Machine Learning Homework 4: Linear Models for Classification

叶璨铭， 12011404@mail.sustech.edu.cn

1 Discriminant Function: Maximum Class Separation

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

$x_{1:N}$ *and*

2 Discriminant Function: Fisher Criterion

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

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

Hint.

$$y = \mathbf{w}^{\mathbf{T}}\mathbf{x}, \qquad m_k = \mathbf{w}^{\mathbf{T}}\mathbf{m_k}, \qquad s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

3 Generative Classification Model

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

4 Discriminative Classification Model

Verify the relation

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

for the derivative of the logistic sigmoid function defined by

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

5 Discriminative Classification Model

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

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

6 Multi-Class

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.

7 Convex Hull

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 $\alpha_n >= 0$ 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.