}{\partial \mathbf{w}} = \mathbf{(m_2 - m_1)} - \mathbf{2}\lambda\mathbf{w} = 0 \\
\Rightarrow \mathbf{w} = \frac{\mathbf{m_1 - m_2}}{2\lambda} \propto \mathbf{(m_2 - m_1)}
\end{aligned} \tag{1}
$$

Question 2

Show that the Fisher criterion

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

can be written in the form

$$
\mathrm{J}(\mathbf{w}) = \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_W w}} \tag{3}
$$

Hint.

$$
y = \mathbf{w^T x}, \tag{4}
$$

$$
m_k = \mathbf{w^T m_k}, \tag{5}
$$

$$
s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2 \tag{6}
$$

Rewrite the formula as

$$
\begin{aligned}
\mathrm{J}(\mathbf{w}) &= \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \\
&= \frac{\mathbf{(w^T m_2 - w^T m_1)^2}}{\sum_{\mathbf{n \in C_1}} \mathbf{(y_n - m_1)^2} + \sum_{\mathbf{n \in C_2}} \mathbf{(y_n - m_2)^2}}
\end{aligned} \tag{7}
$$

We define

$$
\begin{aligned}
\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_2)^T \\
\boldsymbol{S}_W = \sum_{n \in \mathcal{C}_1} (\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^T + \sum_{n \in \mathcal{C}_2} (\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^T
\end{aligned} \tag{8}
$$

So

$$\mathrm{J}(\mathbf{w}) = \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_W w}} \tag{9}$$

Question 3

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-conditional densities $p(\phi|\mathcal{C}_k)$ where ϕ is the input feature vector. Suppose we are given a training data set { $\phi_n, \mathbf{t}_n$ } where $n = 1, \ldots, N$, 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 drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_k = \frac{N_k}{N}, \tag{10}$$

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

$$\begin{aligned} p\left(\{\phi_{\mathbf{n}}, t_n\} \mid \pi_1, \pi_2, \ldots, \pi_K\right) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \left[p\left(\boldsymbol{\phi_n} \mid C_k\right) p\left(C_k\right)\right]^{t_{nk}} \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k p\left(\boldsymbol{\phi_n} \mid C_k\right)\right]^{t_{nk}} \end{aligned} \tag{11}$$

$$\ln p = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left[\ln \pi_k + \ln p\left(\phi_n \mid C_k\right)\right] \propto \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln \pi_k \tag{12}$$

Add a Lagrange Multiplier to the expression

$$\begin{gathered} L = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1\right) \\ \frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda \\ \Rightarrow \pi_k = -\left(\sum_{n=1}^{N} t_{nk}\right) / \lambda = -\frac{N_k}{\lambda} \end{gathered} \tag{13}$$

Because $1 = -\left(\sum_{k=1}^{K} N_k\right) / \lambda = -\frac{N}{\lambda}$, we can obtain $\pi_k = \frac{N_k}{N}$.

Question 4

Verify the relation

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \sigma(1-\sigma) \tag{14}$$

for the derivative of the logistic sigmoid function defined by

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{15}$$

$$\sigma(a) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1-(e^x+1)^{-1} \tag{16}$$

$$\begin{aligned} \frac{\mathrm{d}\sigma}{\mathrm{d}a} &= (-1)(-1)(e^x+1)^{-2}e^x \\ &= \left(1+e^{-x}\right)^{-2}e^{-2x}e^x \\ &= \left(1+e^{-x}\right)^{-1} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(1-\sigma) \end{aligned} \tag{17}$$

Question 5

By making use of the result

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \sigma(1-\sigma) \tag{18}$$

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

$$\nabla\mathbb{E}(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n. \tag{19}$$

Hint.

The error function for the logistic regression model is given by

$$\mathbb{E}(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}. \tag{20}$$

Define $y_n = \sigma\left(a_n\right), a_n = \mathbf{w}^T\boldsymbol{\phi}_n$, we have

$$
\begin{aligned}
\nabla E(\boldsymbol{w}) &= -\nabla \sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \\
&= -\sum_{n=1}^{N} \nabla \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \\
&= -\sum_{n=1}^{N} \frac{d\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}}{dy_n} \frac{dy_n}{da_n} \frac{da_n}{d\mathbf{w}} \\
&= -\sum_{n=1}^{N} \left(\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right) \cdot y_n(1-y_n) \cdot \boldsymbol{\phi}_n \\
&= -\sum_{n=1}^{N} \frac{t_n - y_n}{y_n(1-y_n)} \cdot y_n(1-y_n) \cdot \boldsymbol{\phi}_n \\
&= -\sum_{n=1}^{N} (t_n - y_n)\boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n
\end{aligned} \tag{21}
$$

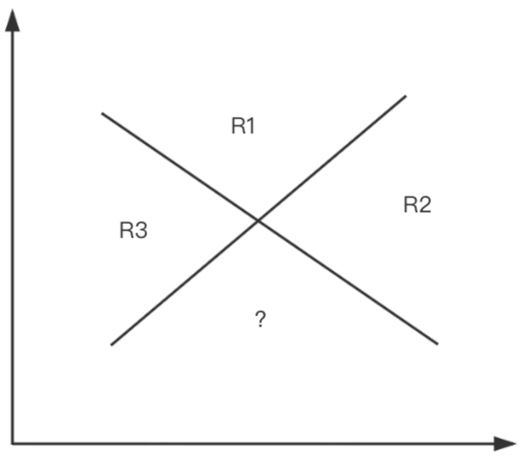
Question 6

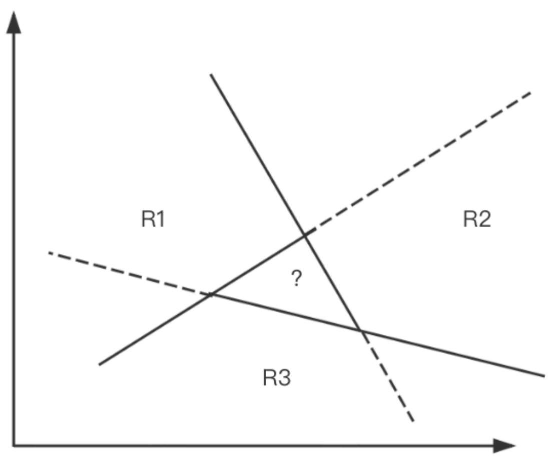
There are several possible ways in which to generalize the concept of linear discriminant functions from two classes to c classes. One possibility would be to use ($c-1$) linear discriminant functions, such that $y_k(\mathbf{x}) > 0$ for inputs $\mathbf{x}$ in class C_k and $y_k(\mathbf{x}) < 0$ for 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.





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

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}^n \tag{22}$$

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.

If the convex hull of $\{\mathbf{x_n}\}$ and $\{\mathbf{z_m}\}$ intersects, we know that there will be a point $\mathbf{y}$ which can be written as $\mathbf{y} = \sum_n \alpha_n \mathbf{x_n}$ and also $\mathbf{y} = \sum_m \beta_m \mathbf{z_m}$. Because $\sum_n \alpha_n = 1$ we can obtain:

$$\begin{aligned} \widehat{\mathbf{w}}^T \mathbf{y} + w_0 &= \widehat{\mathbf{w}}^T \left(\sum_n \alpha_n \mathbf{x_n} \right) + w_0 \\ &= \left(\sum_n \alpha_n \widehat{\mathbf{w}}^T \mathbf{x_n} \right) + \left(\sum_n \alpha_n \right) w_0 \\ &= \sum_n \alpha_n \left(\widehat{\mathbf{w}}^T \mathbf{x_n} + w_0 \right) \end{aligned} \tag{23}$$

Similarly we have

$$\begin{aligned} \widehat{\mathbf{w}}^T \mathbf{y} + w_0 &= \widehat{\mathbf{w}}^T \left(\sum_m \alpha_m \mathbf{z_m} \right) + w_0 \\ &= \left(\sum_m \alpha_m \widehat{\mathbf{w}}^T \mathbf{z_m} \right) + \left(\sum_m \alpha_m \right) w_0 \\ &= \sum_m \alpha_m \left(\widehat{\mathbf{w}}^T \mathbf{z_m} + w_0 \right) \end{aligned} \tag{24}$$

If $\{\mathbf{x_n}\}$ and $\{\mathbf{z_m}\}$ are linearly separable, for $\forall \mathbf{x_n}, \mathbf{z_m}$ we have

$$\begin{aligned} \widehat{\mathbf{w}}^T \mathbf{x_n} + w_0 > 0 \\ \widehat{\mathbf{w}}^T \mathbf{z_m} + w_0 < 0 \end{aligned} \tag{25}$$

which leads to the contradiction.