典型关联分析（Canonical Correlation Analysis）

JerryLead
csxulijie@gmail.com

1. 问题

在线性回归中，我们使用直线来拟合样本点，寻找 n 维特征向量 X 和输出结果（或者叫做 label）Y 之间的线性关系。其中$X \in R^n$，$Y \in R$。然而当 Y 也是多维时，或者说 Y 也有多个特征时，我们希望分析出 X 和 Y 的关系。

当然我们仍然可以使用回归的方法来分析，做法如下：

假设$X \in R^n$，$Y \in R^m$，那么可以建立等式 Y=AX 如下

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

其中$y_i = w_i^T x$，形式和线性回归一样，需要训练 m 次得到 m 个w_i。

这样做的一个缺点是，Y 中的每个特征都与 X 的所有特征关联，Y 中的特征之间没有什么联系。

我们想换一种思路来看这个问题，如果将 X 和 Y 都看成整体，考察这两个整体之间的关系。我们将整体表示成 X 和 Y 各自特征间的线性组合，也就是考察$a^T x$和$b^T y$之间的关系。

这样的应用其实很多，举个简单的例子。我们想考察一个人解题能力 X（解题速度x_1，解题正确率x_2）与他/她的阅读能力 Y（阅读速度y_1，理解程度y_2）之间的关系，那么形式化为：

$u = a_1 x_1 + a_2 x_2$ 和 $v = b_1 y_1 + b_2 y_2$

然后使用 Pearson 相关系数

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

来度量 u 和 v 的关系，我们期望寻求一组最优的解 a 和 b，使得 Corr(u, v)最大，这样得到的 a 和 b 就是使得 u 和 v 就有最大关联的权重。

到这里，基本上介绍了典型相关分析的目的。

2. CCA 表示与求解

给定两组向量x_1和x_2（替换之前的 x 为x_1，y 为x_2），x_1维度为p_1，x_2维度为p_2，默认$p_1 \le p_2$。形式化表示如下：

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad E[x] = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = Var(x) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Σ是 x 的协方差矩阵；左上角是x_1自己的协方差矩阵；右上角是$\mathrm{Cov}(x_1, x_2)$；左下角是$\mathrm{Cov}(x_2, x_1)$，也是Σ_{12}的转置；右下角是x_2的协方差矩阵。

与之前一样，我们从x_1和x_2的整体入手，定义

$$u = a^T x_1 \quad v = b^T x_2$$

我们可以算出 u 和 v 的方差和协方差：

$$\mathrm{Var}(u) = a^T \Sigma_{11} a \qquad \mathrm{Var}(v) = b^T \Sigma_{22} b \qquad \mathrm{Cov}(u, v) = a^T \Sigma_{12} b$$

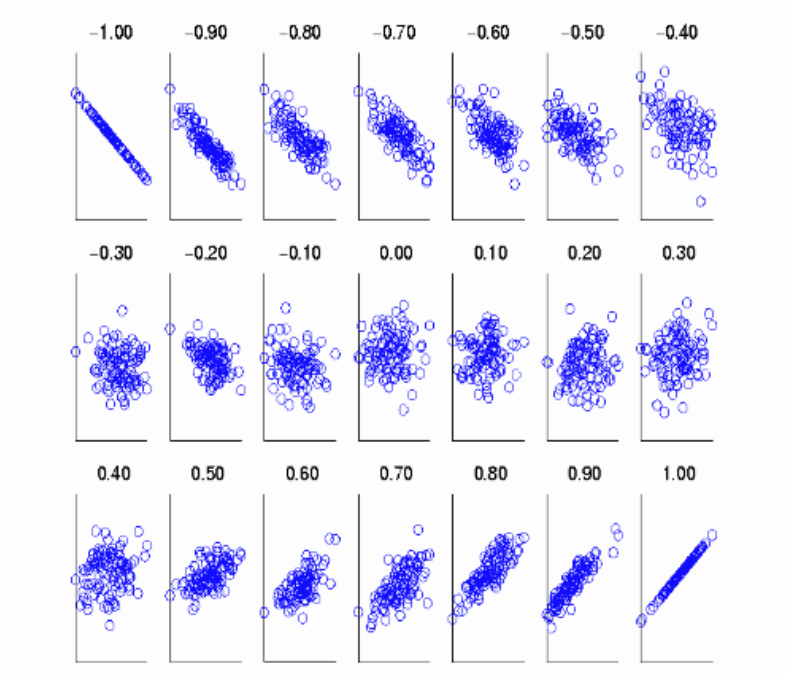
上面的结果其实很好算，推导一下第一个吧：

$$\mathrm{Var}(u) = \mathrm{Var}(a^T x_1) = \frac{1}{N}\sum_{i=1}^{N}(a^T x_1 - a^T \mu_1)^2 = a^T \frac{1}{N}\sum_{i=1}^{N}(x_1 - \mu_1)^2 \, a = a^T \Sigma_{11} a$$

最后，我们需要算 Corr(u,v)了

$$\mathrm{Corr}(u, v) = \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_{11} a}\sqrt{b^T \Sigma_{22} b}}$$

我们期望 Corr(u,v)越大越好，关于 Pearson 相关系数，《数据挖掘导论》给出了一个很好的图来说明：



Scatter plots showing the similarity from –1 to 1.

横轴是 u，纵轴是 v，这里我们期望通过调整 a 和 b 使得 u 和 v 的关系越像最后一个图越好。其实第一个图和最后一个图有联系的，我们可以调整 a 和 b 的符号，使得从第一个图变为最后一个。

接下来我们求解 a 和 b。

回想在 LDA 中，也得到了类似 Corr(u,v)的公式，我们在求解时固定了分母，来求分子（避免 a 和 b 同时扩大 n 倍仍然符号解条件的情况出现）。这里我们同样这么做。

这个优化问题的条件是：

Maximize $a^T\Sigma_{12}b$
Subject to: $a^T\Sigma_{11}a = 1, b^T\Sigma_{22}b = 1$

求解方法是构造 Lagrangian 等式，这里我简单推导如下：

$$\mathcal{L} = a^T\Sigma_{12}b - \frac{\lambda}{2}(a^T\Sigma_{11}a - 1) - \frac{\theta}{2}(b^T\Sigma_{22}b - 1)$$

求导，得

$$\frac{\partial\mathcal{L}}{\partial a} = \Sigma_{12}b - \lambda\Sigma_{11}a$$
$$\frac{\partial\mathcal{L}}{\partial b} = \Sigma_{21}a - \theta\Sigma_{22}b$$

令导数为 0 后，得到方程组：

$$\Sigma_{12}b - \lambda\Sigma_{11}a = 0$$
$$\Sigma_{21}a - \theta\Sigma_{22}b = 0$$

第一个等式左乘a^T，第二个左乘b^T，再根据$a^T\Sigma_{11}a = 1, b^T\Sigma_{22}b = 1$，得到

$$\lambda = \theta = a^T\Sigma_{12}b$$

也就是说求出的λ即是 Corr(u,v)，只需找最大λ即可。

让我们把上面的方程组进一步简化，并写成矩阵形式，得到

$$\Sigma_{11}^{-1}\Sigma_{12}b = \lambda a$$
$$\Sigma_{22}^{-1}\Sigma_{21}a = \lambda b$$

写成矩阵形式

$$\begin{bmatrix}\Sigma_{11}^{-1} & 0\\ 0 & \Sigma_{22}^{-1}\end{bmatrix}\begin{bmatrix}0 & \Sigma_{12}\\ \Sigma_{21} & 0\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix} = \lambda\begin{bmatrix}a\\ b\end{bmatrix}$$

令

$$B = \begin{bmatrix}\Sigma_{11} & 0\\ 0 & \Sigma_{22}\end{bmatrix}, \mathrm{A} = \begin{bmatrix}0 & \Sigma_{12}\\ \Sigma_{21} & 0\end{bmatrix}, \mathrm{w} = \begin{bmatrix}a\\ b\end{bmatrix}$$

那么上式可以写作：

$$B^{-1}\mathrm{A}w = \lambda w$$

显然，又回到了求特征值的老路上了，只要求得B^{-1}A的最大特征值$\lambda_{\max}$，那么 Corr(u,v) 和 a 和 b 都可以求出。

在上面的推导过程中，我们假设了Σ_{11}和Σ_{22}均可逆。一般情况下都是可逆的，只有存在特征间线性相关时会出现不可逆的情况，在本文最后会提到不可逆的处理办法。

再次审视一下，如果直接去计算B^{-1}A的特征值，复杂度有点高。我们将第二个式子代入第一个，得

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}a = \lambda^2 a$$

这样先对$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$求特征值$\lambda^2$和特征向量$a$，然后根据第二个式子求得 b。

待会举个例子说明求解过程。

假设按照上述过程，得到了λ最大时的a_1和b_1。那么a_1和b_1称为典型变量（canonical variates），λ即是 u 和 v 的相关系数。

最后，我们得到 u 和 v 的等式为：

$$\mathrm{u} = a_1^{\mathrm{T}}\mathrm{x}_1 \quad \mathrm{v} = b_1^T\mathrm{x}_2$$

我们也可以接着去寻找第二组典型变量对，其最优化条件是

Maximize $a_2^{\mathrm{T}}\Sigma_{12}b_2$
Subject to: $a_2^{\mathrm{T}}\Sigma_{11}a_2 = 1, b_2^T\Sigma_{22}b_2 = 1$
$a_2^{\mathrm{T}}\Sigma_{11}a_1 = 0, b_2^T\Sigma_{22}b_1 = 0$

其实第二组约束条件就是$\mathrm{Cov}(u_2, u_1) = 0, \mathrm{Cov}(v_2, v_1) = 0$。

计算步骤同第一组计算方法，只不过是λ取$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$的第二大特征值。
得到的a_2和b_2其实也满足

$$a_2^{\mathrm{T}}\Sigma_{12}b_1 = 0, b_2^T\Sigma_{21}a_1 = 0 \text{ 即 } \mathrm{Cov}(u_2, v_1) = 0, \mathrm{Cov}(v_2, u_1) = 0$$

总结一下，i 和 j 分别表示λ_i和λ_j得到结果

$$\mathrm{Corr}(u_i, v_i) = \lambda_i \quad \mathrm{Corr}(u_i, u_j) = 0$$

$$\mathrm{Corr}(v_i, v_j) = 0 \quad \mathrm{Corr}(u_i, v_j) = 0 (i \neq j)$$

3. CCA 计算例子

我们回到之前的评价一个人解题和其阅读能力的关系的例子。假设我们通过对样本计算协方差矩阵得到如下结果：

$$\Sigma = \begin{bmatrix} 1 & .4 & .5 & .6 \\ .4 & 1 & .3 & .4 \\ .5 & .3 & 1 & .2 \\ .6 & .4 & .2 & 1 \end{bmatrix}$$

$$\Sigma_{11} = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix} \quad \Sigma_{12} = \begin{bmatrix} .5 & .6 \\ .3 & .4 \end{bmatrix} \quad \Sigma_{21} = \begin{bmatrix} .5 & .3 \\ .6 & .4 \end{bmatrix} \quad \Sigma_{22} = \begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix}$$

然后求$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$，得

$$\mathrm{A} = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{bmatrix} .452 & .289 \\ .146 & .495 \end{bmatrix}$$

这里的 A 和前面的$B^{-1}\mathrm{A}w = \lambda w$中的 A 不是一回事（这里符号有点乱，不好意思）。
然后对 A 求特征值和特征向量，得到

$$\lambda_1^2 = .5457 \quad \lambda_2^2 = .0009 \quad \mathrm{Vec}\mathbf{A} = \begin{bmatrix} .951 & -.540 \\ .309 & .842 \end{bmatrix}$$

然后求 b，之前我们说的方法是根据$\Sigma_{22}^{-1}\Sigma_{21}a = \lambda b$ 求 b，这里，我们也可以采用类似求 a 的方法来求 b。
回想之前的等式

$$\Sigma_{11}^{-1}\Sigma_{12}b = \lambda a$$
$$\Sigma_{22}^{-1}\Sigma_{21}a = \lambda b$$

我们将上面的式子代入下面的，得

$$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}b = \lambda^2 b$$

然后直接对$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$求特征向量即可，注意$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$和$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$的特征值相同，这个可以自己证明下。

不管使用哪种方法，

$$\mathrm{B} = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \begin{bmatrix} .206 & .251 \\ .278 & .340 \end{bmatrix}$$

$$\mathrm{Vec}\mathbf{B} = \begin{bmatrix} .595 & -.774 \\ .804 & .633 \end{bmatrix}$$

这里我们得到 a 和 b 的两组向量，到这还没完，我们需要让它们满足之前的约束条件

$$a_i^{\mathrm{T}}\Sigma_{11}a_i = 1, b_i^T\Sigma_{22}b_i = 1$$

这里的a_i应该是我们之前得到的 VecA 中的列向量的 m 倍，我们只需要求得 m，然后将 VecA 中的列向量乘以 m 即可。

$$\mathrm{m}^2 a'^{\mathrm{T}}_i \mathrm{R}_{11} a'_i = 1$$

这里的a'_i是 VecA 的列向量。

$$\mathbf{A} = \mathbf{VecA}\begin{pmatrix} 1.23 & 0 \\ 0 & .636 \end{pmatrix}^{-\frac{1}{2}} \text{ and } \quad \mathbf{B} = \mathbf{VecB}\begin{pmatrix} 1.19 & 0 \\ 0 & .804 \end{pmatrix}^{-\frac{1}{2}}$$

因此最后的 a 和 b 为：

$$\mathbf{A} = \begin{bmatrix} .856 & -.677 \\ .278 & 1.055 \end{bmatrix} \text{ and } \quad \mathbf{B} = \begin{bmatrix} .545 & -.863 \\ .737 & .706 \end{bmatrix}$$

第一组典型变量为

$$\mathrm{u}_1 = .856z_1 + .278z_2 \qquad v_1 = .545z_3 + .737z_4$$

相关系数

$$\mathrm{Corr}(u_1, v_1) = \sqrt{\lambda_1^2} = \sqrt{.5457} = .74$$

第二组典型变量为

$$\mathrm{u}_2 = -.677z_1 + 1.055z_2 \qquad v_1 = -.863z_3 + .706z_4$$

相关系数

$$\mathrm{Corr}(u_2, v_2) = \sqrt{\lambda_2^2} = \sqrt{.0009} = .03$$

这里的z_1（解题速度），z_2（解题正确率），z_3（阅读速度），z_4（阅读理解程度）。他们前面的系数意思不是特征对单个 u 或 v 的贡献比重，而是从 u 和 v 整体关系看，当两者关系最密切时，特征计算时的权重。

4. Kernel Canonical Correlation Analysis（KCCA）

通常当我们发现特征的线性组合效果不够好或者两组集合关系是非线性的时候，我们会尝试核函数方法，这里我们继续介绍 Kernel CCA。

在《支持向量机-核函数》那一篇中，大致介绍了一下核函数，这里再简单提一下：

当我们对两个向量作内积的时候

$$\langle x, y\rangle = \sum x_i\, y_i$$

我们可以使用$\Phi(x)$，$\Psi(y)$来替代x和y，比如原来的x特征向量为$(x_1, x_2, x_3)^T$，那么

我们可以定义

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}.$$

如果$\Psi(y)$与$\Phi(x)$的构造一样，那么

$$\langle \Phi(x), \Phi(y)\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_ix_j)\big(y_iy_j\big) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_iy_ix_jy_j = \sum_{i=1}^{n}(x_i\, y_i)\sum_{j=1}^{n}(x_jy_j)$$

$$= (x^Ty)^2 = K\langle x, y\rangle$$

这样，仅通过计算 x 和 y 的内积的平方就可以达到在高维空间（这里为n^2）中计算$\Phi(x)$和$\Phi(y)$内积的效果。

由核函数，我们可以得到核矩阵 K，其中

$$\mathrm{K}_{i,j} = K(x^{(i)}, y^{(i)})$$

即第i行第j列的元素是第i个和第j个样例在核函数下的内积。

一个很好的核函数定义：

$$\phi : \mathbf{x} = (\mathbf{x}_1, \ldots \mathbf{x}_n) \mapsto \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x}))\ (n < N)$$

其中样例 x 有 n 个特征，经过$\Phi(x)$变换后，从 n 维特征上升到了 N 维特征，其中每一个特征是$\phi_i(x) = f(x_1, x_2, \ldots, x_n)$。

回到 CCA，我们在使用核函数之前

$$\mathrm{u} = a^Tx \qquad \mathrm{v} = b^Ty$$

这里假设 x 和 y 都是 n 维的，引入核函数后，$\phi_x(x)$和$\phi_y(y)$变为了 N 维。

使用核函数后，u 和 v 的公式为：

$$\mathrm{u} = c^T\phi_x(x) \qquad \mathrm{v} = d^T\phi_y(y)$$

这里的 c 和 d 都是 N 维向量。
现在我们有样本$\{(x_i, y_i)\}_{i=1}^{M}$，这里的x_i表示样本 x 的第 i 个样例，是 n 维向量。
根据前面说过的相关系数，构造拉格朗日公式如下：

$$\begin{aligned} L_0 \quad = \quad & \mathrm{E}[(u-\mathrm{E}[u])(v-\mathrm{E}[v])] \\ & -\frac{\lambda_1}{2}\mathrm{E}[(u-\mathrm{E}[u])^2] \\ & -\frac{\lambda_2}{2}\mathrm{E}[(v-\mathrm{E}[v])^2]. \end{aligned} \tag{7}$$

其中

$$\mathrm{E[u]} = \frac{1}{M}\sum_i c^T \phi_x(x_i)$$

$$\mathrm{E[uv]} = \frac{1}{M}\sum_i c^T \phi_x(x_i)\, d^T \phi_y(y_i)$$

然后让 L 对 a 求导，令导数等于 0，得到（这一步我没有验证，待会从宏观上解释一下）

$$c = \sum_i \alpha_i \phi_x(x_i)$$

同样对 b 求导，令导数等于 0，得到

$$d = \sum_i \beta_i \phi_y(y_i)$$

求出 c 和 d 干嘛呢？c 和 d 只是ϕ的系数而已，按照原始的 CCA 做法去做就行了呗，为了再引入α和β？

回答这个问题要从核函数的意义上来说明。核函数初衷是希望在式子中有$\phi^T(x)\phi(y)$，然后用 K 替换之，根本没有打算去计算出实际的ϕ。因此即是按照原始 CCA 的方式计算出了 c 和 d，也是没用的，因为根本有没有实际的ϕ让我们去做$c^T\phi(x)$。另一个原因是核函数比如高斯径向基核函数可以上升到无限维，N 是无穷的，因此 c 和 d 也是无穷维的，根本没办法直接计算出来。我们的思路是在原始的空间中构造出权重α和β，然后利用ϕ将α和β上升到高维，他们在高维对应的权重就是 c 和 d。

虽然α和β是在原始空间中（维度为样例个数 M），但其作用点不是在原始特征上，而是原始样例上。看上面得出的 c 和 d 的公式就知道。α通过控制每个高维样例的权重，来控制 c。

好了，接下来我们看看使用α和β后，u 和 v 的变化

$$\mathrm{u} = \langle \mathrm{c}, \phi(x) \rangle = \sum_i \alpha_i \langle \phi_x(x_i), \phi_x(x) \rangle$$

$$\mathrm{v} = \langle \mathrm{d}, \phi(y) \rangle = \sum_i \beta_i \langle \phi_y(y_i), \phi_y(y) \rangle$$

$\phi_x(x_i)$表示可以将第 i 个样例上升到的 N 维向量，$\phi_x(x)$意义可以类比原始 CCA 的 x。鉴于这样表示接下来会越来越复杂，改用矩阵形式表示。

$$\begin{bmatrix} \mathrm{c}_1 \\ \mathrm{c}_2 \\ \vdots \\ \mathrm{c}_N \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ \phi_x(x_1) & \phi_x(x_2) & \cdots & \phi_x(x_m) \\ | & | & | & | \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix}$$

简写为

$$\mathrm{c} = \mathrm{X}^T \alpha$$

其中 X（M×N）为

$$\begin{bmatrix} - & \phi_x^T(x_1) & - \\ - & \phi_x^T(x_2) & - \\ - & \vdots & - \\ - & \phi_x^T(x_m) & - \end{bmatrix}$$

我们发现

$$K_x = XX^T$$

我们可以算出 u 和 v 的方差和协方差（这里实际上事先对样本x和y做了均值归 0 处理）：

$$\mathrm{Var(u)} = c^T Var\big(\phi_x(x)\big)c = c^T X^T X c = \alpha^T XX^T XX^T \alpha = \alpha^T K_x K_x \alpha$$

$$\mathrm{Var(v)} = \beta^T \mathrm{K}_y \mathrm{K}_y \beta$$

$$\mathrm{Cov(u,v)} = c^T Cov\left(\phi_x(x), \phi_y(y)\right) d = c^T X^T Y d = \alpha^T XX^T YY^T \beta = \alpha^T K_x \mathrm{K}_y \beta$$

这里$\phi_x(x)$和$\phi_y(y)$维度可以不一样。

最后，我们得到 Corr(u,v)

$$\mathrm{Corr(u,v)} = \frac{\alpha^{\mathrm{T}} \mathrm{K}_{\mathrm{x}} K_y \beta}{\sqrt{\alpha^{\mathrm{T}} \mathrm{K}_{\mathrm{x}} \mathrm{K}_{\mathrm{x}} \alpha} \sqrt{\beta^{\mathrm{T}} K_y K_y \beta}}$$

可以看到，在将x_1和x_2处理成$\mathrm{E}[x_1] = 0$，$\mathrm{E}[x_2] = 0$后，得到的结果和之前形式基本一样，只是将Σ替换成了两个 K 乘积。

因此，得到的结果也是一样的，之前是

$$B^{-1} \mathrm{A} w = \lambda w$$

其中

$$B = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}, \mathrm{A} = \begin{bmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{bmatrix}, \mathrm{w} = \begin{bmatrix} a \\ b \end{bmatrix}$$

引入核函数后，得到

$$B^{-1} \mathrm{A} w = \lambda w$$

其中

$$B = \begin{bmatrix} K_x K_x & 0 \\ 0 & K_y K_y \end{bmatrix}, \mathrm{A} = \begin{bmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{bmatrix}, \mathrm{w} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

注意这里的两个 w 有点区别，前面的a维度和 x 的特征数相同，b维度和 y 的特征数相同。后面的α维度和 x 的样例数相同，β维度和 y 的样例数相同，严格来说“α维度=β维度”。

5. 其他话题

1、当协方差矩阵不可逆时，怎么办？

要进行 regularization。

一种方法是将前面的 KCCA 中的拉格朗日等式加上二次正则化项，即：

$$L = L_0 + \frac{\eta}{2}(\|c\|^2 + \|d\|^2)$$

这样求导后得到的等式中，等式右边的矩阵一定是正定矩阵。

第二种方法是在 Pearson 系数的分母上加入正则化项，同样结果也一定可逆。

$$\begin{aligned}\rho &= \max_{\alpha,\beta}\frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa\|\mathbf{w}_x\|^2)\cdot(\beta' K_y^2 \beta + \kappa\|\mathbf{w}_y\|^2))}} \\ &= \max_{\alpha,\beta}\frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa\alpha' K_x \alpha)\cdot(\beta' K_y^2 \beta + \kappa\beta' K_y \beta)}}.\end{aligned}$$

2、求 Kernel 矩阵效率不高怎么办？

使用 Cholesky decomposition 压缩法或者部分 Gram-Schmidt 正交化法，。

3、怎么使用 CCA 用来做预测？

先找出 X 和 Y 的典型相关系数，新来一个样例 Xnew，在 X 中使用 KNN，然后找到在 Y 中对应的 N 个样例，求均值或者带权重均值等预测 Ynew。

4、如果有多个集合怎么办？X、Y、Z…？怎么衡量多个样本集的关系？

这个称为 Generalization of the Canonical Correlation。方法是使得两两集合的距离差之和最小。可以参考文献 2。

6. 参考文献

1、 http://www.stat.tamu.edu/~rrhocking/stat636/LEC-9.636.pdf

2、 **Canonical correlation analysis: An overview with application to learning methods**. David R. Hardoon , Sandor Szedmak and John Shawe-Taylor

3、 **A kernel method for canonical correlation analysis**. Shotaro Akaho

4、 **Canonical Correlation a Tutorial.** Magnus Borga

5、 **Kernel Canonical Correlation Analysis**. Max Welling