Bishop, ch 4

Thursday, February 23, 2017 12:09 AM

LINEAR MODELS FOR CLASSIFICATION

For K classes the target can be represented as

$$t = (0, 0, \dots, \underset{\uparrow j}{1}, 0, \dots, 0) \in \mathbb{R}^K$$

indicating that the corresponding $x \in C_j$. We will consider models of the form

$$y(x) = f(w^T x + w_0)$$

where f is a nonlinear function. For $x \in \mathbb{R}^D$, the space is divided into regions where $w^T x + w_0 = \text{const}$. These $D-1$ hyperplanes are the decision boundaries.

Discriminant Functions

A discriminant function is a rule that assigns X to only one class C_k, for $k = 1, 2, \dots, K$. Consider

$$y(x) = w^T x + w_0$$

For a 2-class problem we define

$$\begin{cases} x \in C_1 & \text{if } y(x) > 0 \\ x \in C_2 & \text{if } y(x) \leq 0 \end{cases}$$

Thus $y = 0$ is the decision boundary. If x_A and x_B are any two points on the boundary, $w^T(x_A - x_B) = 0$, thus w is normal to this $D-1$ dimensional hyperplane.

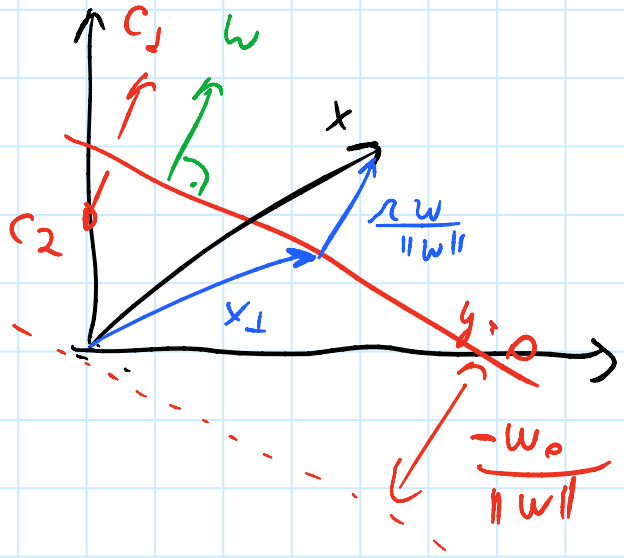
For any x we can write

$$x = x_\perp + r\frac{w}{\|w\|}$$

where $y(x_\perp) = 0$.

Now $y(x) = w^T x + w_0 = \underbrace{w^T x_\perp + w_0}_{y(x_\perp)=0} + r\|w\|$

$$r = \frac{y(x)}{\|w\|}$$



For a K-class problem, we can avoid ambiguities in defining the decision regions by considering K discriminant functions:

$$y_k(x) = w_k^T x + w_{k0} \quad (k = 1, \ldots, K)$$

Then,

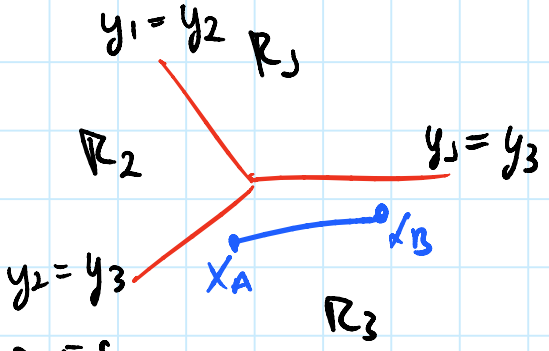
$$x \in C_k \text{ if } y_k(x) > y_j(x), \forall j \neq k.$$

The decision boundary between classes C_k and C_j is given by $y_k(x) = y_j(x)$ which is a $D-1$ dimensional hyperplane determined by

$$(w_k - w_j)^T x + (w_{k0} - w_{j0}) = 0$$

Moreover, any region R_k is convex. To see this, let x_A and x_B be in R_k, and consider the line segment $\hat{x} = \lambda x_A + (1-\lambda) x_B$, $\lambda \in [0,1]$. By linearity

$$y_k(\hat{x}) = \lambda y_k(x_A) + (1-\lambda) y_k(x_B)$$
$$> \lambda y_j(x_A) + (1-\lambda) y_j(x_B) = y_j(\hat{x}) \quad \therefore \hat{x} \in R_k$$



LEAST SQUARES

Now we consider the least-squares solution for $k = 1, \dots, K$. Let $w_k \to \begin{pmatrix} w_{k0} \\ w \end{pmatrix} \equiv w_k$ and $x \to \begin{pmatrix} 1 \\ x \end{pmatrix} \equiv x$, such that

$$y_k(x) = w_k^T x \longrightarrow y(x) = W^T x$$

where

$$y(x) = \begin{pmatrix} y_1(x) \\ \vdots \\ y_k(x) \end{pmatrix}_{K \times 1} \qquad W = \begin{pmatrix} w_1 \mid \cdots \mid w_k \end{pmatrix}_{(D+1) \times K}$$

We have training data $\{(x_n, t_n)\}_{n=1}^N$ and we want to minimize

$$E_D(W) = \frac{1}{2} \sum_{n=1}^{N} \| t_n - W^T x_n \|^2$$

Define $T = \begin{pmatrix} t_1^T \\ \vdots \\ t_N^T \end{pmatrix}$ and $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix}$.

Notice that $W^T x_n = (x_n^T W)^T = (XW)_n^T$ and $t_n = T_n^T$, so the summand is $\| (T - XW)_n^T \|^2$. For any set of vectors $\{v_n\}$ we have $\sum_n \|v_n\|^2 = \sum_n v_n^T v_n = \mathrm{Tr}(V V^T)$ where $V = \begin{pmatrix} v_1^T \\ \vdots \\ v_N^T \end{pmatrix}$.

Thus the least-squares error can be written as

$$E_D(W) = \frac{1}{2} Tr\{(T - XW)^T (T - XW)\}$$

Then we have

$$\frac{\partial E_D(W)}{\partial W} = \frac{\partial}{\partial W} \frac{1}{2} Tr\{-T^T XW - (XW)^T T + W^T X^T X W\}$$

$$= -X^T T + X^T X W = 0$$

where we used $\frac{\partial}{\partial X} Tr(XA) = A^T$ and $\frac{\partial}{\partial X} Tr(X^T B X) = (B + B^T)X$.

The solution is thus

$$W = (X^T X)^{-1} X^T T = X^+ T$$

where $X^+ \equiv (X^T X)^{-1} X^T$ is the left pseudo-inverse, $X^+ X = I$.
The discriminant function is

$$y(x) = (X^+ T)^T x$$

Least squares suffers from being to sensitive to outliers, and also, sometimes, it cannot find the correct regions even when the data is linearly separable.

FISHER LINEAR DISCRIMINANT (LDA)

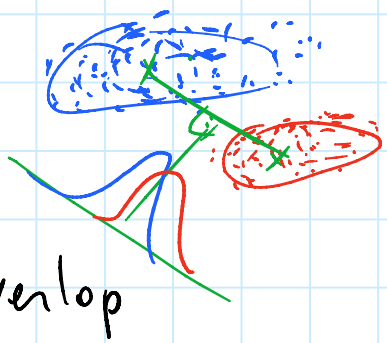
Suppose we have N_1 points from class C_1 and N_2 points from class C_2. We want to project the data into 1D: $y = w^T x$. If we build a classifier such that $y > -w_0 \Rightarrow x \in C_1$, and $x \in C_2$ otherwise, we obtain the previous linear classifier.

One idea is to maximize the projected means:

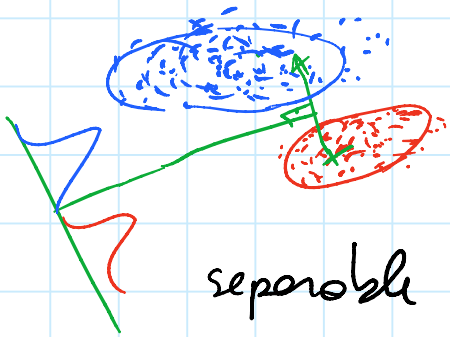
$$w^T(m_1 - m_2)$$

where $m_1 = \frac{1}{N_1} \sum_{x_i \in C_1} x_i$, $m_2 = \frac{1}{N_2} \sum_{x_i \in C_2} x_i$.

It can be shown that $w \propto m_1 - m_2$ is the solution to this problem. This represents a serious drawback if the data has high "off-diagonal" covariance:



overlap



separable

Even when the data is linearly separable, the projection on $m_1 - m_2$ may be highly overlapping.

We need to find a direction such that $\|m_1 - m_2\|$ is maximized, while keeping the variance of the within class projected data as small as possible.

Let $\mu_K = w^T m_K$ be the projected mean and $y_n = w^T x_n$ the projected data point. The within-class variance is

$$s_K^2 = \sum_{y_n \in C_K} (y_n - \mu_K)^2$$

For a two-class problem the Fisher criterion is $J(w) = \frac{(\mu_2 - \mu_1)^2}{s_1^2 + s_2^2}$.

Re-writing this:

$$(\mu_2 - \mu_1)^2 = w^T (m_2 - m_1)(m_2 - m_1)^T w \equiv w^T S_B w$$

$$s_1^2 + s_2^2 = \sum_{x_n \in C_1} w^T (x_n - m_1)(x_n - m_1)^T w + \sum_{x_n \in C_2} w^T (x_n - m_2)(x_n - m_2)^T w$$

$$\equiv w^T S_W w$$

$$\therefore \quad J(w) = \frac{w^T S_B w}{w^T S_W w}$$

S_B is the between class covariance, and S_W is the within class covariance. Thus

$$\frac{\partial J}{\partial w} = 0 = \frac{S_B w}{w^T S_W w} - \frac{w^T S_B w \, S_W w}{(w^T S_W w)^2}$$

$$(w^T S_B w) S_W w = (w^T S_W w) S_B w$$

$$S_B w = (m_2 - m_1)(m_2 - m_1)^T w \propto m_2 - m_1$$

Therefore

$$w \propto S_w^{-1}(m_2 - m_1)$$

This is the Fisher's linear discriminant. We build a classifier from this by deciding:

$$\begin{cases} x \in C_1 \text{ if } y(x) = w^T x \geq y_0 \\ x \in C_2 \text{ otherwise} \end{cases}$$

for some threshold y_0. Since $y = w^T x$ is a sum of random variables we can assume

$$p(y|C_k) = N(\mu_k, \sigma_k^2)$$

From this we can estimate μ_k, σ_k (MLE), and using decision theory we can find the optimal y_0.

Now let us discuss the relation of LDA to least squares. We use a different target scheme. If $x_n \in C_1$ then $t_n = \frac{N}{N_1}$, and if $x_n \in C_2$ then $t_n = -\frac{N}{N_2}$.

Consider

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)^2$$

$$\frac{\partial E}{\partial w_0} = 0 = \sum_{n=1}^{N}(w^T x_n + w_0 - t_n), \quad w_0 = \frac{1}{N}\sum_{n=1}^{N}(t_n - w^T x_n)$$

$$w_0 = \frac{1}{N}\left(N_1 \frac{N}{N_1} - N_2 \frac{N}{N_2}\right) - w^T\left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)$$

$$w_0 = -w^T m \quad \text{where } m = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

$$\frac{\partial E}{\partial w} = 0 = \sum_{n=1}^{N} (w^T x_n + w_0 - t_n) x_n$$

$$0 = \sum_{n=1}^{N} x_n x_n^T w + w_0 \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} t_n x_n$$

$$\sum_{n=1}^{N} t_n x_n = \sum_{x_n \in C_1} \frac{N}{N_1} x_n - \sum_{x_n \in C_2} \frac{N}{N_2} x_n = N(m_1 - m_2)$$

$$w_0 \sum_{n=1}^{N} x_n = -w^T m (N_1 m_1 + N_2 m_2)$$

$$= -(N_1 m_1 + N_2 m_2)\frac{(N_1 m_1 + N_2 m_2)^T}{N} w$$

$$\sum_{n=1}^{N} x_n x_n^T = \sum_{x_n \in C_1} x_n x_n^T + \sum_{x_n \in C_2} x_n x_n^T$$

$$= \sum_{x_n \in C_1} (x_n - m_1)(x_n - m_1)^T + N_1 m_1 m_1^T$$

$$+ \sum_{x_n \in C_2} (x_n - m_2)(x_n - m_2)^T + N_2 m_2 m_2^T$$

Therefore

$$\left(S_W + N_1 m_1 m_1^T + N_2 m_2 m_2^T - \frac{1}{N}(N_1 m_1 + N_2 m_2)(N_1 m_1 + N_2 m_2)^T \right) w = N(m_1 - m_2)$$

$$\left(S_W + \frac{N_1 N_2}{N}\left(m_1 m_1^T + m_2 m_2^T - 2 m_1 m_2^T \right) \right) w = N(m_1 - m_2)$$

Finally,

$$\left(S_W + \frac{N_1 N_2}{N} S_B \right) w = N(m_1 - m_2)$$

Since $S_B w \propto m_2 - m_1$, the solution of this equation is $w \propto S_W^{-1}(m_2 - m_1)$, which is the same as LDA.

The classification will be $x \in C_1$ if $y(x) > 0$ and $x \in C_2$ otherwise, where $y = w^T x + w_0$.

Fisher's LDA can be generalized for multiple classes, but we are not going to consider it here.

Probabilistic generative Models

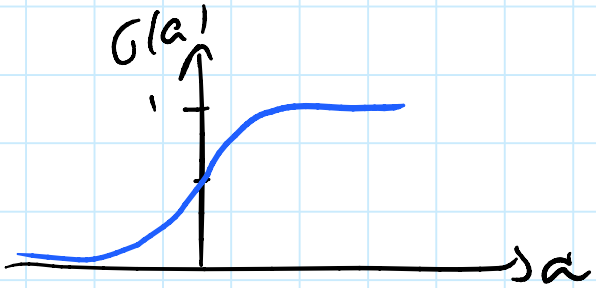
Here we model $p(x|C_k)$ and $p(C_k)$ to finally employ Bayes theorem and use $p(C_k|x)$ for classification. Let us consider a two-class problem:

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} = \frac{1}{1+e^{-a}} \equiv \sigma(a)$$

where

$$a = \log \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \log \frac{p(C_1|x)}{p(C_2|x)}$$

$\sigma(a)$ is called the sigmoid function, and it maps the entire real line on the $(0,1)$ interval



$$1 + e^{-a} = 1/\sigma, \quad e^{-a} = \frac{1}{\sigma} - 1, \quad e^{a} = \frac{1}{1/\sigma - 1} = \frac{\sigma}{1-\sigma}$$

The inverse of σ is the logit function

$$a = \log \frac{\sigma}{1-\sigma} = \log \frac{p(C_1|x)}{p(C_2|x)}$$

For $K > 2$ we have

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} = \frac{e^{a_k}}{\sum_j e^{a_j}}$$

where $a_k = \log p(x|C_k)p(C_k)$. The above function is known as normalized exponential or softmax function since if $a_k \gg a_j$ for any $j \neq k$ then $p(C_k|x) \approx 1$ and $p(C_j|x) \approx 0$.

Now we assume some form of distribution for the class conditional. Let us assume equal covariances:

$$p(x|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}$$

Consider $K=2$. Then $p(C_1|x) = \sigma(a)$ where

$$\begin{aligned} a &= \log \frac{p(x|C_1)}{p(x|C_2)} + \log \frac{p(C_1)}{p(C_2)} \\ &= -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) + \log \frac{p(C_1)}{p(C_2)} \\ &= x^T \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 - x^T \Sigma^{-1}\mu_2 + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \log \frac{p(C_1)}{p(C_2)} \\ &= (\Sigma^{-1}(\mu_1 - \mu_2))^T x - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \log \frac{p(C_1)}{p(C_2)} \end{aligned}$$

Define $w = \Sigma^{-1}(\mu_1 - \mu_2)$

$$w_0 = \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\frac{p(C_1)}{p(C_2)}$$

Then $a = w^T x + w_0$ //

The decision boundary is a linear function in x-space. The priors $p(C_k)$ only enter the bias w_0 and their effect is only to translate the decision boundary.

Now let $K > 2$. Recall that $p(C_k|x) = \frac{e^{a_k}}{\sum_j e^{a_j}}$ where

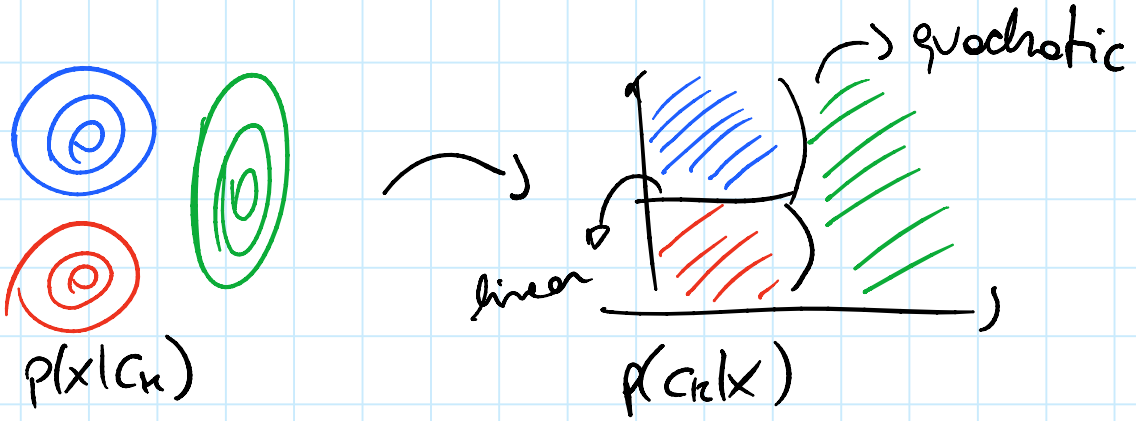
$$a_k = \log p(x|C_k) + \log p(C_k)$$
$$= -\frac{D}{2}\log\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)$$
$$= \underbrace{-\frac{D}{2}\log\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}x^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k}_{\text{this will be the same for all } e^{a_j}\text{, so it cancels with the denominator.}} + \log p(C_k)$$

We thus have

$$a_k = w_k^T x + w_{k0}$$
$$w_k = \Sigma^{-1}\mu_k$$
$$w_{k0} = -\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log p(C_k)$$

The decision boundary will be give by $p(C_k|x) = p(C_j|x)$ where these are the two largest posteriors. This gives $a_k = a_j$ which is linear in x.

Had we not chosen Σ the same for all classes, the quadratic term would not have cancelled, and the decision boundary would be a quadratic function of x. For instance:



Once we specify a form for $p(x|C_k)$ we can use MLE to estimate parameters and the priors. Assume we have training data $\{(x_n, t_n)\}_{n=1}^{N}$, where $t_n = 1$ if $x_n \in C_1$, and $t_n = 0$ if $x_n \in C_2$, where we are considering $K=2$ problem. Assume $x \sim \mathcal{N}(\mu_1, \Sigma)$ for $x \in C_1$ and $p(C_1) = \pi$, and also $x \sim \mathcal{N}(\mu_2, \Sigma)$ for $x \in C_2$ and $p(C_2) = 1-\pi$. Thus,

$$p(x, C_1) = p(x|C_1)\, p(C_1) = \pi\, \mathcal{N}(x|\mu_1, \Sigma)$$
$$p(x, C_2) = p(x|C_2)\, p(C_2) = (1-\pi)\, \mathcal{N}(x|\mu_2, \Sigma)$$

The data likelihood function is thus

$$p(T|\mu_1,\mu_2,\Sigma) = \prod_{n=1}^{N} \left(\pi \mathcal{N}(x_n|\mu_1,\Sigma)\right)^{t_n} \left((1-\pi)\mathcal{N}(x_n|\mu_2,\Sigma)\right)^{1-t_n}$$

where $T = (t_1, \ldots, t_N)^T$.

$$l = \log p = \sum_{n=1}^{N} \{ t_n (\log\pi - \frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1)) + (1-t_n)(\log(1-\pi) - \frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu_2)^T\Sigma^{-1}(x_n-\mu_2)) \}$$

$$\frac{\partial l}{\partial \pi} = 0 = \sum_{n=1}^{N} \{ \frac{t_n}{\pi} - \frac{(1-t_n)}{1-\pi} \} = \sum_{n=1}^{N} \frac{(t_n-\pi)}{\pi(1-\pi)}$$

$$\pi = \frac{1}{N}\sum_{n=1}^{N} t_n = \frac{N_1}{N_1+N_2} //$$

$$\frac{\partial l}{\partial \mu_1} = 0 = \sum_{n=1}^{N} t_n(\Sigma^{-1}x_n - \Sigma^{-1}\mu_1)$$

$$\sum_{n=1}^{N} t_n\mu_1 = \sum_{n=1}^{N} t_n x_n, \quad \mu_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n x_n = \frac{1}{N_1}\sum_{x_n \in C_1} x_n //$$

$$\frac{\partial l}{\partial \mu_2} = 0 = \sum_{n=1}^{N} (1-t_n)(\Sigma^{-1}x_n - \Sigma^{-1}\mu_2)$$

$$\sum_{n=1}^{N} (1-t_n)\mu_2 = \sum_{n=1}^{N} (1-t_n)x_n, \quad \mu_2 = \frac{1}{N_2}\sum_{x_n \in C_2} x_n //$$

the terms involving Σ are

$$= \sum_{n=1}^{N} \{ t_n \left(-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1) \right) + (1-t_n)\left(-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu_2)^T\Sigma^{-1}(x_n-\mu_2) \right) \}$$

$$= -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{x_n \in C_1}(x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1) - \frac{1}{2}\sum_{x_n \in C_2}(x_n-\mu_2)^T\Sigma^{-1}(x_n-\mu_2)$$

$$= -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\mathrm{Tr}\left((N_1 S_1 + N_2 S_2)\Sigma^{-1}\right)$$

where we defined $\begin{cases} S_1 = \frac{1}{N_1}\sum_{x_n \in C_1}(x_n-\mu_1)(x_n-\mu_1)^T \\ S_2 = \frac{1}{N_2}\sum_{x_n \in C_2}(x_n-\mu_2)(x_n-\mu_2)^T \end{cases}$

Setting the $\frac{\partial}{\partial \Sigma}(\dots) = 0$ we have

$$-\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\left(\Sigma^{-1}(N_1 S_1 + N_2 S_2)\Sigma^{-1}\right)^T = 0$$

$$\Sigma^{-1}(N_1 S_1 + N_2 S_2)^T \Sigma^{-T} = N\Sigma^{-1}$$

Therefore $\Sigma = S$, $S = \frac{1}{N}(N_1 S_1 + N_2 S_2)$.

If we have training data we can replace this into our decision rule $a = w^T x + w_0 \gtrless 0$. This is the plug-in classifier.