# 第五章 线性判别函数

4、考虑判别中用的超平面。

(a) 证明在从超平面 $g(\mathbf{x})=\mathbf{w}^t\mathbf{x}+w_0=0$ 到点 $\mathbf{x}_a$ 的距离为 $\left|g(\mathbf{x}_a)\right|/\|\mathbf{w}\|$，且对应的点为约束条件 $g(\mathbf{x})=0$ 下的满足使 $\|\mathbf{x}-\mathbf{x}_a\|^2$ 最小的 $\mathbf{x}$。

(b) 证明 $\mathbf{x}_a$ 到超平面的投影为：

$$\mathbf{x}_p=\mathbf{x}_a-\frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|^2}\mathbf{w}$$

证明：

(a) 令 $\mathbf{x}_p$ 为点 $\mathbf{x}_a$ 在超平面 $g(\mathbf{x})=0$ 上的投影点，即矢量 $\mathbf{x}_a-\mathbf{x}_p$ "垂直于"超平面，令模长为 $r$，即：$\left\|\mathbf{x}_a-\mathbf{x}_p\right\|=r$。同时权矢量 $\mathbf{w}$ 为超平面 $g(\mathbf{x})=0$ 的法矢量，也"垂直于"超平面，因此 $\mathbf{x}_a-\mathbf{x}_p$ 与 $\mathbf{w}$ 方向相同（假设 $\mathbf{x}_a$ 处在超平面的正侧），因此有：

$$\mathbf{x}_a-\mathbf{x}_p=r\frac{\mathbf{w}}{\|\mathbf{w}\|},\qquad\text{（其中}\frac{\mathbf{w}}{\|\mathbf{w}\|}\text{为单位矢量）}$$

$$\mathbf{x}_a=\mathbf{x}_p+r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

等式两边左乘 $\mathbf{w}^t$：

$$\mathbf{w}^t\mathbf{x}_a=\mathbf{w}^t\mathbf{x}_p+r\frac{\mathbf{w}^t\mathbf{w}}{\|\mathbf{w}\|}=\mathbf{w}^t\mathbf{x}_p+r\|\mathbf{w}\|$$

而：

$$\begin{aligned}g(\mathbf{x}_a)&=\mathbf{w}^t\mathbf{x}_a+w_0\\&=\mathbf{w}^t\mathbf{x}_p+r\|\mathbf{w}\|+w_0\\&=g(\mathbf{x}_p)+r\|\mathbf{w}\|=r\|\mathbf{w}\|\qquad(\mathbf{x}_p\text{为超平面上点},\ g(\mathbf{x}_p)=0)\end{aligned}$$

因此：

$$r=\frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|}$$

当 $\mathbf{x}_a$ 处于超平面的负侧时，$g(\mathbf{x}_a)<0$，同理可得 $r=-\frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|}$，因此：

$$r=\frac{|g(\mathbf{x}_a)|}{\|\mathbf{w}\|}$$

令 $\mathbf{x}_q$ 为超平面上的点，$\mathbf{x}_p$ 为点 $\mathbf{x}_a$ 在超平面上的投影点，则有：

$$\mathbf{x}-\mathbf{x}_q=(\mathbf{x}-\mathbf{x}_p)+(\mathbf{x}_p-\mathbf{x}_q)$$

两边分别左乘 $(\mathbf{x}-\mathbf{x}_q)^t$ 和 $\left[(\mathbf{x}-\mathbf{x}_p)+(\mathbf{x}_p-\mathbf{x}_q)\right]^t$，则有等式：

$$\begin{aligned}(\mathbf{x}-\mathbf{x}_q)^t(\mathbf{x}-\mathbf{x}_q)&=\left[(\mathbf{x}-\mathbf{x}_p)+(\mathbf{x}_p-\mathbf{x}_q)\right]^t\left[(\mathbf{x}-\mathbf{x}_p)+(\mathbf{x}_p-\mathbf{x}_q)\right]\\&=(\mathbf{x}-\mathbf{x}_p)^t(\mathbf{x}-\mathbf{x}_p)+(\mathbf{x}-\mathbf{x}_p)^t(\mathbf{x}_p-\mathbf{x}_q)\\&\quad+(\mathbf{x}_p-\mathbf{x}_q)^t(\mathbf{x}-\mathbf{x}_p)+(\mathbf{x}_p-\mathbf{x}_q)^t(\mathbf{x}_p-\mathbf{x}_q)\end{aligned}$$

其中：

$$(\mathbf{x}-\mathbf{x}_q)^t(\mathbf{x}-\mathbf{x}_q)=\|\mathbf{x}-\mathbf{x}_q\|^2$$

$$(\mathbf{x}-\mathbf{x}_p)^t(\mathbf{x}-\mathbf{x}_p)=\|\mathbf{x}-\mathbf{x}_p\|^2$$

$$(\mathbf{x}_p-\mathbf{x}_q)^t(\mathbf{x}_p-\mathbf{x}_q)=\|\mathbf{x}_p-\mathbf{x}_q\|^2$$

由于矢量 $\mathbf{x}-\mathbf{x}_p$ 与超平面正交，而 $\mathbf{x}_p-\mathbf{x}_q$ 处在超平面内，因此：

$$(\mathbf{x}-\mathbf{x}_p)^t(\mathbf{x}_p-\mathbf{x}_q)=(\mathbf{x}_p-\mathbf{x}_q)^t(\mathbf{x}-\mathbf{x}_p)=0$$

所以：

$$\|\mathbf{x}-\mathbf{x}_q\|^2=\|\mathbf{x}-\mathbf{x}_p\|^2+\|\mathbf{x}_p-\mathbf{x}_q\|^2\geq\|\mathbf{x}-\mathbf{x}_p\|^2$$

只有当 $\mathbf{x}_p=\mathbf{x}_q$ 时取得等号。因此 $\mathbf{x}_p$ 是使约束条件 $g(\mathbf{x})=0$ 下的满足使 $\|\mathbf{x}-\mathbf{x}_a\|^2$ 最小的 $\mathbf{x}$。

(b) 当 $\mathbf{x}_a$ 处于超平面的正侧时：

$$\mathbf{x}_a-\mathbf{x}_p=r\frac{\mathbf{w}}{\|\mathbf{w}\|}=\frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|}=\frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|^2}\mathbf{w}$$

当 $\mathbf{x}_a$ 处于超平面的负侧时：

$$\mathbf{x}_a - \mathbf{x}_p = -r\frac{\mathbf{w}}{\|\mathbf{w}\|} = -\frac{-g(\mathbf{x}_a)}{\|\mathbf{w}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|} = \frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|^2}\mathbf{w}$$

因此：

$$\mathbf{x}_p = \mathbf{x}_a - \frac{g(\mathbf{x}_a)}{\|\mathbf{w}\|^2}\mathbf{w}$$

14、考虑平方误差准则函数：

$$J_s(\mathbf{a}) = \sum_{i=1}^{n}\left(\mathbf{a}^t\mathbf{y}_i - b_i\right)^2$$

令 $b_i = b$，取如下 6 个训练点：

$w_1$： $(1,5)^t, (2,9)^t, (-5,-3)^t$

$w_2$： $(2,-3)^t, (-1,-4)^t, (0,2)^t$

(a) 计算它的Hessian矩阵；

(b) 假定二次准则函数，计算最优学习率 $h$ 。

解：

(a)

$$\nabla J = \frac{dJ_s(\mathbf{a})}{d\mathbf{a}} = \sum_{i=1}^{n} 2\left(\mathbf{a}^t\mathbf{y}_i - b_i\right)\mathbf{y}_i$$

$$\mathbf{H} = 2\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^t$$

将样本变为增广向量：

$$\mathbf{y}_1 = \begin{pmatrix}1\\1\\5\end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix}1\\2\\9\end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix}1\\-5\\-3\end{pmatrix}$$

$$\mathbf{y}_4 = \begin{pmatrix}-1\\-2\\3\end{pmatrix}, \mathbf{y}_5 = \begin{pmatrix}-1\\1\\4\end{pmatrix}, \mathbf{y}_6 = \begin{pmatrix}-1\\0\\-2\end{pmatrix}$$

Hessian 矩阵为：

$$\mathbf{H} = \begin{pmatrix}12 & -2 & 12\\ -2 & 70 & 72\\ 12 & 72 & 288\end{pmatrix}$$

(b) 准则函数为二次函数，因此 $J_s$ 在 $\mathbf{a}(k)$ 附近的 Taylor 展开式为：

$$J_s\left(\mathbf{a}(k+1)\right)=J_s\left(\mathbf{a}(k)\right)+\nabla J_s^t\left(\mathbf{a}(k)\right)\left(\mathbf{a}-\mathbf{a}(k)\right)+\frac{1}{2}\left(\mathbf{a}-\mathbf{a}(k)\right)^t\mathbf{H}\left(\mathbf{a}-\mathbf{a}(k)\right)$$

而：$\mathbf{a}(k+1)=\mathbf{a}(k)-h(k)\nabla J\left(\mathbf{a}(k)\right)$

因此：

$$J_s\left(\mathbf{a}(k+1)\right)=J\left(\mathbf{a}(k)\right)-h(k)\left\|\nabla J\left(\mathbf{a}(k)\right)\right\|^2+\frac{1}{2}h^2(k)\nabla J^t\left(\mathbf{a}(k)\right)\mathbf{H}\nabla J\left(\mathbf{a}(k)\right)$$

寻找使得$J_s\left(\mathbf{a}(k+1)\right)$最小的$h(k)$，上式对$h(k)$求导数：

$$\frac{dJ_s\left(\mathbf{a}(k+1)\right)}{dh(k)}=-\left\|\nabla J\left(\mathbf{a}(k)\right)\right\|^2+h(k)\nabla J^t\left(\mathbf{a}(k)\right)\mathbf{H}\nabla J\left(\mathbf{a}(k)\right)=0$$

因此最优学习率：

$$h(k)=\frac{\left\|\nabla J\left(\mathbf{a}(k)\right)\right\|^2}{\nabla J^t\left(\mathbf{a}(k)\right)\mathbf{H}\nabla J\left(\mathbf{a}(k)\right)}$$

15、在感知器算法的收敛证明中(定理 5.1)取尺度因子$a=b^2/g$。

(a) 用 5.5 节的符号，证明如果$a>b^2/(2g)$，那么需要校正的次数的最大值为：

$$k_0=\frac{\left\|\mathbf{a}_1-a\,\mathbf{a}\right\|^2}{2ag-b^2}$$

(b) 当$\mathbf{a}_1=\mathbf{0}$时，$a$ 为何值时使得$k_0$最小？

证明：

(a) 在定理 5.1 证明中，有如下不等式成立：

$$\left\|\mathbf{a}(k+1)-a\,\hat{\mathbf{a}}\right\|^2\le\left\|\mathbf{a}(k)-a\,\hat{\mathbf{a}}\right\|^2-2ag+b^2$$

如果取$a>b^2/(2g)$，则有$2ag>b^2$，$2ag-b^2>0$。即每次迭代之后$\mathbf{a}(k+1)$到$a\,\hat{\mathbf{a}}$距离的平方减小了$2ag-b^2$，经$k$次迭代之后：

$$\left\|\mathbf{a}(k+1)-a\,\hat{\mathbf{a}}\right\|^2\le\left\|\mathbf{a}(1)-a\,\hat{\mathbf{a}}\right\|^2-k\left(2ag-b^2\right)$$

因$\left\|\mathbf{a}(1)-a\,\hat{\mathbf{a}}\right\|^2-k\left(2ag-b^2\right)\ge0$，所以经过不超过$k_0$此的迭代之后，终止：

$$k_0=\frac{\left\|\mathbf{a}(1)-a\,\hat{\mathbf{a}}\right\|^2}{2ag-b^2}$$

(b) 当$\mathbf{a}(1)=\mathbf{0}$时：

$$k_0=\frac{\left\|-a\,\hat{\mathbf{a}}\right\|^2}{2ag-b^2}=\frac{\left\|\hat{\mathbf{a}}\right\|^2}{2g/a-b^2/a^2}$$

求 $k_0$ 的最小值，即求 $2g/a-b^2/a^2$ 的最大值，令 $q=\frac{1}{a}$，构造目标函数：

$$J(q)=2gq-b^2q^2$$

$$\frac{dJ(q)}{dq}=2g-2b^2q=0$$

因此：$q=\frac{g}{b^2}$，即当 $a=\frac{b^2}{g}$ 时，$k_0$ 取得极值，可以验证此极值为 $k_0$ 的最小值。

# 第 10 章 无监督学习和聚类

14、$\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n$ 为 $n$ 个 $d$ 维样本，$\boldsymbol{\Sigma}$ 是任意的大小为 $d\times d$ 的非奇异矩阵，证明使得：

$$\sum_{k=1}^{n}\left(\mathbf{x}_k-\mathbf{x}\right)^t\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_k-\mathbf{x}\right)$$

最小的 $\mathbf{x}$ 就是样本的均值 $\overline{\mathbf{x}}=1/n\sum_{k=1}^{n}\mathbf{x}_k$。

证明：令：

$$J(\mathbf{x})=\sum_{k=1}^{n}\left(\mathbf{x}_k-\mathbf{x}\right)^t\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_k-\mathbf{x}\right)$$

则：

$$\frac{dJ(\mathbf{x})}{d\mathbf{x}}=\sum_{k=1}^{n}\left(\boldsymbol{\Sigma}^{-1}+\left(\boldsymbol{\Sigma}^{-1}\right)^t\right)\left(\mathbf{x}-\mathbf{x}_k\right)=\mathbf{0}$$

因此：

$$n\left(\boldsymbol{\Sigma}^{-1}+\left(\boldsymbol{\Sigma}^{-1}\right)^t\right)\mathbf{x}=\left(\boldsymbol{\Sigma}^{-1}+\left(\boldsymbol{\Sigma}^{-1}\right)^t\right)\sum_{k=1}^{n}\mathbf{x}_k$$

假设 $\left(\boldsymbol{\Sigma}^{-1}+\left(\boldsymbol{\Sigma}^{-1}\right)^t\right)$ 非奇异，等式两边左乘 $\left(\boldsymbol{\Sigma}^{-1}+\left(\boldsymbol{\Sigma}^{-1}\right)^t\right)^{-1}$，则有：

$$\mathbf{x}=1/n\sum_{k=1}^{n}\mathbf{x}_k\text{。}$$