CS329 Machine Learning

Homework #4

Fan Site
fanst2021@mail.sustech.edu.cn

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 $\mathbf{w}$, using a Lagrange multiplier to enforce the constraint $\mathbf{w}^{\mathrm{T}}\mathbf{w} = \mathbf{1}$, leads to the result that $\mathbf{w} \propto \mathbf{m_2} - \mathbf{m_1}$.

Solution

Construct the Lagrange function and let its gradient be 0,

$$L = \mathbf{w}^{\mathrm{T}(\mathbf{m}_2 - \mathbf{m}_1)} + \lambda(\mathbf{w}^{\mathrm{T}}\mathbf{w} - 1)$$

$$\nabla L = \mathbf{m}_2 - \mathbf{m}_1 + 2\lambda\ \mathbf{w}$$

$$\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m}_2 - \mathbf{m}_1)$$

Therefore $\mathbf{w} \propto \mathbf{m_2} - \mathbf{m_1}$

Question 2

Show that the Fisher criterion

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

can be written in the form

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$$

Hint.

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

$$m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k$$

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

Solution

Expand the Fisher criterion using the hint,

$$J(\mathbf{w}) = \frac{\|\mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)\|^2}{\sum_{n \in C_1} (\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - m_1)^2 + \sum_{n \in C_2} (\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - m_2)^2}$$

And we have that

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}$$

Substitute back to the equation above,

$$
\begin{aligned}
J(\mathbf{w}) &= \frac{\mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}} \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}}\mathbf{w} + \mathbf{w}^{\mathrm{T}} \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}\mathbf{w}} \\
&= \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}
\end{aligned}
$$

Question 3

Consider a generative classification model for K classes defined by prior class probabilities $p(\mathcal{C}_k) = \pi_k$ and general class-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 $n = 1, ..., N$, 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

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

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

Solution

The log-likelood function

$$
\begin{aligned}
\ln p(\{\phi_n, t_n\}|\pi_1, ..., \pi_K) &= \ln \prod_{n=1}^{N} \prod_{k=1}^{K} [p(\phi_n|\mathcal{C}_k)p(\mathcal{C}_k)]^{t_{nk}} \\
&= \ln \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k p(\phi_n|\mathcal{C}_k)]^{t_{nk}} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}[\ln \pi_k + \ln p(\phi_n|\mathcal{C}_k)]
\end{aligned}
$$

Use Lagrange's method,

$$
\begin{aligned}
L &= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln \pi_k + \lambda \left(1 - \sum_{k=1}^{K} \pi_k \right) \\
\frac{\partial L}{\partial \pi_k} &= \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} - \lambda = 0 \\
\pi_k &= \sum_{n=1}^{N} \frac{t_{nk}}{\lambda} = \frac{N_k}{\lambda} \\
\text{Therefore } \lambda &= \sum_{k=1}^{K} \frac{N_k}{\pi_k} = N
\end{aligned}
$$

So we have

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

Question 4

Verify the relation

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

for the derivation of the logistic function defined by

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

Solution

Derive the logistic function,

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a} = \frac{\mathrm{d}}{\mathrm{d}a}\left(\frac{1}{1+\exp(-a)}\right) = \frac{\exp(-a)}{(1+\exp(-a))^2} = \sigma(1-\sigma)$$

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 the logistic regression model is given by

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

Solution

We have

$$y_n = \sigma(a_n),\ a_n = \mathbf{w}^{\mathrm{T}}\phi_n$$

Derive the error function for the logistic regression model,

$$\begin{aligned}
\nabla \ln p(\mathbf{t}|\mathbf{w}) &= -\sum_{n=1}^{N}\nabla\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \\
&= -\sum_{n=1}^{N}\frac{\mathrm{d}}{\mathrm{d}y_n}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}\frac{\mathrm{d}y_n}{\mathrm{d}a_n}\frac{\mathrm{d}a_n}{\mathrm{d}\mathbf{w}} \\
&= -\sum_{n=1}^{N}\left(\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right)y_n(1-y_n)\phi_n \\
&= \sum_{n=1}^{N}\frac{y_n - t_n}{y_n(1-y_n)}y_n(1-y_n)\phi_n
\end{aligned}$$

So we have

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

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 $\frac{c(c-1)}{2}$ discriminant functions.

Again, by drawing a specific example in two dimensions for $c = 3$, show that this approach can also lead to ambiguous regions.

Solution

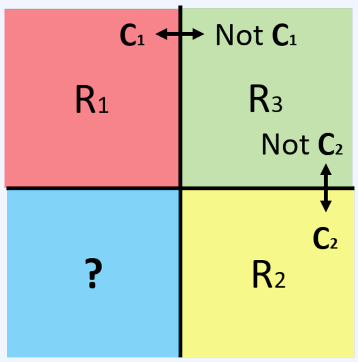

Figure 1: Ambiguous classification using 2 discriminant functions C_1 and C_2

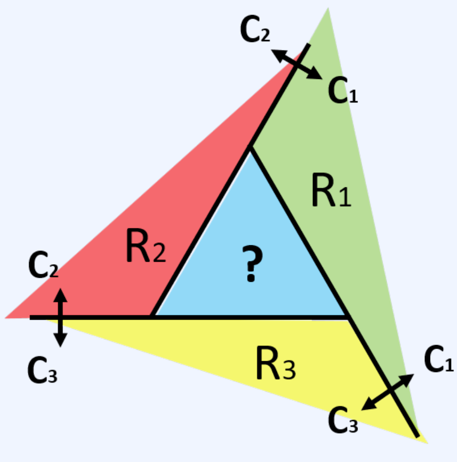

Figure 2: Ambiguous classification using 3 discriminant functions

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

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}_n$$

where $\alpha_n \geq 0$ 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.

Solution

1. Convex hulls intersect → Point sets not linear separable

As the convex hulls intersect, there must exists a point p such that

$$p = \sum_n \alpha_n \mathbf{x}_n = \sum_m \beta_m \mathbf{z}_m$$

where

$$\sum_n \alpha_n = \sum_m \beta_m = 1$$

Therefore,

$$\begin{aligned}
\hat{\mathbf{w}}^{\mathrm{T}} p + w_0 &= \hat{\mathbf{w}}^{\mathrm{T}} \sum_n \alpha_n \mathbf{x}_n + \left(\sum_n \alpha_n\right) w_0 \\
&= \sum_n \alpha_n (\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{x}_n + w_0) \\
\hat{\mathbf{w}}^{\mathrm{T}} p + w_0 &= \hat{\mathbf{w}}^{\mathrm{T}} \sum_m \beta_m \mathbf{z}_m + \left(\sum_m \beta_m\right) w_0 \\
&= \sum_n \beta_n (\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{z}_m + w_0)
\end{aligned}$$

Then assume they are linearly separable,

$$\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_n + w_0 > 0 \text{ for all } \mathbf{x}_n \text{ , while } \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{z}_m + w_0 < 0 \text{ for all } \mathbf{z}_m.$$

which leads to a contradiction:

$$\begin{aligned}
\hat{\mathbf{w}}^{\mathrm{T}} p + w_0 &= \sum_n \alpha_n (\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{x}_n + w_0) > 0 \\
\hat{\mathbf{w}}^{\mathrm{T}} p + w_0 &= \sum_m \alpha_m (\hat{\mathbf{w}}^{\mathrm{T}} \mathbf{z}_m + w_0) < 0
\end{aligned}$$

Therefore they are not linearly separable if the two convex hulls intersect.

2. Point sets linear separable → Convex hulls not intersect

This is the contra-positive proposition of the previous one.

The truth value of the contra-positive proposition agrees with that of the original proposition.

Therefore the convex hulls do not intersect if the two point sets are linear separable.