HW4

12011923 张旭东

Q1：

To show that

$$\mathbf{w} \propto \mathbf{m_2} - \mathbf{m_1} \tag{1}$$

Starting with the class separation criterion

$$m_2 - m_1 = \mathbf{w}^{\mathbf{T}}(\mathbf{m_2} - \mathbf{m_1}) \tag{2}$$

With the constraint $\mathbf{w}^{\mathbf{T}}\mathbf{w} = \mathbf{1}$, the Lagrangian for this optimization problem is given by

$$L(\mathbf{w}, \lambda) = \mathbf{w}^{\mathbf{T}}(\mathbf{m_2} - \mathbf{m_1}) - \lambda(\mathbf{w}^{\mathbf{T}}\mathbf{w} - \mathbf{1}) \tag{3}$$

Taking the derivative of L with respect to $\mathbf{w}$ and setting it to zero, we can get:

$$\frac{\partial L}{\partial \mathbf{w}} = (\mathbf{m_2} - \mathbf{m_1}) - 2\lambda\mathbf{w} = 0 \tag{4}$$

Then we can get:

$$\mathbf{w} = \frac{\mathbf{m_2} - \mathbf{m_1}}{2\lambda} \tag{5}$$

The constraint $\mathbf{w}^{\mathbf{T}}\mathbf{w} = \mathbf{1}$ implies that

$$(\frac{\mathbf{m_2} - \mathbf{m_1}}{2\lambda})^{\mathbf{T}}(\frac{\mathbf{m_2} - \mathbf{m_1}}{2\lambda}) = 1 \tag{6}$$

$$(\mathbf{m_2} - \mathbf{m_1})^{\mathbf{T}}(\mathbf{m_2} - \mathbf{m_1}) = 4\lambda^2 \tag{7}$$

Then we can get the expression of λ:

$$\lambda = \frac{\pm\sqrt{(\mathbf{m_2} - \mathbf{m_1})^{\mathbf{T}}(\mathbf{m_2} - \mathbf{m_1})}}{2} \tag{8}$$

Substituting λ back in to the expression for $\mathbf{w}$:

$$\mathbf{w}=\frac{\mathbf{m_2}-\mathbf{m_1}}{\pm\sqrt{(\mathbf{m_2}-\mathbf{m_1})^{\mathbf{T}}(\mathbf{m_2}-\mathbf{m_1})}} \tag{9}$$

The $\pm$ sign indicates that the direction of $\mathbf{w}$ is either in the same or opposite direction as $\mathbf{m_2}-\mathbf{m_1}$. Hence, $\mathbf{w}$ is proportional to $\mathbf{m_2}-\mathbf{m_1}$, showing that the maximization of the class separation criterion leads to the result $\mathbf{w}\propto\mathbf{m_2}-\mathbf{m_1}$

Q2:

The Fisher criterion is

$$J(\mathbf{w})=\frac{(m_2-m_1)^2}{s_1^2+s_2^2} \tag{10}$$

Because $m_k=\mathbf{w^T m_k}$, we can get:

$$(m_2-m_1)^2=(\mathbf{w^T m_2}-\mathbf{w^T m_1})^2=\mathbf{w^T}(\mathbf{m_2}-\mathbf{m_1})(\mathbf{m_2}-\mathbf{m_1})^{\mathbf{T}}\mathbf{w} \tag{11}$$

The between-class matrix is defined as :

$$\mathbf{S_B}=(\mathbf{m_2}-\mathbf{m_1})(\mathbf{m_2}-\mathbf{m_1})^{\mathbf{T}} \tag{12}$$

So

$$(m_2-m_1)^2=\mathbf{w^T S_B w} \tag{13}$$

Because $m_k=\mathbf{w^T m_k}$, $s_k^2=\sum_{n\in C_k}(y_n-m_k)^2$ and $y=\mathbf{w^T x}$, we can get:

$$\begin{aligned} s_1^2+s_2^2&=\sum_{n\in C_1}(y_n-m_1)^2+\sum_{n\in C_2}(y_n-m_2)^2\\ &=\sum_{n\in C_1}(\mathbf{w^T x_1}-\mathbf{w^T m_1})^2+\sum_{n\in C_2}(\mathbf{w^T x_2}-\mathbf{w^T m_2})^2\\ &=\mathbf{w^T}(\sum_{n\in C_1}(\mathbf{x_1}-\mathbf{m_1})^2+\sum_{n\in C_2}(\mathbf{x_2}-\mathbf{m_2})^2)\mathbf{w} \end{aligned} \tag{14}$$

The within-class scatter matrix is defined as:

$$\mathbf{S_W}=\sum_{n\in C_1}(\mathbf{x_1}-\mathbf{m_1})^2+\sum_{n\in C_2}(\mathbf{x_2}-\mathbf{m_2})^2 \tag{15}$$

So

$$s_1^2 + s_2^2 = \mathbf{w^T S_W w} \tag{16}$$

So, the Fisher criterion can be written in the form:

$$J(\mathbf{w}) = \frac{\mathbf{w^T S_B w}}{\mathbf{w^T S_W w}} \tag{17}$$

Q3:

The likelihood function for the dataset:

$$L = p(t_n|\phi_n, C_k) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k)^{t_{nk}} \cdot p(\phi_n|C_k)^{t_{nk}} \tag{18}$$

Take the natural logarithm of the likelihood function, we can get the log-likelihood function:

$$\ln p(t_n|\phi_n, C_k) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}[\ln \pi_k + \ln p(\phi_n|C_k)] \tag{19}$$

In order to maximize the log-likelihood, we need to preserve the fact that $\sum_{k=1}^{K} \pi_k$=1, so we introduce a Lagrange multiplier as follows:

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}[\ln \pi_k + \ln p(\phi_n|C_k)] - \lambda(\sum_{k=1}^{K} \pi_k - 1) \tag{20}$$

To obtain the maximum likelihood, differentiate the log-likelihood with respect to π_k and set the derivative to 0 :

$$\sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} - \lambda = 0 \tag{21}$$

Therefore,

$$\sum_{n=1}^{N} t_{nk} = \pi_k N + \lambda \tag{22}$$

$$\pi_k = \frac{N_k}{N} \tag{23}$$

Therefore, 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 C_k.

Q4:

$$\begin{gathered}
\frac{d\sigma}{da} = \frac{d(\frac{1}{1+\exp(-a)})}{da} \\
= -\frac{1}{(1+\exp(-a))^2}(0-\exp(-a)) \\
= \frac{\exp(-a)}{(1+\exp(-a))^2} \\
= \frac{1}{1+\exp(-a)}\frac{\exp(-a)}{1+\exp(-a))} \\
= \frac{1}{1+\exp(-a)}\frac{1+\exp(-a)-1}{1+\exp(-a))} \\
\frac{1}{1+\exp(-a)}(1-\frac{1}{1+\exp(-a))}) \\
= \sigma(1-\sigma)
\end{gathered} \tag{24}$$

Q5:

The error function for the logistic regression model is given by

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}[t_N \ln y_n + (1-t_n)\ln(1-y_n)] \tag{25}$$

where y_n is the model's output and t_n is the target. The output can be written as:

$$y_n = \sigma(\mathbf{w}^{\mathbf{T}}\phi_n) \tag{26}$$

where σ is the logistic sigmoid function and ϕ_n is the n_{th} input vector.

Taking the derivative of the error function with respect to the weights $\mathbf{w}$, we can get:

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial E(\mathbf{w})}{\partial y_n} \frac{\partial y_n}{\partial (\mathbf{w}^{\mathbf{T}} \phi_n)} \frac{\partial (\mathbf{w}^{\mathbf{T}} \phi_n)}{\partial \mathbf{w}} \tag{27}$$

$$\frac{\partial E(\mathbf{w})}{\partial y_n} = -\sum_{n=1}^{N} (\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}) \tag{28}$$

Using the result $\frac{d\sigma}{da} = \sigma(1-\sigma)$ for the derivative of logistic sigmoid,

$$\frac{\partial y_n}{\partial (\mathbf{w}^{\mathbf{T}} \phi_n)} = y_n(1-y_n) \tag{29}$$

And the last component is :

$$\frac{\partial (\mathbf{w}^{\mathbf{T}} \phi_n)}{\partial \mathbf{w}} = \phi_n \tag{30}$$

So

$$\begin{aligned} \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} &= -\sum_{n=1}^{N} (\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}) y_n(1-y_n)\phi_n \\ &= \sum_{n=1}^{N} (y_n - t_n)\phi_n \end{aligned} \tag{31}$$

In conclusion, the derivative of the error function for the logistic regression model is given by

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\phi_n \tag{32}$$

Q6:

1. Using $(c-1)$ Linear Discriminant Functions

   Consider three classes C_1, C_2, C_3 in two dimensions. Using two discriminant functions $y_1(x), y_2(x)$:

   - $y_1(x) > 0$ for $C_1, y_1(x) < 0$ for not in C_1
   - $y_2(x) > 0$ for $C_2, y_2(x) < 0$ for not in C_2

   Ambiguity:

In regions where $y_1(x) < 0$ and $y_2(x) < 0$, the classification is ambiguous, as it's unclear whether it belongs to class C_1, C_2, or C_3.

2. Using $c(c-1)/2$ Discriminant Functions

   Consider three classes C_1, C_2, C_3 in two dimensions. Using two discriminant functions $y_{12}(x), y_{13}(x), y_{23}(x)$:

   - $y_{12}(x) > 0$ for C_1, $y_{12}(x) < 0$ for C_2
   - $y_{13}(x) > 0$ for C_1, $y_{13}(x) < 0$ for C_3
   - $y_{23}(x) > 0$ for C_2, $y_{23}(x) < 0$ for C_3

   Ambiguity:

   In regions where $y_{12}(x) < 0, y_{13}(x) < 0$ and $y_{23}(x) < 0$, the classification is ambiguous, leading to uncertainty in assigning the input to one of the three classes.

Q7:

1.Suppose the convex hull of two sets of points, $\{\mathbf{x}^n\}$ and $\{\mathbf{z}^m\}$, intersect. This means that there exists some points $\{\mathbf{p}\}$ can be written as a convex combination of both sets of points:

$$p = \sum_n \alpha^n \mathbf{x}^n = \sum_m \beta^m \mathbf{z}^m \tag{33}$$

where

$$\alpha^n >= 0, \sum_n \alpha^n = 1 \tag{34}$$

$$\beta^m >= 0, \sum_m \beta^m = 1 \tag{35}$$

Assume that the two sets are linearly separable, that means 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 $\mathbf{z}^m$. Then, for points $\{\mathbf{p}\}$,

$$\hat{\mathbf{w}}^T\mathbf{p} + w_0 = \hat{\mathbf{w}}^T(\sum_n \alpha^n \mathbf{x}^n) + w_0 = \sum_n \alpha^n \hat{\mathbf{w}}^T \mathbf{x}^n + w_0 > 0 \tag{36}$$

and

$$\hat{\mathbf{w}}^T\mathbf{p} + w_0 = \hat{\mathbf{w}}^T(\sum_m \beta^m \mathbf{z}^m) + w_0 = \sum_n \beta^m \hat{\mathbf{w}}^T \mathbf{z}^m + w_0 < 0 \tag{37}$$

But this is a contradiction, and thus the two sets of points can't be linearly separable.

2.Conversely, suppose the convex hull of two sets of points, {$\mathbf{x}^n$} and {$\mathbf{z}^m$}, are linearly separable. Then as discussed, 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 $\mathbf{z}^m$. If the convex hull intersect, there exists some points, {$\mathbf{p}$} in the intersection such that {$\mathbf{p}$} can be written as a convex combination of both sets. But as the above derived, this leads to a contradiction, thus the convex hulls can't be intersect.

In conclusion, 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.