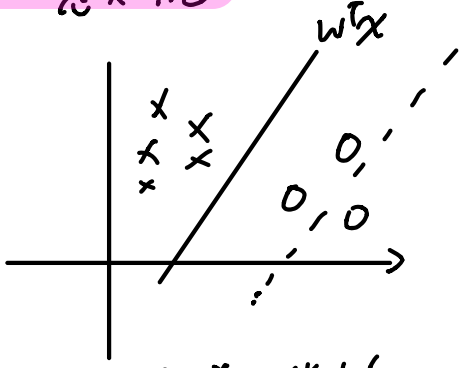
线性分类

- 硬输出 $y=\{+1,-1\}$
  - 感知机 (Perceptron)
  - 线性判别分析 (LDA / Linear Discriminant Analysis)
- 软输出 $P(y=1)=p$, $P(y=-1)=1-p$
  - 概率判别模型：Logistic Regression
  - 概率生成模型
    - Gaussian Discriminant Analysis (连续) ☆
    - Naive Bayes (离散)

一. 感知机



思想：错误驱动

模型：$f(x)=sign(w^Tx)$，$x\in R^p$，$w\in R^p$

$$sign(a)=\begin{cases}+1, & a\geq 0\\ -1, & a<0\end{cases}$$

D: 被错误分类的样本

样本集：$\{(x_i,y_i)\}_{i=1}^N$

策略：Loss Function：

$$L(w)=\sum_{i=1}^{N}\{y_i w^T x_i<0\}$$ 不可导

$$\hookrightarrow L(w)=\sum_{x_i\in D}-y_i w^T x_i$$

Loss Function：将随机事件或其有关随机变量的取值映射为非负数以表示该随机事件的"风险"或"损失"的函数.

求解感知机，可以用SGD算法.（见统计学习方法第2章）

二. 线性判别分析 (LDA)

思想 ① 相同类内部的训练样本距离越小越好

② 不同类别之间距离越大越好.

实质：找一个合适的投影方向

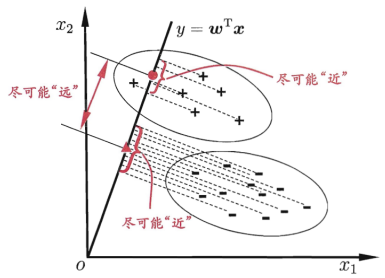
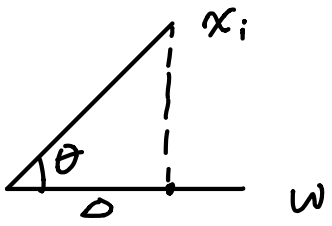


图 3.3 LDA 的二维示意图. "+"、"−"分别代表正例和反例, 椭圆表示数据簇的外轮廓, 虚线表示投影, 红色实心圆和实心三角形分别表示两类样本投影后的中心点.

$S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})$

- 样本在w上的投影



投影长度为 $\Delta = x_i \cos\theta$

当 $\|w\|=1$ 时，$w^T x = |x_i| \cdot |w| \cdot \cos\theta = |x_i| \cdot \cos\theta$

即 $\|w\|=1$ 时，$w^T x_i$ 为投影长度.

- 如何表示同种样本类间距离最小？ ⟶ 用方差表示.

由上可知投影长度为 $z_i = w^T x_i$

则平均投影长度为 $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} w^T x_i$

则方差为 $S = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^T$

$= \frac{1}{N}\sum_{i=1}^{N}(w^T x_i - \bar{z})(w^T x_i - \bar{z})^T$

设分为 C_1 和 C_2 两组。$\bar{z}_1, S_1$ 为组 C_1 的平均值和方差，

$\bar{z}_2, S_2$ 为 C_2 的 ~~~ 。

有 C_1 组：$\bar{z}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} w^T x_i$ ①

$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(w^T x_i - \bar{z}_1)(w^T x_i - z_1)^T$ ②

联立①②：

$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(w^T x_i - \frac{1}{N}\sum_{j=1}^{N_1} w^T x_j)(w^T x_i - \frac{1}{N}\sum_{j=1}^{N_1} w^T x_j)^T$

$= w^T \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \frac{1}{N}\sum_{j=1}^{N_1} x_j)(x_i - \frac{1}{N}\sum_{j=1}^{N_1} x_j)^T$

$= w^T \left[\frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \overline{x_{c1}})(x_i - \overline{x_{c1}})^T\right] w$

x_{c1} 的方差

$S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})$

$S_1 = W^T S_{c1} W$

同理，C_2 组算得 $S_2 = W^T S_{c2} W$

两个类的类内分散度之和.

则类内距离可以表示成 $S_1 + S_2 = W^T (S_{c1} + S_{c2}) W$

- 如何表示不同样本类间距离最大？ ⟶ 用平均值表示.

$$d = (\bar{z}_1 - \bar{z}_2)^2$$

$$= \left(\frac{1}{N_1}\sum_{i=1}^{N_1} W^T x_i - \frac{1}{N_2}\sum_{i=1}^{N_2} W^T x_i\right)^2$$

$$= \left(W^T\left(\frac{1}{N}\sum_{i=1}^{N_1} x_i - \frac{1}{N_2}\sum_{i=1}^{N_2} x_i\right)\right)^2$$

$$= \left[W^T(\bar{x}_{c1} - \bar{x}_{c2})\right]^2$$

$$d = W^T(\bar{x}_{c1} - \bar{x}_{c2})(\bar{x}_{c1} - \bar{x}_{c2}) W$$

- 综合类间大，类内小思想，可得目标函数

$$J(W) = \frac{(\bar{z}_1 - \bar{z}_2)^2}{S_1 + S_2}$$

$$= \frac{W^T(\bar{x}_{c1} - \bar{x}_{c2})(\bar{x}_{c1} - \bar{x}_{c2}) W}{W^T(S_{c1} + S_{c2}) W}$$

$$\hat{W} = \arg\max_W J(W)$$

令 $S_b = (\bar{x}_{c1} - \bar{x}_{c2})(\bar{x}_{c1} - \bar{x}_{c2})^T$

$S_W = S_{c1} + S_{c2}$

$$\therefore \hat{W} = \arg\max_W \frac{W^T S_b W}{W^T S_W W}$$

矩阵求导公式

$$\frac{\partial x^T A x}{\partial x} = 2Ax$$

$$\Rightarrow \frac{\partial J(W)}{}$$

$J(w)$

方法一

$$\Rightarrow \frac{\partial}{\partial w} \quad w^T S_b w \cdot (w^T S_w w)^{-1} = 0$$

$$2 S_b w (w^T S_w w)^{-1} - 2 w^T S_b w (w^T S_w w)^{-2} \cdot S_w w = 0$$

$$\Rightarrow S_b w (w^T S_w w)^{-1} = w^T S_b w (w^T S_w w)^{-2} S_w w$$

$$\Rightarrow \underbrace{w^T S_b w}_{R} S_w w = S_b w \underbrace{w^T S_w w}_{R}$$

标量同乘

$w \to p \times 1$
$w^T \to 1 \times p$
$S_w \to p \times p$

$$\Rightarrow S_w w = \frac{w^T S_w w}{w^T S_b w} \cdot S_b w$$

$$w = \frac{w^T S_w w}{w^T S_b w} \cdot S_w^{-1} \cdot S_b w$$

$$\propto S_w^{-1} \underbrace{S_b \cdot w}$$

$$(\overline{x_{c1}} - \overline{x_{c2}})\underbrace{(\overline{x_{c1}} - \overline{x_{c2}})^T \cdot w}_{R}$$

$$\Rightarrow \propto \boxed{S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})}$$

就是 w 的方向

方法二（西瓜书）

$$J = \frac{w^T S_b w}{w^T S_w w}$$

因为 J 的解 w 的长度无关，只与它的方向有关，令 $w^T S_w w = 1$

故问题转化为最优化问题

$$\min \quad -w^T S_b w$$
$$\text{s.t.} \quad w^T S_w w = 1$$

$$\frac{\partial x^T A x}{\partial x} = (A + A^T) x$$

对称：$\frac{\partial x^T A x}{\partial x} = 2Ax$

$$\Rightarrow L(w, \lambda) = -w^T S_b w + \lambda (w^T S_w w - 1)$$

$$\frac{\partial L(w,\lambda)}{\partial w} = -(S_b + S_b^T)w + \lambda(S_w + S_w^T)w$$

又：$S_b = S_b^T$，$S_w = S_w^T$（对称）

$$\therefore \frac{\partial L(w,\lambda)}{\partial w} = -2S_b w + 2\lambda S_w w = 0$$

$$\Rightarrow \underline{S_b w} = \lambda S_w w$$

$$(\overline{x_{c1}} - \overline{x_{c2}})(\overline{x_{c1}} - \overline{x_{c2}})^T w = \lambda S_w w$$

令 $(\overline{x_{c1}} - \overline{x_{c2}})^T w = \varphi$

有 $\varphi(\overline{x_{c1}} - \overline{x_{c2}}) = \lambda S_w w$

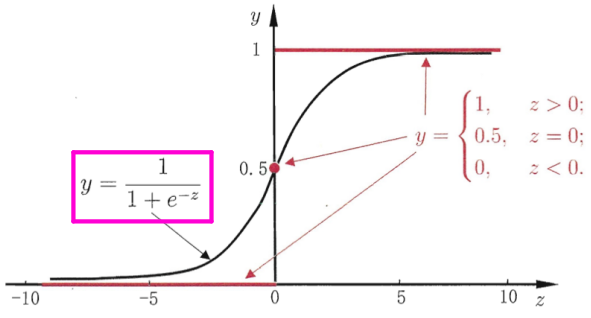
$$w = \frac{\varphi}{\lambda} S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})$$

∵ w 不必关心其值，取 $\frac{\varphi}{\lambda} = 1$，得 $\boxed{w = S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})}$

三. logistics Regression

- 原理：线性回归 $w^T x$ —激活函数→ 线性分类 $(0, 1)$

- logistics Regression 使用的激活函数为 Sigmoid function：$\sigma(z) = \frac{1}{1+e^{-z}}$



Sigmoid Function

Data：$\{(x_i, y_i)\}_{i=1}^{n}$

$x_i \in R^p$，$y_i \in (0,1)$

$\boxed{P_1 = P(y=1|x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}}$，$y=1$

$P_0 = P(y=0|x) = 1 - P_1 = 1 - \frac{1}{1+e^{-w^T x}} = \frac{e^{-w^T x}}{1+e^{-w^T x}}$

$P(y|x) = P_1^{\,y} P_0^{\,1-y}$

即可用 MLE 估计 w 的最大值.

MLE: $\arg\max_w \ln P(y|x)$

$= \arg\max_w \ln \prod_{i=1}^{N} P(y_i|x_i)$

$= \arg\max_w \sum_{i=1}^{N} \ln (P_1^{\,y_i} P_0^{\,1-y_i})$

$= \arg\max_w \sum_{i=1}^{N} (y_i \log P_1 + (1-y_i) \ln P_0)$ (Cross entropy) (P_0 处 $1-P_1$)

$= \arg\max_w \sum_{i=1}^{N} (y_i \ln P_1 + \ln P_0 - y_i \ln P_0)$

$= \arg\max_w \sum_{i=1}^{N} (y_i (\ln \frac{P_1}{P_0}) + \ln P_0)$

将 $P_1 = \frac{1}{1+e^{-w^T x}}$, $P_0 = \frac{e^{-w^T x}}{1+e^{-w^T x}}$ 代入上式,

得: $\arg\max_w \sum_{i=1}^{N} (y_i \ln (e^{w^T x}) + \ln (\frac{1}{1+e^{w^T x}})$

$= \arg\max_w \sum_{i=1}^{N} y_i w^T x - \ln (1+e^{w^T x})$

$\Rightarrow \arg\min_w \sum_{i=1}^{N} [-y_i w^T x + \ln (1+e^{w^T x})]$

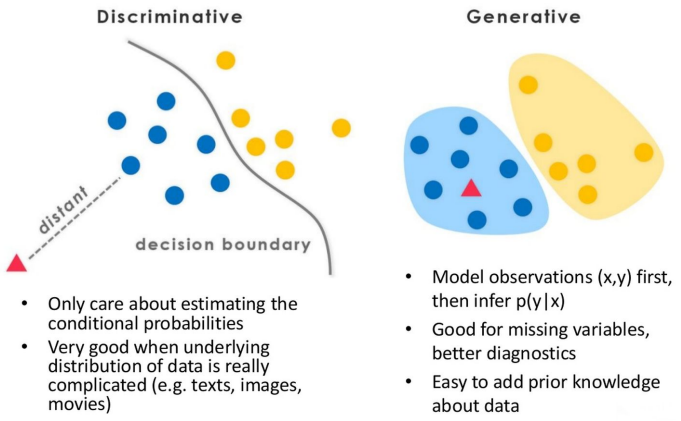
求解此问题. 可以用 gradient descent method 或 Newton method

四. 高斯判别分析 (GDA) 生成模型

- 判别模型和生成模型的区别

判别模型: 求得 $P(Y|X)$. 对未见示例 X. 根据 $P(Y|X)$ 可以求得标记 Y (对二分类.

Discriminative vs. Generative



- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

- Model observations (x,y) first, then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

其实是得到一个score，若score > threshold，则为正类）。常见的模型有 logistics Regression、SVM。

生成模型：求得 $P(Y,X)$。对未见示例 x，你要求出 x 与各个同标记之间的联合概率分布，然后大的那个获胜。常见模型：GDA、HMM、Naive Bayes

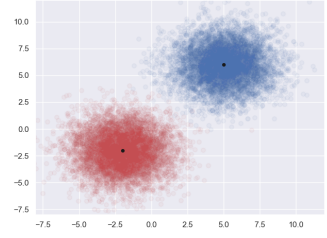
即对联合概率 $P(x,y)$ 的直接建模 ⟶

$$P(y|x) \propto \underbrace{P(x|y)\cdot P(y)}_{\text{似然}\quad\text{先验}} \longrightarrow P(x,y)$$

（$P(y|x)$：后验）

$$\hat{y} = \arg\max_{y\in\{0,1\}} P(y|x) \approx \arg\max_{y} P(y)\cdot P(x|y)$$

最大后验概率 (MAP)



· GDA 模型定义

· 定义 $y \in Bernoulli(\phi) \Rightarrow \phi^y(1-\phi)^{1-y}$

$$\left.\begin{array}{ll} x|y=1 & \sim N(\mu_1,\Sigma) \\ x|y=0 & \sim N(\mu_2,\Sigma) \end{array}\right\} \Rightarrow N(\mu_1,\Sigma)^y \cdot N(\mu_2,\Sigma)^{1-y}$$

强假设：基于不同分类的条件概率满足高斯分布，它们拥有不同的均值，但它们的方差是相同的（$P(x|y=C_1)$ 是一系列的值和但协方差相同的矩阵！）

$$\text{log-likelihood}: \quad L(\theta) = \log \prod_{i=1}^{N} P(x_i, y_i)$$

$$= \prod_{i=1}^{N} \log\left[P(y_i)P(x_i|y_i)\right]$$

$$= \prod_{i=1}^{N} \log P(y_i) + \log P(x_i|y_i)$$

$y_i \sim B(\phi)$

$$= \prod_{i=1}^{N} \log\left(\phi^{y_i}(1-\phi)^{1-y_i}\right) + \log\left(N(\mu_1,\Sigma)^{y_i}\cdot N(\mu_2,\Sigma)^{1-y_i}\right)$$

$$L(\theta) = \prod_{i=1}^{N} \left(\log \phi^{y_i} + \log(1-\phi)^{1-y_i} + \log\left(N(\mu_1,\Sigma)^{y_i}\right) + \log\left(N(\mu_2,\Sigma)^{1-y_i}\right)\right.$$

$\theta = (\mu_1, \mu_2, \Sigma, \phi)$

$\hat{\theta} = \arg\max_{\theta} L(\theta)$

设 $y=1 \Rightarrow N_1$.
$y=0 \Rightarrow N_2$
$N = N_1 + N_2$

- GDA 模型的求解

  - 首先是对 ϕ 求解，有 ϕ 的只有前两项，对其求偏导.

即 $\frac{\partial}{\partial \phi}\left[\sum_{i=1}^{N} y_i \log\phi + (1-y_i)\log(1-\phi)\right]$

$$= \sum_{i=1}^{N}\left(\frac{y_i}{\phi} + \frac{1-y_i}{1-\phi}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N}\left[(1-\theta)y_i - \theta(1-y_i)\right] = 0$$

$$\Rightarrow \sum_{i=1}^{N}(y_i - \theta) = 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i - N\theta = 0$$

$$\Rightarrow \boxed{\theta = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{N_1}{N}}$$

  - 然后对 μ_1 求解.

$$\sum_{i=1}^{N} y_i \log N(\mu_1, \Sigma)$$

$$\sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right)$$

（$\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}$ 无关）

$$\sum_{i=1}^{N} y_i \left(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right)$$

求要求 $\hat{\mu}_1 = \arg\max_{\mu_1} \sum_{i=1}^{N} y_i \left(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right) \Rightarrow \Delta$

$$\Delta = -\frac{1}{2}\sum_{i=1}^{N} y_i (x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)$$

$(A\pm B)^T = A^T \pm B^T$

$$= -\frac{1}{2}\sum_{i=1}^{N} y_i \left(x_i^T \Sigma^{-1} - \mu_1^T \Sigma^{-1} \right) (x_i - \mu_1)$$

(实数)^T = 实数

$$= -\frac{1}{2}\sum_{i=1}^{N} y_i \left(x_i^T \Sigma^{-1} x_i - \underbrace{x_i^T \Sigma^{-1} \mu_1}_{1\times p \times p\times p\times p\times 1 \text{ 为实数}} - \underbrace{\mu_1^T \Sigma^{-1} x_i}_{(x_i^T \Sigma^{-1}\mu_1)^T} + \mu_1^T \Sigma^{-1} \mu_1 \right)$$

$$= -\frac{1}{2}\sum_{i=1}^{N} y_i \left(\underbrace{x_i^T \Sigma^{-1} x_i}_{1\times p\times p\times p\times p\times 1 \text{ 为 } R} - 2\mu_1^T \Sigma^{-1} x_i + \underbrace{\mu_1^T \Sigma^{-1} \mu_1}_{②} \right)$$

① ②

$$\frac{\partial \ell}{\partial \mu_1} = -\frac{1}{2}\sum_{i=1}^{N} y_i \left(\underbrace{-2\Sigma^{-1} x_i - 2\Sigma^{-1}\mu_1}_{①} \right) = 0$$

$\frac{\partial 2x^T A}{\partial x} = A$ ①

$\frac{\partial 2x^T A x}{\partial x} = 2Ax$, 当 A is symmetric

$$= \sum_{i=1}^{N} y_i \left(\Sigma^{-1} x_i + \Sigma^{-1} \mu_1 \right) = 0$$

$$= \sum_{i=1}^{N} y_i (x_i - \mu_1) = 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i x_i = \sum_{i=1}^{N} \mu_1 y_i$$

$$\Rightarrow \boxed{\mu_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} x_i y_i}{N_1}}$$

同对称性. $\boxed{\mu_2 = \frac{\sum_{i=1}^{N} (1-y_i) x_i}{N_2}}$

最后对 Σ 求解:

$$L(\theta) = \sum_{i=1}^{N} \left(\log \phi^{y_i} + \log(1-\phi)^{1-y_i} + \underbrace{\log\left(N(\mu_1,\Sigma)^{y_i}\right)}_{①} + \underbrace{\log\left(N(\mu_2,\Sigma)^{1-y_i}\right)}_{②} \right)$$

有 Σ 的只有后两项. $\hat{\Sigma} = \text{argmax}$ ① + ②

$$= \text{argmax} \quad y_i \log N(\mu_1,\Sigma) + (1-y_i)\log N(\mu_2,\Sigma)$$

记 $C_1 = \{x_i \mid y_i = 1, x_i = 1, \dots N\}$

$C_2 = \{x_i \mid y_i = 0, x_i = 1, \dots, N\}$

$C_1 = |N_1|$, $C_2 = |N_2|$. $N_1 + N_2 = N$

则有 ① + ② $= \underbrace{\sum_{x \in C_1} \log N(\mu_1, \Sigma)}_{①} + \sum_{x \in C_2} \log N(\mu_2, \Sigma)$

先看①：

$$\sum_{i=1}^{N} \log N(\mu_1, \Sigma)$$

$$= \sum_{i=1}^{N} \log \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right\}$$

$$= \sum_{i=1}^{N} \log \underbrace{\frac{1}{(2\pi)^{\frac{p}{2}}}}_{C} - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

$$= C - N\frac{1}{2}\log|\Sigma| - \sum_{i=1}^{N} \frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

$1 \times p \quad p \times p \quad p \times 1 \Rightarrow R$

$\because \sum_{i=1}^{N} tr\left[(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right]$

$tr(ABC) = tr(CAB) = trCBA)$

$$= \sum_{i=1}^{N} tr\left[(x_i - \mu)(x_i - \mu)^T \Sigma^{-1}\right]$$

$$= tr\left[\underbrace{\sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T}_{NS} \Sigma^{-1}\right]$$

$S = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$

$\therefore$ 原式 $= -\frac{1}{2} N \log|\Sigma| - \frac{1}{2} N \cdot tr(S \cdot \Sigma^{-1}) + C$

① + ② $\Rightarrow -\frac{1}{2} N_1 \log|\Sigma| - \frac{1}{2} N_1 \cdot tr(S_1 \cdot \Sigma^{-1}) + C$

$-\frac{1}{2} N_2 \log|\Sigma| - \frac{1}{2} N_2 tr(S_2 \Sigma^{-1}) + C$

$$L(\Sigma) = -\frac{1}{2} N \log|\Sigma| - \frac{1}{2} N_1 tr(S_1 \cdot \Sigma^{-1}) - \frac{1}{2} N_2 tr(S_2 \cdot \Sigma^{-1}) + C$$

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

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

$\frac{\partial tr(AB)}{\partial A} = B^T$

$\frac{\partial |A|}{\partial A} = |A| \cdot A^{-1}$

$$\Rightarrow \Sigma = \frac{N_1 S_1 + N_2 S_2}{N}$$