SUSTech

MACHINE-LEARNING

Fall 2020

Homework 4

Liu Yuxi 2020 年 11 月 2 日

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

Solution. We construct the Lagrangian function

$$L = w^T(m_2 - m_1) + \lambda(w^T w - 1)$$

Taking the gradient of L we obtain

$$\delta L = m_2 - m_1 + 2\lambda w$$

setting this gradient to zeros

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

and we get

$$w \propto (m_2 - m_1)$$

4.2. Show that the Fisher criterion

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

can be written in the form

$$j(w) = \frac{w^T S_B w}{w^T S_W w}$$

Solution.

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

$$\begin{aligned} s_1^2 + s_2^2 &= \sum_{n \in C_1} (y_n - m_1)^2 + \sum_{k \in C_2} (y_k - m_2)^2 \\ &= \sum_{n \in C_1} (w^T(x_n - m_1))^2 + \sum_{k \in C_2} (w^T(x_k - m_2))^2 \\ &= \sum_{n \in C_1} w^T(x_n - m_1)(x_n - m_1)^T w + \sum_{k \in C_2} w^T(x_k - m_2)(x_k - m_2)^T w \\ &= w^T S_W w \end{aligned}$$

Finally, we obtain

$$j(w) = \frac{w^T S_B w}{w^T S_W w}$$

4.3. Consider a agenerative classification model for K classes defined by prior class probabilities $p(C_k) = \pi_k$ and general class-conditional dendities $p(\phi|C_k)$ where ϕ is the input feature vector. Suppose we are given a training data set$\{\phi_n, t_n\}$ where $n = 1, ..., N$, and 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 data points are drown 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.

Solution. The likelihood is

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

taking the logarithm

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

with the constraint $\sum_k \pi_k = 1$, and we use Lagrange

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

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

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

4.4. Verify the relation

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

for the derivative of the logistic sigmoid function defined by

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

Solution.

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

4.5. 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

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

Solution.

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

and use

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

we get

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

$$\nabla a_n = \phi_n$$

Using the chain rule, we obtain

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

4.6. 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 $y_k(x) > 0$ for inputs x in class C_k and $y_k(x) < 0$ for 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 functino $y_{jk}(x)$ for each possible pair of classes C_j and C_k, such that $y_{jk}(x) > 0$ for patterns in class C_j and $y_{jk}(x) < 0$ 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.

Solution. As for the approach1, we have two linear discriminant function. and $y_k(x) > 0$ for inputs x in class C_k and $y_k(x) < 0$ for not

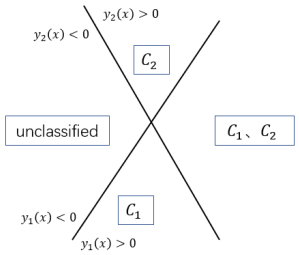

图 1: approach1 with c-1 linear discriminant functions

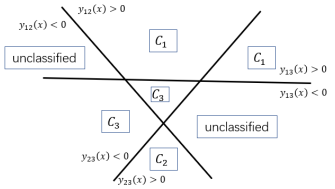

图 2: approach2 with three linear discriminant

in class C_k. But there is still several problem, when $y_1(x) > 0$ and $y_2(x) > 0$, we can't judge the point belong to C_1 or C_2.

As for the approach2, we have the classification as follows:

(a) If $y_{12}(x) > 0$ and $y_{13}(x) > 0$, then x $\in C_1$

(b) If $y_{12}(x) < 0$ and $y_{23}(x) > 0$, then x $\in C_2$

(c) If $y_{13}(x) < 0$ and $y_{23}(x) < 0$, then x $\in C_3$

The follow regions are still unclassified.

(a) $y_{12}(x) < 0$ and $y_{13}(x) > 0$. $y_{12}(x) > 0$ and $y_{23}(x) > 0$ and $y_{13} < 0$

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

$$x = \sum_n \alpha_n x^n$$

where $\alpha_n >= 0$ and $\sum_n \alpha_n = 1$. Consider a second set of point $\{z^m\}$and its corresponding convex hull. The two sets of points will

be linearly separable if there exists a vector $\hat{w}$ and a scalar ω_0 such that $\hat{w}^T x^n + \omega_0 > 0$ for all x^n, and $\hat{w}^T z^m + \omega_0 < 0$ for all 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. we have

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

$$x = \sum_n \alpha_n x^n$$

then we get

$$y(x) = \hat{w}^T (\sum_n \alpha_n x_n) + w_0$$

with $\sum_n \alpha_n = 1$

$$y(x) = \sum_n \alpha (\hat{w}^T x^n + w_0)$$

Similarly we get

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

If the convex hulls intersect, there must be at least one point exist in x and z. We denote this point as xz. And we get

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

But for linear separability, we have

$$y(x^n) = \hat{w}^T x^n + w_0 > 0$$
$$y(z^m) = \hat{w}^T z^m + w_0 > 0$$

There is no possible for the above formulations establish simultaneously.