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 $\mathbf{w} \propto (\mathbf{m_2} - \mathbf{m_1})$.

Important Note:

- The notation is extremely unfriendly for handwritten homeworks. We should **never** distinguish two entirely different symbols merely by their boldness.
- Therefore, in this paper we shall use μ for the **before-projection mean**, while using m to denote the **after-projection mean**. As for variance, we use $\sum$ to denote the before and S to denote the after.
- We advocate that does not indicate anything, in case anyone may write it or read it wrong.

Solution: The problem of the maximization of the class separation can be formulated as follows:

$$\max_w f(w) = w^T(\mu_2 - \mu_1)$$
$$s.t.\, w^T w = 1$$

Using Lagrange Multiplier λ, we can transform the problem to an unconstrained one.

$$\nabla f(w) = \lambda \nabla g(w)$$
$$g(w) = w^T w - 1$$
$$g(w) = 0$$

Since $\nabla f(w) = (\mu_2 - \mu_1)^T$ and $\nabla g(w) = 2w^T$, we obtain

$$w = \frac{1}{2\lambda}(\mu_2 - \mu_1) \sim (\mu_2 - \mu_1)$$

2 Discriminant Function: Fisher Criterion

Show that the Fisher criterion

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

can be written in the form

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

Hint.

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

Analysis:

- m_1, m_2, S_1, S_2 are the mean and variance after the projection. The lower case s_1 and s_2 denotes the standard deviation.
- S_B and S_W are "scatter matrix", which means how different the vectors are in the vector space.

- The B in S_B denotes "between-class", and W in S_W denotes "within-class".

$$S_B = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (\mu_\mathbf{i} - \mu_\mathbf{j})(\mu_\mathbf{i} - \mu_\mathbf{j})^T$$

$$S_W = \sum_{i=1}^{k} \Sigma_k$$

Solution:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{(\mathbf{w^T}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1}))^2}{\mathbf{w^T}\boldsymbol{\Sigma_1}\mathbf{w} + \mathbf{w^T}\boldsymbol{\Sigma_2}\mathbf{w}} = \frac{w^T(\mu_2 - \mu_1)(\mu_2 - \mu_1)^T w}{w^T(\Sigma_1 + \Sigma_2)w} = \frac{w^T S_B w}{w^T S_W w}$$

which is also refereed to as the generalized Rayleigh quotient.

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 $\mathcal{C}_k$.

Analysis.

Solution.

$$-\ln P(D|\pi_k) = -\sum_{n=1}^{N} \ln P(\phi|C_k) + \ln P(C_k|\pi_k)$$

$$\nabla_{\pi_k} - lnP(D|\pi_k) = -\sum_{n=1}^{N} (\ln P(C_k|\pi_k))` = -\sum_{n=1}^{N} \pi_k^{-1} = -\sum_{k=1}^{K} \frac{N_k}{\pi_k}$$

4 Discriminative Classification Model

Verify the relation

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

for the derivative of the logistic sigmoid function defined by

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

Solution.

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

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

and we know that

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

So

$$\frac{d(1+\exp(-a))}{da} = \sigma^3(1-\sigma)$$

5 Discriminative Classification Model

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

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

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.

8 Glossary

- scatter
  - [*VN*] ~ sth (on/over/around sth)~ sth (with sth)to throw or drop things in different directions so that they cover an area of ground 撒；撒播
  - [sing.] 散点，散落的样子。
- overstriking
  - 加粗
- bold
  - 加粗、勇敢。

9 References