CS329 Homework #4

Course: Machine Learning(H)(CS329) - Instructor: Qi Hao

Name: Jianan Xie(谢嘉楠)

SID: 12110714

Question 1

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

Ans:

Using Lagrange multiplier, we need to maximize $\mathrm{L}(\lambda, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}(\mathbf{m_2} - \mathbf{m_1}) + \lambda(\mathbf{w}^{\mathrm{T}}\mathbf{w} - \mathbf{1})$.

Then we get the derivatives:

$$\frac{\partial \mathrm{L}(\lambda, \mathbf{w})}{\partial \lambda} = \mathbf{w}^{\mathrm{T}}\mathbf{w} - \mathbf{1} = 0$$
$$\frac{\partial \mathrm{L}(\lambda, \mathbf{w})}{\partial \mathbf{w}} = \mathbf{m_2} - \mathbf{m_1} + 2\lambda\mathbf{w} = 0$$

Then we derive that $\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m_2} - \mathbf{m_1})$, thus $\mathbf{w} \propto (\mathbf{m_2} - \mathbf{m_1})$.

Question 2

Show that the Fisher criterion

$$\mathrm{J}(\mathbf{w}) = \frac{(\mathrm{m}_2 - \mathrm{m}_1)^2}{\mathrm{s}_1^2 + \mathrm{s}_2^2}$$

can be written in the form

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

> **Hint.**
>
> $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}, \quad m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m_k}, \quad s_k^2 = \sum_{n \in C_k}(y_n - m_k)^2$

Ans:

We define some measures of the scatter as following:

- The scatter in feature space-x: $S_k = \sum_{n \in C_k}(x_n - m_k)(x_n - m_k)^{\mathrm{T}}$
- Within-class scatter matrix: $S_W = S_1 + S_2$
- Between-class scatter matrix: $S_B = (\mathbf{m_2} - \mathbf{m_1})(\mathbf{m_2} - \mathbf{m_1})^{\mathrm{T}}$

By hints:

$$
\begin{aligned}
\mathrm{J}(\mathbf{w}) &= \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \\
&= \frac{(\mathbf{w}^{\mathrm{T}}\mathbf{m_2} - \mathbf{w}^{\mathrm{T}}\mathbf{m_1})^2}{\sum_{n \in C_1}(\mathbf{w}^{\mathrm{T}}\mathbf{x_n} - \mathbf{w}^{\mathrm{T}}\mathbf{m_1})^2 + \sum_{n \in C_2}(\mathbf{w}^{\mathrm{T}}\mathbf{x_n} - \mathbf{w}^{\mathrm{T}}\mathbf{m_2})^2} \\
&= \frac{[\mathbf{w}^{\mathrm{T}}(\mathbf{m_2} - \mathbf{m_1})]^{\mathrm{T}}[\mathbf{w}^{\mathrm{T}}(\mathbf{m_2} - \mathbf{m_1})]}{\sum_{n \in C_1}[\mathbf{w}^{\mathrm{T}}(\mathbf{x_n} - \mathbf{m_1})]^{\mathrm{T}}[\mathbf{w}^{\mathrm{T}}(\mathbf{x_n} - \mathbf{m_1})] + \sum_{n \in C_2}[\mathbf{w}^{\mathrm{T}}(\mathbf{x_n} - \mathbf{m_2})]^{\mathrm{T}}[\mathbf{w}^{\mathrm{T}}(\mathbf{x_n} - \mathbf{m_2})]} \\
&= \frac{[(\mathbf{m_2} - \mathbf{m_1})^{\mathrm{T}}\mathbf{w}]^{\mathrm{T}}[(\mathbf{m_2} - \mathbf{m_1})^{\mathrm{T}}\mathbf{w}]}{\sum_{n \in C_1}[(\mathbf{x_n} - \mathbf{m_1})^{\mathrm{T}}\mathbf{w}]^{\mathrm{T}}[(\mathbf{x_n} - \mathbf{m_1})^{\mathrm{T}}\mathbf{w}] + \sum_{n \in C_2}[(\mathbf{x_n} - \mathbf{m_2})^{\mathrm{T}}\mathbf{w}]^{\mathrm{T}}[(\mathbf{x_n} - \mathbf{m_2})^{\mathrm{T}}\mathbf{w}]} \\
&= \frac{\mathbf{w}^{\mathrm{T}}(\mathbf{m_2} - \mathbf{m_1})(\mathbf{m_2} - \mathbf{m_1})^{\mathrm{T}}\mathbf{w}}{\sum_{n \in C_1}\mathbf{w}^{\mathrm{T}}(\mathbf{x_n} - \mathbf{m_1})(\mathbf{x_n} - \mathbf{m_1})^{\mathrm{T}}\mathbf{w} + \sum_{n \in C_2}\mathbf{w}^{\mathrm{T}}(\mathbf{x_n} - \mathbf{m_2})(\mathbf{x_n} - \mathbf{m_2})^{\mathrm{T}}\mathbf{w}} \\
&= \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S_B}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S_1}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{S_2}\mathbf{w}} \\
&= \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S_B}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S_W}\mathbf{w}}
\end{aligned}
$$

Question 3

Consider a generative classification model for *K* classes defined by prior class probabilities $p(C_k) = \pi_k$ and general class-conditonal dendities $p(\phi|C_k)$ 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 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 C_k.

Ans:

$p(\phi, C_k) = p(C_k)p(\phi|C_k) = \pi_k p(\phi|C_k)$, so the likelihood function is:

$$p(\{\phi_n, \mathbf{t_n}\}|\pi_1, \pi_2, .., \pi_K) = \prod_{n=1}^{N}\prod_{k=1}^{K}[\pi_k p(\phi_n|C_k)]^{t_{nk}}$$

Then take the log-likelihood:

$$\ln p(\{\phi_n, \mathbf{t_n}\}|\pi_1, \pi_2, .., \pi_K) = \sum_{k=1}^{K}\sum_{n=1}^{N}[t_{nk}\ln\pi_k + t_{nk}\ln p(\phi_n|C_k)]$$

As $\sum_{k=1}^{K} \pi_k = 1$, we use the Lagrange Multiplier to maximize

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

Then we get the derivatives:

$$\frac{\partial L(\pi_k, \lambda)}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0$$
$$\frac{\partial L(\pi_k, \lambda)}{\partial \pi_k} = \frac{\sum_{n=1}^{N} t_{nk}}{\pi_k} + \lambda = 0$$

We derive that $\pi_k = -\frac{\sum_{n=1}^{N} t_{nk}}{\lambda} = -\frac{N_k}{\lambda}$, then we need to get the value of variable λ. By doing sum according k on both sides, we get $\sum_{k=1}^{K} \pi_k = -\sum_{k=1}^{K} \frac{N_k}{\lambda}$. Thus,
$\lambda = -\frac{\sum_{k=1}^{K} N_k}{\sum_{k=1}^{K} \pi_k} = -\frac{N}{1} = -N$.

Finally, we get the MLE of $\pi_k = -\frac{N_k}{\lambda} = \frac{N_k}{N}$.

Question 4

Verify the relation

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

for the derivative of the logistic sigmoid function defined by

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

Ans:

Some basic rules of derivates:

- $\frac{d}{dx} \frac{1}{f(x)} = -\frac{\frac{df(x)}{dx}}{f(x)^2}$ (*)
- $\frac{d}{dx} \exp(ax) = a \exp(ax)$ (**)

So, the verification is shown below:

$$\begin{aligned} \frac{d\sigma}{da} &= \frac{d}{da} \frac{1}{1 + \exp(-a)} \\ &= -\frac{\frac{d}{da}[1 + \exp(-a)]}{[1 + \exp(-a)]^2} && (*) \\ &= \frac{\exp(-a)}{[1 + \exp(-a)]^2} && (**) \\ &= \sigma(a)[1 - \sigma(a)] \end{aligned}$$

The proof is done.

Question 5

By making use of the result

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

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 the logistic regression model is given by**
>
> $\mathbb{E}(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$.

Ans:

By making use of the result $\frac{d\sigma}{da} = \sigma(1-\sigma)$, we know that:

$$\frac{da_n}{d\mathbf{w}} = \frac{d}{d\mathbf{w}}\mathbf{w}^{\mathrm{T}}\phi_{\mathrm{n}} = \phi_{\mathrm{n}}$$
$$\frac{dy_n}{da_n} = \frac{d}{da_n}\sigma(a_n) = \sigma(a_n)(1-\sigma(a_n)) = y_n(1-y_n)$$

Therefore the derivative of the error function for the logistic regression model is:

$$\begin{aligned}
\nabla \mathbb{E}(\mathbf{w}) &= \frac{d}{d\mathbf{w}}\{-\sum_{n=1}^{N}[t_n \ln y_n + (1-t_n)\ln(1-y_n)]\} \\
&= -\sum_{n=1}^{N}\{\frac{d}{dy_n}[t_n \ln y_n + (1-t_n)\ln(1-y_n)]\frac{dy_n}{da_n}\frac{da_n}{d\mathbf{w}}\} \\
&= -\sum_{n=1}^{N}(\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n})y_n(1-y_n)\phi_n \\
&= \sum_{n=1}^{N}\frac{y_n - t_n}{y_n(1-y_n)}y_n(1-y_n)\phi_n \\
&= \sum_{n=1}^{N}(y_n - t_n)\phi_n
\end{aligned}$$

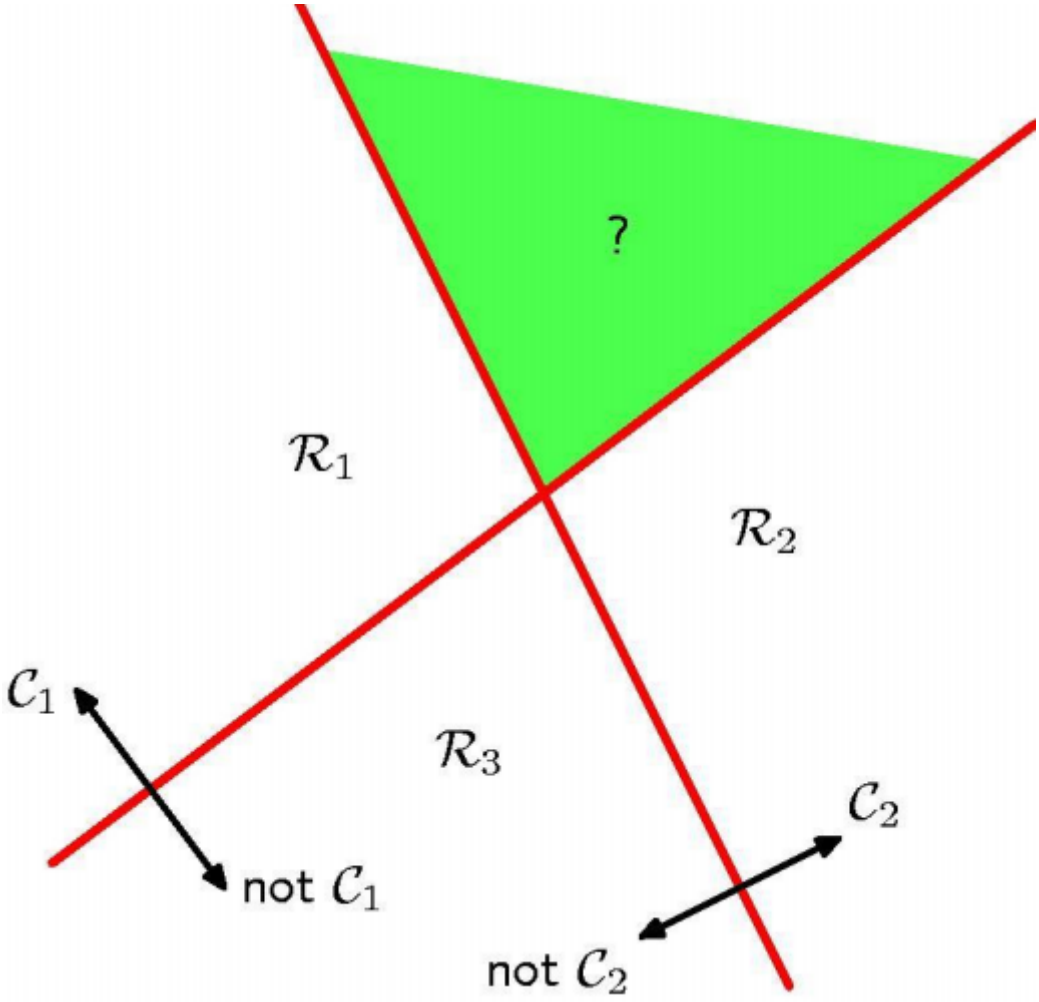
The proof is done.

Question 6

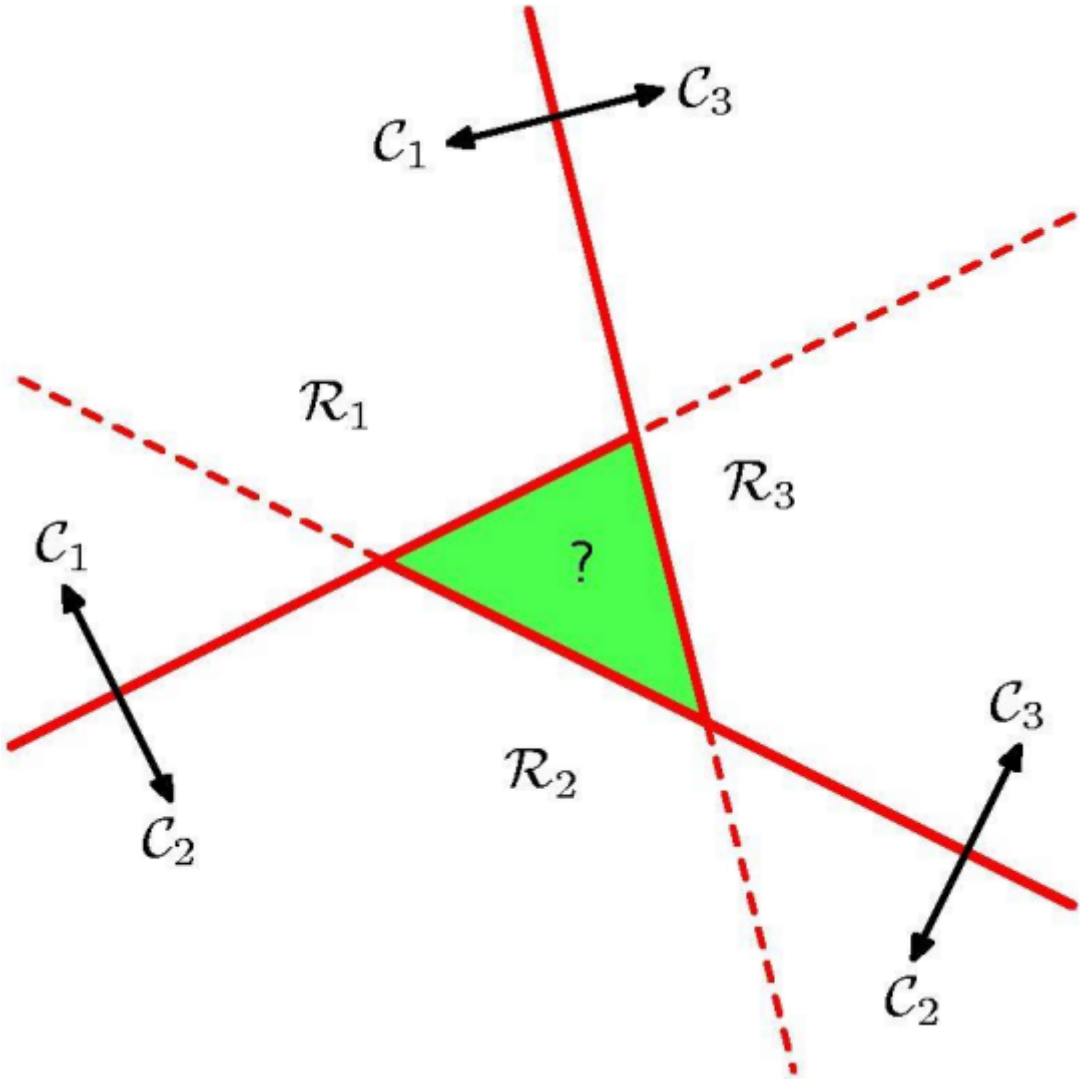
There are several possible ways in which to generalize the concept of linear discriminant functions from two classes to *c* classes. One possibility would be to use (*c* − 1) linear discriminant functions, such that $y_k(\mathbf{x}) > 0$ for inputs **x** in class C_k and $y_k(\mathbf{x}) < 0$ for inputs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

Ans:

(1) For c=3, if we use (c-1) linear discriminant functions to tell C_1, C_2, C_3 apart through the way $y_k(\mathbf{x}) > 0$ for inputs **x** in class C_k and $y_k(\mathbf{x}) < 0$ for inputs not in class C_k, we will find the problem that we cannot tell which class the data points belong, which satisfy $y_1(\mathbf{x}) > 0, y_2(\mathbf{x}) > 0$. The intuitive graphical representation is below:



(2) For c=3, if we use c(c-1)/2 discriminant functions to tell C_1, C_2, C_3 apart through the way $y_{jk}(\mathbf{x})$ 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, we will find the problem that we cannot tell which class the data points belong, which satisfy $y_{12}(\mathbf{x}) < 0, y_{23}(\mathbf{x}) < 0, y_{31}(\mathbf{x}) < 0$. The intuitive graphical representation is below:



Question7

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

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

where $\alpha \geq 0$ 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 $\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}^n + w_0 > 0$ for all $\mathbf{x}^n$, and $\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{z}^m + w_0 < 0$ 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.

Ans:

(Tips: the superscript in question represents id instead of power exponent)

If their convex hulls intersect, that means we can find that
$\exists y, \quad s.t.\ \mathbf{y} = \sum_n \alpha_n \mathbf{x}^n = \sum_m \beta_m z^m$, where $\alpha, \beta \geq 0$ and $\sum_n \alpha_n = 1, \sum_m \beta_m = 1$.
Through proof of contradiction, we suppose the two sets of points can be linearly separable, which means there exists a vector $\hat{\mathbf{w}}$ and a scalar w_0 such that $\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}^n + w_0 > 0$ for all $\mathbf{x}^n$, and $\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{z}^m + w_0 < 0$ for all $\mathbf{z}^m$. We figure that $\sum_n \hat{\mathbf{w}}^{\mathrm{T}}\alpha_n\mathbf{x}^n + \sum_n \alpha_n w_0 > 0$, and

$\sum_m \hat{\mathbf{w}}^{\mathrm{T}} \beta_m \mathbf{z}^m + \sum_m \beta_m w_0 < 0$. We get that $\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{y} + w_0 > 0$, and $\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{y} + w_0 < 0$, which is a contradiction. Thus, the two sets of points cannot be linearly separable.

If they are linearly separable, that means there exists a vector $\hat{\mathbf{w}}$ and a scalar w_0 such that $\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{x}^n + w_0 > 0$ for all $\mathbf{x}^n$, and $\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{z}^m + w_0 < 0$ for all $\mathbf{z}^m$. Through proof of contradiction, we suppose their convex hulls intersect, which means $\exists y, \quad s.t. \ \mathbf{y} = \sum_n \alpha_n \mathbf{x}^n = \sum_m \beta_m z^m$, where $\alpha, \beta \geq 0$ and $\sum_n \alpha_n = 1, \sum_m \beta_m = 1$. Then we figure that
$\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{y} + w_0 = \sum_n \hat{\mathbf{w}}^{\mathrm{T}} \alpha_n \mathbf{x}^n + \sum_n \alpha_n w_0 > 0$, and
$\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{y} + w_0 = \sum_m \hat{\mathbf{w}}^{\mathrm{T}} \beta_m \mathbf{z}^m + \sum_m \beta_m w_0 < 0$, which is a contradiction. Thus, their convex hulls do not intersect.