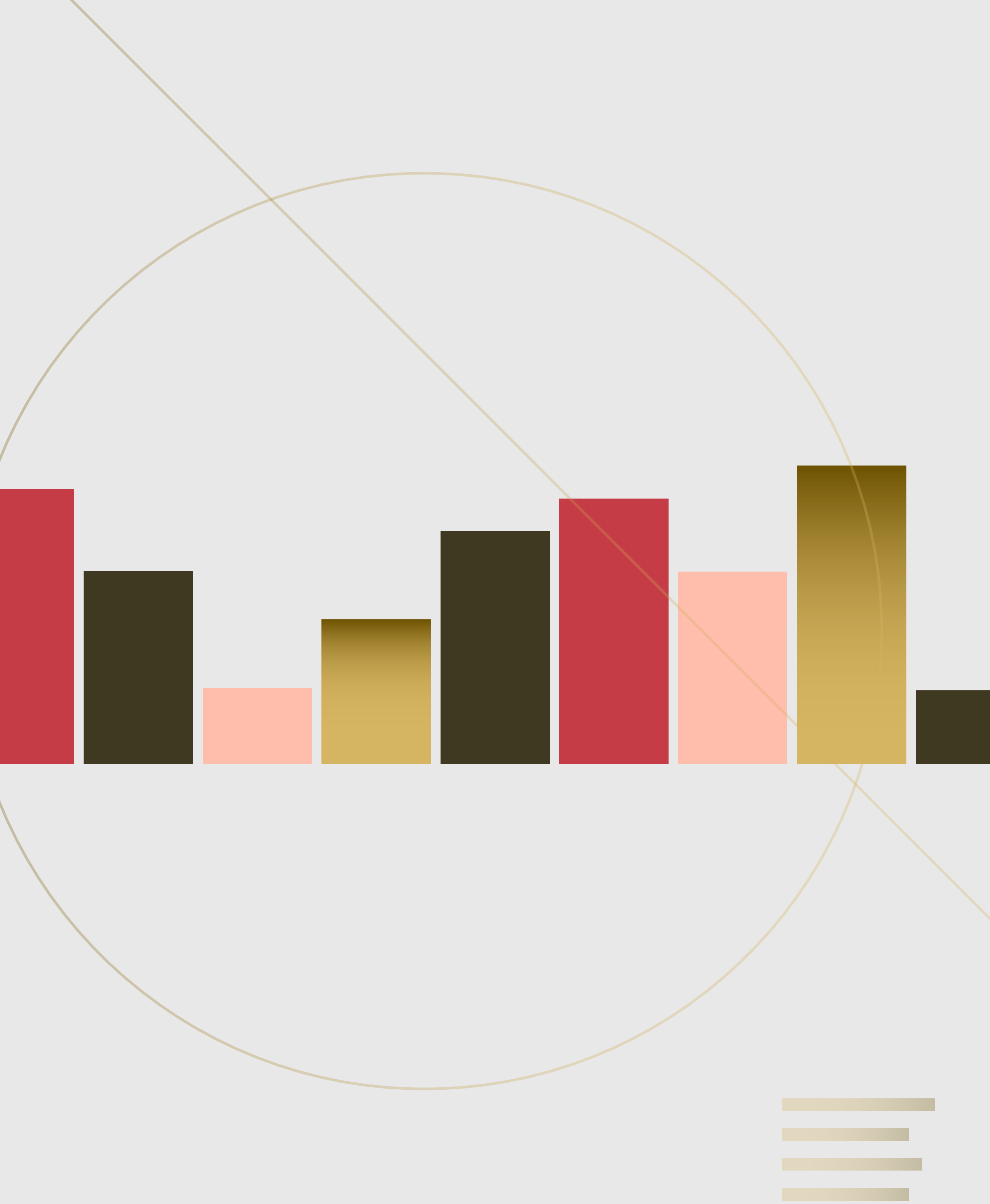


一、线性回归

线性回归的目的是试图学得一线性模型，即

$$f(x_i)=wx_i+b，使得 f(x_i)\simeq y_i$$

我们需要确定 w 和 b，称为均方误差，试图将均方误差最小化，即

$$(w^*,b^*)=\underset{(w,b)}{\arg\min}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

$$=\underset{(w,b)}{\arg\min}\sum_{i=1}^{m}(y_i-wx_i-b)^2$$

使用最小二乘法来求解均方误差最小化。求解 w 和 b 使 $E(w,b)=\sum_{i=1}^{m}(y_i-wx_i-b)^2$ 最小化的过程，称为线性回归模型的最小二乘"参数估计"。

我们将 $E(w,b)$ 分别对 w 和 b 求导，得：

$$\frac{\partial E(w,b)}{\partial w}=2\sum_{i=1}^{m}(y_i-wx_i-b)(-x_i)$$

$$=2\left(w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}(y_i-b)x_i\right)$$

$$\frac{\partial E(w,b)}{\partial b}=2\sum_{i=1}^{m}(y_i-wx_i-b)(-1)$$

$$=2\left(mb-\sum_{i=1}^{m}(y_i-wx_i)\right)$$

令 $\frac{\partial E(w,b)}{\partial w}=0$，$\frac{\partial E(w,b)}{\partial b}=0$，得

$$\frac{\partial E(w,b)}{\partial b}=2\left(mb-\sum_{i=1}^{m}(y_i-wx_i)\right)=0$$

$$mb=\sum_{i=1}^{m}(y_i-wx_i)$$

$$b=\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)$$

$$\frac{\partial E(w,b)}{\partial w}=2\left(w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}(y_i-b)x_i\right)=0$$

$$w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}y_ix_i+\sum_{i=1}^{m}bx_i=0$$

$$w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}y_ix_i+\sum_{i=1}^{m}x_i\left(\frac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)\right)=0$$

$$w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}y_ix_i+\sum_{i=1}^{m}x_i\left(\frac{1}{m}\sum_{i=1}^{m}y_i-\frac{1}{m}\sum_{i=1}^{m}wx_i\right)=0$$

$$w\sum_{i=1}^{m}x_i^2-\frac{w}{m}\left(\sum_{i=1}^{m}x_i\right)^2=\sum_{i=1}^{m}y_ix_i-\sum_{i=1}^{m}y_i\bar{x}$$

$$w=\frac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2}$$

多元线性回归

更一般地，$f(x_i)=w^Tx_i+b$，使 $f(x_i)\simeq y_i$

同样的，利用最小二乘法对 w,b 进行估计。便于讨论，将 w 和 b 写成向量形式 $\hat{w}=(w;b)$，将数据集 D 表示为一个 $m\times(d+1)$ 大小的矩阵 X，即

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} & 1 \\ x_{21} & x_{22} & \cdots & x_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix}=\begin{pmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{pmatrix}$$

故此时，有 $\hat{w}^*=\arg\min_{\hat{w}}(y-X\hat{w})^T(y-X\hat{w})$

令 $E(\hat{w})=(y-X\hat{w})^T(y-X\hat{w})$

$$\begin{aligned}E(\hat{w})&=(y-X\hat{w})^T(y-X\hat{w})\\&=(y^T-\hat{w}^TX^T)(y-X\hat{w})\\&=y^Ty-y^TX\hat{w}-\hat{w}^TX^Ty+\hat{w}^TX^TX\hat{w}\end{aligned}$$

对 $\hat{w}$ 求导，得

$$\frac{\partial E(\hat{w})}{\partial\hat{w}}=\frac{\partial(y^Ty-y^TX\hat{w}-\hat{w}^TX^Ty+\hat{w}^TX^TX\hat{w})}{\partial\hat{w}}$$

$$\frac{\partial\hat{w}^TX^TX\hat{w}}{\partial\hat{w}}=\frac{\partial(X^TX)(\hat{w}_1^2+\hat{w}_2^2+\dots+\hat{w}_{d+1}^2)}{\partial\hat{w}_i}=2X^TX\hat{w}$$

$$\frac{\partial w^TX^Ty}{\partial\hat{w}}=X^Ty \qquad \frac{\partial y^TX\hat{w}}{\partial\hat{w}}=X^Ty$$

$$\frac{\partial E(\hat{w})}{\partial\hat{w}}=2X^TX\hat{w}-X^Ty-X^Ty$$

如果X^TX为满秩矩阵或正定矩阵，令$\frac{\partial E(\hat{w})}{\partial \hat{w}}=0$，有

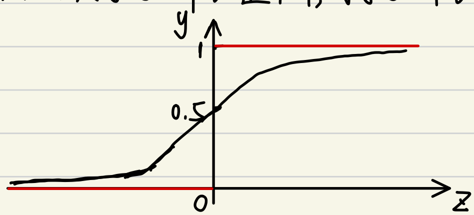
$$\hat{w}=(X^TX)^{-1}X^Ty$$

二、对数几率回归

对数几率回归常用于分类任务。首先先考虑二分类任务，其输出标记$y\in\{0,1\}$，而线性回归产生的预测值$z=w^Tx+b$是实值，我们需要将z转换为0/1值。我们采用单位阶跃函数。

$$y=\begin{cases}0 & z<0\\ 0.5 & z=0\\ 1 & z>0\end{cases}$$

即若预测值z大于0则为正例，小于0则为反例，等于0则可任意判别



单位阶跃函数与对数几率函数

从上图可以得知单位阶跃函数不连续，故我们采用一个常用的替代函数

$$y=\frac{1}{1+e^{-z}}$$

又$z=w^Tx+b$

此时

$$y=\frac{1}{1+e^{-(w^Tx+b)}}$$

$$\frac{1}{y}=1+e^{-(w^Tx+b)}$$

$$\frac{1-y}{y}=e^{-(w^Tx+b)}$$

$$\ln\frac{1-y}{y}=-(w^Tx+b)$$

$$\ln\frac{y}{1-y}=w^Tx+b$$

若将y视为样本x作为正例的可能性，则1-y是其反例的可能，两者的比值称为"几率"，反映了x作为正例的可能性. 对几率取对数则得到"对数几率"：

$$\ln\frac{y}{1-y}$$

此时我们来确定前面的w和b. 如果将 $y=\frac{1}{1+e^{-(w^Tx+b)}}$ 中的y视为类后验概率估计 $p(y=1|x)$，则

$$\ln\frac{p(y=1|x)}{p(y=0|x)}=w^Tx+b$$

那么有

$$p(y=1|x)=\frac{e^{w^Tx+b}}{1+e^{w^Tx+b}}$$

$$p(y=0|x)=\frac{1}{1+e^{w^Tx+b}}$$

于是，我们可以通过极大似然法来估计w和b. 给定数据集 $\{(x_i,y_i)\}_{i=1}^m$，对率回归模型最大化"对数似然"

$$\ell(w,b)=\sum_{i=1}^m\ln p(y_i|x_i;w,b)$$

为了便于讨论，令 $\beta=(w;b)$，$\hat{x}=(x;1)$，则 w^Tx+b 可简化为 $\beta^T\hat{x}$，再令 $p_1(\hat{x};\beta)=p(y=1|\hat{x};\beta)$，$p_0(\hat{x};\beta)=p(y=0|\hat{x};\beta)=1-p_1(\hat{x};\beta)$，则

$$p(y_i|x_i;w,b)=y_ip_1(\hat{x}_i;\beta)+(1-y_i)p_0(\hat{x}_i;\beta)$$

$$\ell(\beta)=\sum_{i=1}^m\ln\left(y_ip_1(\hat{x}_i;\beta)+(1-y_i)p_0(\hat{x}_i;\beta)\right)$$

$$=\sum_{i=1}^m\left(-y_i\beta^T\hat{x}_i+\ln(1+e^{\beta^T\hat{x}_i})\right)$$

推导：因为 $p_1(\hat{x}_i;\beta)=p(y=1|\hat{x};\beta)=\frac{e^{w^Tx+b}}{1+e^{w^Tx+b}}$

$$p_0(\hat{x}_i;\beta)=p(y=0|\hat{x};\beta)=\frac{1}{1+e^{w^Tx+b}}$$

$$\ell(\beta)=\sum_{i=1}^m\ln\left(\frac{y_ie^{\beta^T\hat{x}_i}+1-y_i}{1+e^{\beta^T\hat{x}_i}}\right)$$

$$\ell(\beta)=\begin{cases}\sum_{i=1}^m\left(-\ln(1+e^{\beta^T\hat{x}_i})\right) & y_i=0\\ \sum_{i=1}^m\left(\beta^T\hat{x}_i-\ln(1+e^{\beta^T\hat{x}_i})\right) & y_i=1\end{cases}$$

综合可得

$$\ell(\beta)=\sum_{i=1}^m\left(y_i\beta^T\hat{x}_i-\ln(1+e^{\beta^T\hat{x}_i})\right)$$

所得出来的 $l(\beta)$ 是关于 β 的高阶可导连续凸函数，于是有

$$\beta^* = \arg\min_{\beta} l(\beta)$$

以牛顿法为例求解，其第 $t+1$ 轮迭代解的更新公式为

$$\beta^{t+1} = \beta^t - \left(\frac{\partial^2 l(\beta)}{\partial\beta\,\partial\beta^T}\right)^{-1} \frac{\partial l(\beta)}{\partial\beta}$$

牛顿法原理：

牛顿法是一种线性化方法，其基本思想是将非线性方程 $f(x)=0$ 逐步归结为某种线性方程来求解。

设已知方程 $f(x)=0$ 有近似根 x_k（假定 $f'(x_k)\neq 0$），将函数 $f(x)$ 在点 x_k 展开，有

$$f(x) \approx f(x_k) + f'(x_k)(x-x_k)$$

于是方程 $f(x)=0$ 可近似地表示为

$$f(x_k) - f'(x_k)(x-x_k) = 0$$

这是个线性方程，记其根 x_{k+1}，则 x_{k+1} 的计算公式为

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\quad k=0,1,\cdots,$$

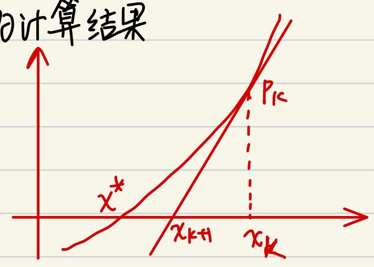
牛顿法的几何解释：

方程 $f(x)=0$ 的根 x^* 可解释为曲线 $y=f(x)$ 与 x 轴的交点的横坐标。设 x_k 是根 x^* 的某个近似值，过曲线 $y=f(x)$ 上横坐标为 x_k 的点 P_k 引切线，并将该切线与 x 轴的交点的横坐标 x_{k+1} 作为 x^* 的新的近似值。

注意到切线方程为

$$y = f(x_k) + f'(x_k)(x-x_k)$$

此时求得的值 x_{k+1} 就是牛顿公式的计算结果



设$f(x)$二次连续可微，$x_k \in R^n$，Hesse矩阵$\nabla^2 f(x_k)$正定。我们在x_k附近用二次泰勒展开近似$f(x)$.

$$f(x)=f(x_k)+\nabla f(x_k)^T(x-x_k)+\frac{1}{2}(x-x_k)^T\nabla^2 f(x_k)(x-x_k)$$

令$s=x-x_k$　　$x=s+x_k$

$$f(s+x_k)\approx f(x_k)+\nabla f(x_k)^T s+\frac{1}{2}s^T\nabla^2 f(x_k)s$$

令$q^k(s)=f(x_k)+\nabla f(x_k)^T s+\frac{1}{2}s^T\nabla^2 f(x_k)s$为$f(x)$的二次近似，求解自变量迭代增量$s$使$q^k(s)$极小化，即对$q^k(s)$求导并令导数等于0：

$$\nabla q^k(s)=\nabla f(x_k)+\nabla^2 f(x_k)s=0$$

$$\Rightarrow\quad s=-[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$$

代入 $x=s+x_k$，有 $x_{k+1}=x_k-[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$

对应到式中

$$\beta^t=x_k,\quad [\nabla^2 f(x_k)]^{-1}=\left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}\right)^{-1}\quad \nabla f(x_k)=\frac{\partial l(\beta)}{\partial\beta}$$

其关于β的一阶、二阶导数分别为

$$\frac{\partial l(\beta)}{\partial\beta}=-\sum_{i=1}^{m}\hat{x}_i(y_i-p_1(\hat{x}_i;\beta))$$

推导：

$$\begin{aligned}\frac{\partial l(\beta)}{\partial\beta}&=\frac{\partial\sum_{i=1}^{m}(-y_i\beta^T\hat{x}_i+\ln(1+e^{\beta^T\hat{x}_i}))}{\partial\beta}\\&=\sum_{i=1}^{m}\left(\frac{\partial(-y_i\beta^T\hat{x}_i)}{\partial\beta}+\frac{\partial(\ln(1+e^{\beta^T\hat{x}_i}))}{\partial\beta}\right)\\&=\sum_{i=1}^{m}\left(-y_i\hat{x}_i+\frac{1}{1+e^{\beta^T\hat{x}_i}}\cdot e^{\beta^T\hat{x}_i}\cdot\hat{x}_i\right)\\&=-\sum_{i=1}^{m}\hat{x}_i\left(y_i-\frac{1}{1+e^{\beta^T\hat{x}_i}}\cdot e^{\beta^T\hat{x}_i}\right)\\&=-\sum_{i=1}^{m}\hat{x}_i(y_i-p_1(\hat{x}_i;\beta))\end{aligned}$$

$$\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}=\sum_{i=1}^{m}\hat{x}_i\hat{x}_i^T p_1(\hat{x}_i;\beta)(1-p_1(\hat{x}_i;\beta))$$

推导：

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}&=\frac{-\partial\sum_{i=1}^{m}\hat{x}_i\left(y_i-\frac{e^{\beta^T\hat{x}_i}}{1+e^{\beta^T\hat{x}_i}}\right)}{\partial\beta^T}\\&=-\sum_{i=1}^{m}\hat{x}_i\frac{\partial\left(y_i-\frac{e^{\beta^T\hat{x}_i}}{1+e^{\beta^T\hat{x}_i}}\right)}{\partial\beta^T}\\&=-\sum_{i=1}^{m}\hat{x}_i\left(\frac{\partial y_i}{\partial\beta^T}-\frac{\partial\left(\frac{e^{\beta^T\hat{x}_i}}{1+e^{\beta^T\hat{x}_i}}\right)}{\partial\beta^T}\right)\end{aligned}$$

第一项 $\frac{\partial y_1}{\partial \beta}=0$，第二项 $\frac{\partial\left(\frac{e^{\beta^T\hat{x}_i}}{1+e^{\beta^T\hat{x}_i}}\right)}{\partial \beta^T}=\frac{e^{\beta^T\hat{x}_i}\cdot\hat{x}_i^T(1+e^{\beta^T\hat{x}_i})-e^{\beta^T\hat{x}_i}\cdot e^{\beta^T\hat{x}_i}\cdot\hat{x}_i^T}{(1+e^{\beta^T\hat{x}_i})^2}$

$$=\hat{x}_i^T e^{\beta^T\hat{x}_i}\cdot\frac{(1+e^{\beta^T\hat{x}_i})-e^{\beta^T\hat{x}_i}}{(1+e^{\beta^T\hat{x}_i})^2}$$

$$=\hat{x}_i^T\cdot\frac{e^{\beta^T\hat{x}_i}}{1+e^{\beta^T\hat{x}_i}}\cdot\frac{1}{1+e^{\beta^T\hat{x}_i}}$$

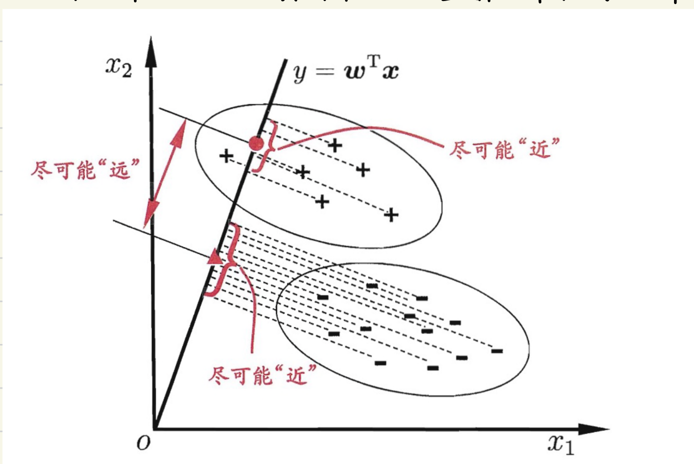
则 $\frac{\partial^2 l(\beta)}{\partial\beta\,\partial\beta^T}=-\sum_{i=1}^m\hat{x}_i\left(-\hat{x}_i^T\cdot\frac{1}{1+e^{\beta^T\hat{x}_i}}\cdot\frac{e^{\beta^T\hat{x}_i}}{1+e^{\beta^T\hat{x}_i}}\right)$

$$=\sum_{i=1}^m\hat{x}_i\hat{x}_i^T p_1(\hat{x}_i;\beta)(1-p_1(\hat{x}_i;\beta))$$

三、线性判别分析

LDA既是一种二分类算法，同时也是一种降维方法。

线性判别分析（LDA）的思想是：给定训练集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近、异类样例的投影点尽可能远离；在对新样本分类时，将其投影到同样的这条直线上，再根据投影点的位置来确定新样本的类别。



LDA示意图.

协方差的定义：用于衡量两个变量的总体误差。期望值分别为E(X)与E(Y)的两个实随机变量X与Y之间的协方差 Cov(X,Y) 定义为：

$$\mathrm{Cov}(X,Y)=E(X-E(X))(Y-E(Y))$$
$$=E(XY)-2EXEY+EXEY$$
$$=E(XY)-EXEY$$

设 $X=(X_1,X_2,\cdots,X_n)^T$ 为n维随机变量，称矩阵 $C=(C_{ij})_{n\times n}=\begin{pmatrix}C_{11}&C_{12}&\cdots&C_{1n}\\C_{21}&C_{22}&\cdots&C_{2n}\\\vdots&\vdots&&\vdots\\C_{n1}&C_{n2}&\cdots&C_{nn}\end{pmatrix}$ 为n维随机变量X的协方差矩阵，也记为D(X)，其中 $C_{ij}=\mathrm{Cov}(X_i,X_j)$，$i,j=1,2,\cdots,n$ 为X的分量 X_i 和 X_j 的协方差。由于 $C_{ij}=C_{ji}$，所以协方差矩阵为对称非负定矩阵。

给定数据集 $D=\{(x_i,y_i)\}_{i=1}^m$，$y_i\in\{0,1\}$，令 X_i，μ_i，Σ_i 分别表示第 $i\in\{0,1\}$ 类示例的集合、均值向量、协方差矩阵。若将数据投影到直线w上，则两类样本的中心在直线上的投影分别为 $w^T\mu_0$ 和 $w^T\mu_1$，两类样本的协方差分别为 $w^T\Sigma_0 w$ 和 $w^T\Sigma_1 w$。由于直线是一维空间，因此 $w^T\mu_0$、$w^T\mu_1$、$w^T\Sigma_0 w$ 和 $w^T\Sigma_1 w$ 均为实数。

如果想要使同类样例的投影点尽可能接近，可以让同类样例投影点的协方差尽可能小，即$w^T\Sigma_0 w + w^T\Sigma_1 w$尽可能小；而想要使异类样例的投影点尽可能远离，可以让类中心之间的距离尽可能大，即$\|w^T\mu_0 - w^T\mu_1\|_2^2$尽可能大。即我们同时考虑二者，则得到欲最大化的目标：

$$J = \frac{\|w^T\mu_0 - w^T\mu_1\|_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w}$$
$$= \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T(\Sigma_0+\Sigma_1)w}$$

推导：
$$J = \frac{\|w^T\mu_0 - w^T\mu_1\|_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w}$$
$$= \frac{\|(w^T\mu_0 - w^T\mu_1)^T\|_2^2}{w^T(\Sigma_0+\Sigma_1)w}$$
$$= \frac{\|(\mu_0-\mu_1)^T w\|_2^2}{w^T(\Sigma_0+\Sigma_1)w}$$
$$= \frac{[(\mu_0-\mu_1)^T w]^T(\mu_0-\mu_1)^T w}{w^T(\Sigma_0+\Sigma_1)w}$$
$$= \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T(\Sigma_0+\Sigma_1)w}$$

此时定义"类内散度矩阵"
$$S_w = \Sigma_0 + \Sigma_1$$
$$= \sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^T + \sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^T$$

以及"类间散度矩阵"
$$S_b = (\mu_0-\mu_1)(\mu_0-\mu_1)^T$$

则 $J = \frac{w^T S_b w}{w^T S_w w}$，这就是LDA欲最大化的目标，即$S_b$与$S_w$的"广义瑞利商"。

可以发现上式的分子和分母都是关于w的二次项，因此它的解与w的长度无关，只与其方向有关。不失一般性，令$w^T S_w w = 1$，则目标可改写为

$$\min_w \; -w^T S_b w \quad s.t. \; w^T S_w w = 1$$

此时，我们采用拉格朗日乘子法进行求解，则

$$S_b w = \lambda S_w w$$

其中λ为拉格朗日乘子。因为$S_b w$的方向恒为$\mu_0-\mu_1$，则令

$$S_b w = \lambda(\mu_0-\mu_1)$$

故 $$w = S_w^{-1}(\mu_0-\mu_1)$$