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 $m_2 - m_1 = \mathbf{w^T}(\mathbf{m_2} - \mathbf{m_1})$ with respect to $\mathbf{w}$, using a Lagrange multiplier to enforce the constraint $\mathbf{w^T w} = \mathbf{1}$, 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

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

Hint.

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

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".

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

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

Solution:

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

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 $\mathcal{C}_k$.

Note: 这里题目再次符号滥用, I_{jk} 是 if pattern n is from class C_k 这个概率, I 表示 $\begin{cases} 1 & \text{satisfy} \\ 0 & \text{else} \end{cases}$, 而 t_{nj} 表示矩阵。

Solution:

① let $z_n = \text{argmax}(t_n)$

实际上就 $\pi_{1:K}$

$$like(\pi_k \mid \varphi_{1:n}, t_{1:n}) = p(\varphi_{1:n}, t_{1:n} \mid \pi_k)$$

$$= \prod_{n=1}^{N} p(\varphi_n, t_n \mid \pi_k)$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} \left(p(\varphi_n \mid C_k)\pi_k\right)^{t_{nk}}$$

$$\ln like = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left[\ln p(\varphi_n|C_k) + \ln \pi_k\right]$$

② $\underset{\pi_k}{\text{argmax}}\ \ln like \quad s.t. \quad \sum_{k=1}^{K} \pi_k = 1$

Using Lagrange multiplier method,

$$\begin{cases} L(\pi_k, \lambda) = \ln like + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right) \\ \frac{\partial L(\pi_k, \lambda)}{\partial \pi_k} = \frac{\partial L(\pi_k, \lambda)}{\partial \lambda} = 0 \end{cases}$$

$$\Rightarrow \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}\left(\frac{1}{\pi_k}\right) + \lambda \sum_{k=1}^{K} 1 = 0$$

$\pi_k = \pi_k$ 时才相等, 故 k 确定, Σ 消去。

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

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

两边 $\sum_{k=1}^{K}$

$$-\lambda = N$$

故 $(-\lambda)\pi_k = N\pi_k = N_k \Rightarrow \pi_k = \frac{N_k}{N}$

4 Discriminative Classification Model

Verify the relation

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

for the derivative of the logistic sigmoid function defined by

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

$$\frac{d\sigma(a)}{da} = \left([\sigma(a)^{-1}]^{-1} \right)'$$

$$= (-1)\,[\sigma(a)^{-1}]^{-2}\,\frac{d(1+e^{-a})}{da}$$

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

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

$$\therefore \frac{d\sigma(a)}{da} = \sigma(a)^2 e^{-a}$$

而 $\sigma(a)e^{-a} = \frac{1}{1+e^{-a}} e^{-a} = \frac{1}{e^a+1}$

而 $\sigma(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{e^a+1} = 1 - \frac{1}{e^a+1}$

$\therefore \sigma(a)e^{-a} = (1-\sigma(a))$ 代回

$$\frac{d\sigma(a)}{da} = \sigma(a)\cdot\sigma(a)e^{-a} = \sigma(a)(1-\sigma(a))$$

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

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

Note: E 会写作 Err 或 J 或 Loss。

E 表示期望。

$$p(y_n|w) \sim \text{Bern}(y|t_n)$$

Solution: $Err(w) = -\ln p(D|w)$

$$= -\sum_{n=1}^{N} \ln\left[y_n^{t_n} (1-y_n)^{(1-t_n)} \right]$$

$$= -\sum_{n=1}^{N} t_n \ln y_n + (1-t_n)\ln(1-y_n)$$

其中 $y_n = \sigma(w^T \varphi(x_n))$

$$\nabla_w Err(w) = -\sum_{n=1}^{N} \frac{t_n}{y_n} y_n(1-y_n)\varphi^T(x_n) + \frac{(1-t_n)}{(1-y_n)}(-1)\,y_n(1-y_n)\,\varphi^T(x_n)$$

$$= \sum_{n=1}^{N} (y_n - t_n)\varphi^T(x_n)$$

(w为列向量, ∇_w为行向量, 故为 φ^T)

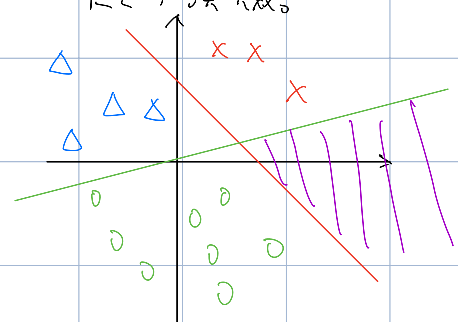
6 Multi-Class

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.

(a) ① (c-1)个分类器，对第c类，是当且仅当前面的c-1类都<0。

② c类必须是等价类，满足等价关系，是不相关

③ 本题要构造 线性分类器 对2类收敛，但是对3类失效。

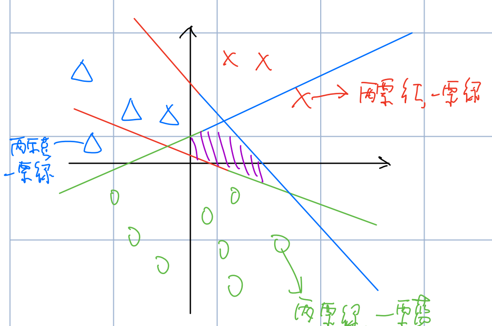


紫色区域被绿线认定 0类

被红线认定 x类

因此出现一个在这个区域的样本时，无法判定属于哪一类，为ambiguous.

(b) ① 两两用分类器，投票决定是哪一类



紫色区域被3个分类器分别认为蓝，红，绿，各得一票，无法决策，为ambiguous情况。

7 Convex Hull

凸包(外壳)

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

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

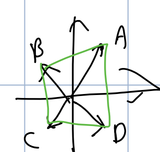
where $\alpha_n >= 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{\mathbf{z}^m\}$ and its corresponding convex hull. The two sets of points will be linearly separable if there exists a vector $\hat{\mathbf{w}}$ and a scalar w_0 such that $\hat{\mathbf{w}}^T\mathbf{x}^n + w_0 > 0$ for 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.

Analysis: ① $\{x^n\}$表示有n个x，n仅为下标，不是次方。

② 每个x^i是q维向量，$x^i \in R^q$

③ $\{z^m\}$是m个z，z也为q维向量。

④ 凸包来自计算几何，二维中，有{A,B,C,D}，那么这四个点是凸包的一部分



(设结 convex hull是对$\forall \alpha_n$，x组成的集合，而不是一个x)

二维中这个集合就是凸多边形。

intersect指的是集合区域内部相交的概念。(包括边的相交都算的)

Solution: ① 这两个命题为逆否关系，只用证明其中一个。

② 若凸包相交，则∃点y，系数α_n, β_m 使

$y = \sum_n \alpha_n x_n$ 且 $y = \sum_m \beta_m y_n$，y为凸包{x}与{z}的某一交点。

③ 假设∃w, w_0，使 $w^T x_n + w_0 > 0$ 且 $w^T z_m + w_0 < 0$

$$w^T y + w_0 = w^T\left(\sum_n \alpha_n x_n\right) + w_0 = \sum_n \alpha_n w^T x_n + \left(\sum_n \alpha_n\right) w_0 = \sum_n \alpha_n \left[w^T x_n + w_0\right]$$

由于 $w^T x_n + w_0 > 0$，$\alpha_n \geq 0$ (且不全为0)，∴ $w^T y + w_0 > 0$

④ 同理，$w^T y + w_0 = \sum_m \beta_m \left[w^T z_m + w_0\right] < 0$

与③矛盾

⑤ 综上，命题得证，此定理又称超平面分离定理。