模式识别与机器学习

Pattern Recognition & Machine Learning

第六讲 主成分分析与相关的谱方法

- 本节学习目标

  - ✓能够熟练运用主成分分析
  - ✓理解概率主成分分析的原理
  - ✓理解核主成分分析的原理
  - ✓能够熟练运用线性判别分析和典型相关分析

目录

- 最大化方差

假定观测数据$\{\mathbf{x}_i\}$是在欧式空间上的D维数据。

“信号通常具有较大的方差”

PCA的目标是将数据投影到一个维度为$M(M \le D)$的子空间，使得投影后的数据在各个维度上的方差的和最大。

首先考虑投影到一维空间$M = 1$，投影变换$\mathbf{u}_1$是一个D维变量。

PCA只关注投影的方向，假设投影向量满足 $\mathbf{u}_1^\top \mathbf{u}_1 = 1$.

每个数据点$\mathbf{x}_i$经过投影得到 $z_i = \mathbf{u}_1^\top \mathbf{x}_i$，投影后数据的方差为

$$Var(z) = \frac{1}{N}\sum_{i=1}^{N}\{z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\}^2 = \mathbf{u}_1^\top S \mathbf{u}_1,$$

其中S表示原始空间数据的协方差矩阵

$$S = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m}(\mathbf{x}))(\mathbf{x}_i - \mathbf{m}(\mathbf{x}))^\top,$$

$$\mathbf{m}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i.$$

带约束的优化问题可以利用拉格朗日乘子法得到非约束优化问题。引入拉格朗日乘子λ_1之后，可得等价的优化目标为

$$\arg\max_{\mathbf{u}_1} \mathbf{u}_1^{\top} S\mathbf{u}_1 + \lambda_1(1-\mathbf{u}_1^{\top}\mathbf{u}_1).$$

对上式关于$\mathbf{u}_1$求导，并使导数为0，可得

$$S\mathbf{u}_1 = \lambda_1\mathbf{u}_1, \qquad Var(z) = \mathbf{u}_1^{\top} S\mathbf{u}_1 = \lambda_1.$$

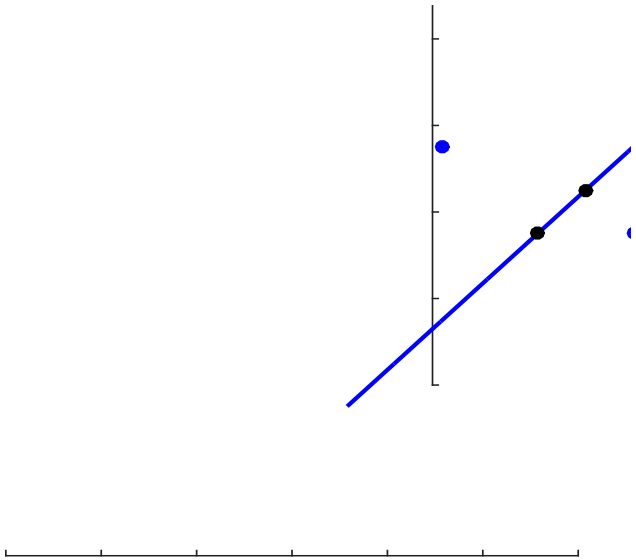

图11-1 使用PCA将数据投影到一维空间的示例

考虑更一般化的情况，当$M > 1$时，投影变换是一个$D \times M$的矩阵U，且满足各个投影方向相互正交 $U^\top U = \mathbf{I}$.
优化目标定义为投影后所得的子空间中数据在每个维度上的方差总和最大

$$\arg\max_U \mathrm{Tr}(U^\top S U),$$

$$s.t.\ U^\top U = \mathbf{I}.$$

使用拉格朗日乘子法，引入M个拉格朗日乘子$\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \cdots, \lambda_M\}$

$$\arg\max_U \mathrm{Tr}\left[U^\top S U + \mathrm{diag}(\boldsymbol{\lambda})(\mathbf{I} - U^\top S U)\right],$$

对上式序投影矩阵求导，设置导数为零

$$SU = \mathrm{diag}(\boldsymbol{\lambda})U.$$

$$\mathrm{Tr}(U^\top S U) = \sum_{m=1}^{M} \lambda_m.$$

最优的M维子空间投影应该是原始数据的协方差矩阵$\mathbf{S}$的M个最大特征值$\lambda_1, \lambda_2, \cdots, \lambda_M$对应的特征向量$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_M$构成的矩阵$U$.

- 最小化误差

以最小化原始数据与投影后数据的平方和误差为优化目标。

首先，考虑数据$\mathbf{x}_n$由原始D维空间$\mathcal{A}$转换到新的D维空间$\mathcal{B}$的变换关系，其中新的D维空间$\mathcal{B}$由标准正交基$U=[\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_D]$构成，且$U^{\top}U=\mathbf{I}$.
数据$\mathbf{x}_n$可以表示为标准正交基的线性组合，且组合系数是数据点在新空间中的坐标$\mathbf{x}'_n$

$$\mathbf{x}_n=\sum_{d=1}^{D}x'_{nd}\mathbf{u}_d.$$

$$x'_{nd}=\mathbf{x}_n^{\top}\mathbf{u}_d,\ d=1,2,...,D.$$

$$\mathbf{x}_n=\sum_{d=1}^{D}(\mathbf{x}_n^{\top}\mathbf{u}_d)\mathbf{u}_d.$$

其次，考虑数据$\mathbf{x}_n$投影到最优$M(M \leq D)$维子空间的变换关系。
假设最优子空间使用U中$M(M \leq D)$个标准正交基向量$\{\mathbf{u}_m\}_{m=1}^{M}$表示，数据点$\mathbf{x}_n$投影到子空间之后的数据表示为 $\mathbf{z}_n = [z_{n1}, z_{n2}, \ldots, z_{nM}]^{\top}$.
为了得到投影后的数据在原始空间的重构表示，需要使用D组正交基构建包含M维子空间的D维空间，这D组正交基包括子空间的标准正交基向量$\{\mathbf{u}_m\}_{m=1}^{M}$和其正交补$\{\mathbf{u}_m\}_{m=M+1}^{D}$。投影后的数据$\mathbf{z}_n$在原始空间中重构后的表示为

$$\tilde{\mathbf{x}}_n = \sum_{m=1}^{M} z_{nm}\mathbf{u}_m + \sum_{m=M+1}^{D} b_m \mathbf{u}_m,$$

其中，$\{b_m\}$是一组不依赖单个样本的变量，即为了误差最小允许增加来自正交补空间的一个常值向量。

最后，可得最小化投影损失的优化目标为

$$J=\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n-\tilde{\mathbf{x}}_n\|_2^2=\frac{1}{N}\sum_{n=1}^{N}\|\sum_{d=1}^{D}(\mathbf{x}_n^\top\mathbf{u}_d)\mathbf{u}_d-\sum_{m=1}^{M}z_{nm}\mathbf{u}_m-\sum_{m=M+1}^{D}b_m\mathbf{u}_m\|_2^2,$$

$$s.t.\ U^\top U=\mathbf{I},$$

其中变量包括$\{\mathbf{u}_i\},\{z_{nm}\},\{b_m\}$. 分别关于$\{z_{nm}\}$和$\{b_m\}$求导可得

$$z_{nm}=\mathbf{x}_n^\top\mathbf{u}_m,\ n=1,2,...,N,\ m=1,2,...,M,$$

$$b_m=\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^\top\right)\mathbf{u}_m,\ m=M+1,M+2,...,D.$$

将上两式代入优化目标中可得

$$J=\frac{1}{N}\sum_{n=1}^{N}\sum_{m=M+1}^{D}\left(\mathbf{x}_n^\top\mathbf{u}_m-\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^\top\right)\mathbf{u}_m\right)^2=\sum_{m=M+1}^{D}\mathbf{u}_m^\top S\mathbf{u}_m$$

$$s.t.\ U^\top U=\mathbf{I}.$$

通过引入拉格朗日乘子 $\lambda_1, \lambda_2, \cdots, \lambda_D$ ，可以得到对应的无约束优化目标，并且优化目标的解$\{\mathbf{u}_d\}$满足如下表示：

$$S\mathbf{u}_d = \lambda_d \mathbf{u}_d,\ d = 1, 2, ..., D,$$

其中$\{\lambda_d\}$是协方差矩阵S的特征值，$\{\mathbf{u}_d\}$是对应的特征向量。目标损失可以进一步化简为

$$J = \sum_{m=M+1}^{D} \lambda_m.$$

因此，当$\{\lambda_{m+1}, \lambda_{m+2}, \cdots, \lambda_D\}$是$D - M$个最小特征值时，投影损失最小，此时最优子空间的基向量（即投影向量）是协方差矩阵S的M个最大特征值$\{\lambda_1, \lambda_2, \cdots, \lambda_M\}$对应的特征向量。

- 主成分分析与K-L变换
  - PCA是一种特殊的K-L变换，PCA的正交投影矩阵通过对协方差矩阵进行特征值分解获得。
  - 当K-L变换使用协方差矩阵时，K-L变换等同于PCA。

自相关矩阵与数据的类别无关，假设所有数据统一表示为$X=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N]^\top$，那么数据的自相关矩阵为 $R_x=\mathbb{E}[XX^\top]=\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^\top.$

总类内散度矩阵与数据的类别相关，假设数据具有C个类别，且类别中的数据表示为c，总类内散度矩阵为$X^c=[\mathbf{x}_1^c,\mathbf{x}_2^c,...,\mathbf{x}_{N_c}^c]^\top$

$$S_w=\frac{1}{C}\sum_{k=1}^{C}S_w^c,$$

其中S_w^c表示每个类别内的散度矩阵

$$S_w^c=\frac{1}{N_c}\sum_{n=1}^{N_c}(\mathbf{x}_n^c-\mathbf{m}^c)(\mathbf{x}_n^c-\mathbf{m}^c)^\top,$$

$$\mathbf{m}^c=\frac{1}{N_c}\sum_{n=1}^{N_c}\mathbf{x}_n^c.$$

目录

概率PCA是一种线性高斯模型，它的边缘概率分布和条件概率分布均假设为高斯分布。为了表示主成分子空间，模型引入潜变量$\mathbf{z}$，并且定义潜变量具有标准高斯先验分布

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}\,|\,\mathbf{0},\mathbf{I}).$$

此外，假设在给定潜变量的条件下观测变量$\mathbf{x}$的条件分布$p(\mathbf{x}|\mathbf{z})$也是高斯分布

$$p(\mathbf{x}\,|\,\mathbf{z}) = \mathcal{N}(\mathbf{x}\,|\,W\mathbf{z}+\boldsymbol{\mu},\sigma^{2}\mathbf{I}),$$

概率PCA的生成过程如下：先从先验分布中生成一个潜变量$\mathbf{z}$，然后通过线性变换得到包含噪声$\boldsymbol{\epsilon}$的观测$\mathbf{x}$：

$$\mathbf{x} = W\mathbf{z}+\boldsymbol{\mu}+\boldsymbol{\epsilon}.$$

概率PCA的模型参数可通过最大似然估计来进行求解。
已知模型的先验和似然，可得观测数据的边缘概率分布$p(\mathbf{x})$为

$$p(\mathbf{x}) = \int p(\mathbf{x}\,|\,\mathbf{z})\,p(\mathbf{z})d\mathbf{z}.$$

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}\,|\,\boldsymbol{\mu}, WW^{\top} + \sigma^{2}\mathbf{I}\right).$$

对边缘概率似然分布关于模型参数求导

$$W_{\mathrm{m}\ell} = U_{M}\left(\mathrm{diag}(\boldsymbol{\lambda}) - \sigma^{2}_{\mathrm{m}\ell}\mathbf{I}\right)^{1/2} R,$$

$$\sigma^{2}_{\mathrm{m}\ell} = \frac{1}{D-M}\sum_{m=M+1}^{D}\lambda_{m},$$

$$\boldsymbol{\mu}_{\mathrm{m}\ell} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_{n},$$

通过贝叶斯公式，可得关于潜变量$\mathbf{z}$的后验概率分布为

$$p(\mathbf{z}\,|\,\mathbf{x}) = \mathcal{N}(\mathbf{z}\,|\,M^{-1}W^{\top}(\mathbf{x}-\boldsymbol{\mu}), \sigma^{2}M^{-1}),$$

其中$M = WW^{\top} + \sigma^{2}\mathbf{I}$.

目录

核PCA：对数据进行非线性投影并且使用核技术的PCA方法。

假设训练数据$\{\mathbf{x}_n\}$是在欧式空间上的D维数据。
首先对数据进行中心化处理 $\sum_{n=1}^{N}\mathbf{x}_n = 0$
主成分子空间的基向量满足

$$S\mathbf{u}_n = \lambda_n \mathbf{u}_n,$$

$$S = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\top},$$

并且特征向量构成的矩阵满足正交约束 $\mathbf{u}_i^{\top}\mathbf{u}_j = 0,\ i \neq j.$

$\phi(\mathbf{x})$表示数据从原始空间到高维空间的非线性映射，如果投影到M维子空间，核PCA寻找投影向量$\{\mathbf{v}_m\}_{m=1}^{M}$，使得映射后的数据$\phi(X)=[\phi(\mathbf{x}_1),\phi(\mathbf{x}_2),\cdots,\phi(\mathbf{x}_N)]$经投影后在子空间各个维度的表示$\{\mathbf{v}_m^{\top}\phi(X)\}_{m=1}^{M}$的方差总和最大。假设变换后的数据均值为零，即 $\sum_{n=1}^{N}\phi(\mathbf{x}_n)=0.$
那么在高维空间上的协方差矩阵为

$$S'=\frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^{\top},$$

与PCA类似，可得其主成分子空间的基向量$\mathbf{v}_m$满足 $S'\mathbf{v}_m=\lambda_m\mathbf{v}_m.$
根据上述分析可得

$$\frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^{\top}\mathbf{v}_m=\lambda_m\mathbf{v}_m$$

由于 $\phi(\mathbf{x}_n)^{\top}\mathbf{v}_m$ 是一个标量，$\mathbf{v}_m$可以表示为$\{\phi(\mathbf{x}_N)\}$的线性组合$\mathbf{v}_m=\sum_{n=1}^{N}\alpha_{mn}\phi(\mathbf{x}_n)$
将协方差矩阵S'和$\mathbf{v}_m$代入 $S'\mathbf{v}_m=\lambda_m\mathbf{v}_m$，可得

$$\frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^{\top}\sum_{n'=1}^{N}\alpha_{mn'}\phi(\mathbf{x}_{n'})=\lambda_m\sum_{n=1}^{N}\alpha_{mn}\phi(\mathbf{x}_n).$$

$$\frac{1}{N}\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^{\top}\sum_{n'=1}^{N}\alpha_{mn'}\phi(\mathbf{x}_{n'})=\lambda_m\sum_{n=1}^{N}\alpha_{mn}\phi(\mathbf{x}_n).$$

使用核技巧，对上式两边左乘 $\phi(\mathbf{x}_\ell)^{\top}$，$\ell=1,2,...,N$
并且引入核函数 $k(\mathbf{x},\mathbf{x}')=\phi(\mathbf{x})^{\top}\phi(\mathbf{x}')$

$$\frac{1}{N}\sum_{n=1}^{N}k(\mathbf{x}_\ell,\mathbf{x}_n)\sum_{n'=1}^{N}\alpha_{mn'}k(\mathbf{x}_n,\mathbf{x}_{n'})=\lambda_m\sum_{n=1}^{N}\alpha_{mn}k(\mathbf{x}_\ell,\mathbf{x}_n),\ \ell=1,2,...,N.$$

矩阵形式

$$KK\boldsymbol{\alpha}_m=\lambda_m NK\boldsymbol{\alpha}_m,$$

非零特征值机器对应的特征向量满足 $K\boldsymbol{\alpha}_m=\lambda_m N\boldsymbol{\alpha}_m$，
因此核PCA等价于求解 $K\boldsymbol{\alpha}_m=\lambda_m N\boldsymbol{\alpha}_m$，
其中K是核函数在所有训练数据上构建的核矩阵，$\boldsymbol{\alpha}_m=[\alpha_{m1},\alpha_{m2},...,\alpha_{mN}]^{\top}$
是需要求解的特征向量
在得到所需的特征向量之后，投影后的主成分表示为

$$z_m=\mathbf{v}_m^{\top}\phi(\mathbf{x})=\sum_{n=1}^{N}\alpha_{mn}k(\mathbf{x},\mathbf{x}_n),\ m=1,2,...,M.$$

目录

- 二类数据的线性判别分析

假设$\mathbf{x}$表示原始D维二类数据，通过线性投影表示变量y

$$y = \mathbf{w}^{\top}\mathbf{x}.$$

假设类别一的均值为$\mathbf{m}_1$，类别二的均值是$\mathbf{m}_2$，投影后两类均值的距离

$$|\mathbf{w}^{\top}(\mathbf{m}_1 - \mathbf{m}_2)|.$$

假设类别c中的数据表示为 $X^c = [\mathbf{x}_1^c, \mathbf{x}_2^c, ..., \mathbf{x}_{N_c}^c],\ c = 1, 2$
原始数据的类内协方差表示为

$$\Sigma_c = \sum_{n=1}^{N_c}(\mathbf{x}_n^c - \mathbf{m}_c)(\mathbf{x}_n^c - \mathbf{m}_c)^{\top},\ c = 1, 2.$$

投影后数据的类内方差表示为

$$S_c = \sum_{n=1}^{N_c}(\mathbf{w}^{\top}\mathbf{x}_n^c - \mathbf{w}^{\top}\mathbf{m}_c)^2,\ c = 1, 2.$$

结合类间距离最大与类内方差最小，LDA求解的优化问题为

$$\arg\max_{\mathbf{w}} J(\mathbf{w}) = \frac{\left[\mathbf{w}^{\top}(\mathbf{m}_1 - \mathbf{m}_2)\right]^2}{S_1 + S_2}$$

$$\arg\max_{\mathbf{w}} J(\mathbf{w}) = \frac{\mathbf{w}^{\top}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\top}\mathbf{S}_W\mathbf{w}}$$

其中 $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\top}$，$\mathbf{S}_W = \Sigma_1 + \Sigma_2$.

由于LDA只关注最终的投影方向，可约束 $\mathbf{w}^{\top}\mathbf{S}_W\mathbf{w} = 1$

引入拉格朗日乘子λ，可得等价的优化问题为

$$\arg\max_{\mathbf{w}} J(\mathbf{w}) = \mathbf{w}^{\top}\mathbf{S}_B\mathbf{w} + \lambda(1 - \mathbf{w}^{\top}\mathbf{S}_W\mathbf{w}).$$

关于w求导可得

$$S_B\mathbf{w} = \lambda S_W\mathbf{w},$$

$$S_W^{-1}S_B\mathbf{w} = \lambda\mathbf{w}.$$

根据运算可得 $S_B\mathbf{w}$ 与 $(\mathbf{m}_1 - \mathbf{m}_2)$ 方向相同，即

$$S_B\mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\top}\mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)\alpha.$$

$$S_W^{-1}S_B\mathbf{w} = S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\alpha.$$

综上可得 $\mathbf{w} \propto S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$.

- 多类数据的线性判别分析

当数据是多类时，数据可以被投影到一个M维空间，此时投影变换不再是向量，是由一组基向量构成的矩阵W。

当使用投影矩阵时，$W^{\top}\mathbf{S}_B W$ 和 $W^{\top}\mathbf{S}_W W$ 是矩阵，无法直接相除作为优化目标，通常使用这些矩阵的特征值的和，即矩阵的迹替换

$$\arg\max_W J(W) = \frac{\mathrm{Tr}(W^{\top}\mathbf{S}_B W)}{\mathrm{Tr}(W^{\top}\mathbf{S}_W W)},$$

其中，$S_B = \sum_{c=1}^{C} N_c(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^{\top}$，$S_W = \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{x}_n^c - \mathbf{m}_c)(\mathbf{x}_n^c - \mathbf{m}_c)^{\top}$.

m表示所有数据的均值

$$\arg\max_{W} J(W) = \frac{\mathrm{Tr}(W^{\top}\mathbf{S}_{B}W)}{\mathrm{Tr}(W^{\top}\mathbf{S}_{W}W)}$$

对上式W同时缩放不影响最终求解，增加约束 $\mathrm{Tr}(W^{\top}\mathbf{S}_{W}W)=1$
引入拉格朗日乘子α，可得等价的无约束优化目标为

$$\arg\max_{W} J(W) = \mathrm{Tr}(W^{\top}S_{B}W) + \alpha\left[1-\mathrm{Tr}(W^{\top}S_{W}W)\right],$$

关于W求导可得

$$S_{B}W = S_{W}W\mathrm{diag}(\boldsymbol{\lambda}),$$

其中，$\mathrm{diag}(\boldsymbol{\lambda})$表示对角线由$\boldsymbol{\lambda}=\{\lambda_1,\lambda_2,\cdots,\lambda_M\}$构成的对角阵，且 $\sum_{m=1}^{M}\lambda_m=\alpha$. 投影矩阵$W$中的列向量可以通过求解特征值问题 $S_{W}^{-1}S_{B}\mathbf{w}=\lambda\mathbf{w}$ 获得，此时优化目标的值为

$$\mathrm{Tr}(W^{\top}S_{B}W) = \mathrm{Tr}[\mathrm{diag}(\boldsymbol{\lambda})] = \sum_{m=1}^{M}\lambda_m$$

因此最优解是前M个最大特征值对应的特征向量构成的矩阵。这里需要注意S_B的秩最大为$C-1$，所以 $S_{W}^{-1}S_{B}$ 的非零特征值对应的特征向量数目不会超过$C-1$， 所以投影后的空间维度满足约束$M\leq C-1$.

- 典型相关分析（canonical correlation analysis，CCA）

寻找两个投影矩阵，分别将原始数据的两组表示投影到一个公共空间中，并且最大化投影后的两组数据的相关性。

给定两组数据 $X=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]^\top$ 和 $Y=[\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N]^\top$，如果投影到一维空间，CCA寻找一堆线性投影向量$\mathbf{w}_x$和$\mathbf{w}_y$，使得两组数据在投影后的相关性最大，投影后两组数据之间的相关性为

$$\mathrm{corr}(\mathbf{w}_x^\top\mathbf{x},\mathbf{w}_y^\top\mathbf{y})=\frac{\mathrm{cov}(\mathbf{w}_x^\top\mathbf{x},\mathbf{w}_y^\top\mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{w}_x^\top\mathbf{x})\mathrm{var}(\mathbf{w}_y^\top\mathbf{y})}}=\frac{\mathbf{w}_x^\top C_{xy}\mathbf{w}_y}{\sqrt{(\mathbf{w}_x^\top C_{xx}\mathbf{w}_x)(\mathbf{w}_y^\top C_{yy}\mathbf{w}_y)}},$$

$$C_{xy}=\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{m}_x)(\mathbf{y}_i-\mathbf{m}_y)^\top,$$

$$\mathbf{m}_x=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i,\ \mathbf{m}_y=\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}_i.$$

$$\mathrm{corr}(\mathbf{w}_x^\top \mathbf{x}, \mathbf{w}_y^\top \mathbf{y}) = \frac{\mathrm{cov}(\mathbf{w}_x^\top \mathbf{x}, \mathbf{w}_y^\top \mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{w}_x^\top \mathbf{x})\mathrm{var}(\mathbf{w}_y^\top \mathbf{y})}} = \frac{\mathbf{w}_x^\top C_{xy} \mathbf{w}_y}{\sqrt{(\mathbf{w}_x^\top C_{xx} \mathbf{w}_x)(\mathbf{w}_y^\top C_{yy} \mathbf{w}_y)}},$$

由于$\mathbf{w}_x$和$\mathbf{w}_y$的尺度对目标式的值没有影响，约束分母中两个因子为1

$$(\mathbf{w}_x^\star, \mathbf{w}_y^\star) = \underset{\mathbf{w}_x, \mathbf{w}_y}{\mathrm{argmax}}\, \mathbf{w}_x^\top C_{xy} \mathbf{w}_y$$

$$\text{s.t. } \mathbf{w}_x^\top C_{xx} \mathbf{w}_x = 1,$$

$$\mathbf{w}_y^\top C_{yy} \mathbf{w}_y = 1.$$

引入拉格朗日乘子λ_1, λ_2，转化成无约束的优化问题

$$(\mathbf{w}_x^\star, \mathbf{w}_y^\star) = \underset{\mathbf{w}_x, \mathbf{w}_y}{\mathrm{argmax}}\, \mathbf{w}_x^\top C_{xy} \mathbf{w}_y + \lambda_1 (1 - \mathbf{w}_x^\top C_{xx} \mathbf{w}_x) + \lambda_2 (1 - \mathbf{w}_y^\top C_{yy} \mathbf{w}_y).$$

关于$\mathbf{w}_x$和$\mathbf{w}_y$求导

$$C_{xy} \mathbf{w}_y - 2\lambda_1 C_{xx} \mathbf{w}_x = \mathbf{0},$$

$$C_{xy}^\top \mathbf{w}_x - 2\lambda_2 C_{yy} \mathbf{w}_y = \mathbf{0}.$$

$$\begin{aligned} \mathbf{0} &= \mathbf{w}_x^\top C_{xy} \mathbf{w}_y - 2\mathbf{w}_x^\top \lambda_1 C_{xx} \mathbf{w}_x - \mathbf{w}_y^\top C_{xy}^\top \mathbf{w}_x + 2\mathbf{w}_y^\top \lambda_2 C_{yy} \mathbf{w}_y \\ &= 2\lambda_2 \mathbf{w}_y^\top C_{yy} \mathbf{w}_y - 2\lambda_1 \mathbf{w}_x^\top C_{xx} \mathbf{w}_x. \end{aligned}$$

$$\mathbf{0} = \mathbf{w}_x^{\top} C_{xy} \mathbf{w}_y - 2\mathbf{w}_x^{\top} \lambda_1 C_{xx} \mathbf{w}_x - \mathbf{w}_y^{\top} C_{xy}^{\top} \mathbf{w}_x + 2\mathbf{w}_y^{\top} \lambda_2 C_{yy} \mathbf{w}_y$$

$$= 2\lambda_2 \mathbf{w}_y^{\top} C_{yy} \mathbf{w}_y - 2\lambda_1 \mathbf{w}_x^{\top} C_{xx} \mathbf{w}_x .$$

结合上式以及约束可得$\lambda_1 = \lambda_2$，令$\lambda = \lambda_1 = \lambda_2$

根据 $C_{xy}^{\top} \mathbf{w}_x - 2\lambda_2 C_{yy} \mathbf{w}_y = \mathbf{0}$ 可得$\mathbf{w}_x$和$\mathbf{w}_y$的关系为

$$\mathbf{w}_y = \frac{C_{yy}^{-1} C_{xy}^{\top} \mathbf{w}_x}{2\lambda}.$$

代入 $C_{xy} \mathbf{w}_y - 2\lambda_1 C_{xx} \mathbf{w}_x = \mathbf{0}$ 可得$\mathbf{w}_x$需满足

$$C_{xy} C_{yy}^{-1} C_{xy}^{\top} \mathbf{w}_x = 4\lambda^2 C_{xx} \mathbf{w}_x .$$

至此，求解优化问题等价于求解上述公式所示的广义特征值问题，即求解形如 $A\hat{\mathbf{x}} = \hat{\lambda} B\hat{\mathbf{x}}$ 的问题。

- 对标准CCA使用核方法进行非线性扩展可得核CCA。
- 核方法的关键思想：将数据映射到更高维度的特征空间，且高维空间中向量的内积可以通过核函数计算。

假设映射函数为 $\phi_x, k_x(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi_x(\mathbf{x}_i), \phi_x(\mathbf{x}_j) \rangle,\ i, j = 1, 2, ..., N$
核CCA对两组数据分别引入两个映射函数，$\phi_x(\mathbf{x})$和$\phi_y(\mathbf{y})$
首先对原始数据进行中心化处理 $X = X - m(X),\ Y = Y - m(Y)$

$$C_{xy} = \frac{1}{N} X^\top Y, C_{xx} = \frac{1}{N} X^\top X, C_{yy} = \frac{1}{N} Y^\top Y$$

由于$\mathbf{w}_x$和$\mathbf{w}_y$可分别表示为数据$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$和$\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N$的线性组合

$$\mathbf{w}_x = X^\top \boldsymbol{\alpha},\ \boldsymbol{\alpha} \in R^N$$

$$\mathbf{w}_y = Y^\top \boldsymbol{\beta},\ \boldsymbol{\beta} \in R^N$$

CCA的优化问题表示为

$$\arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\mathbf{w}_x^\top C_{xy} \mathbf{w}_y}{\sqrt{(\mathbf{w}_x^\top C_{xx} \mathbf{w}_x)(\mathbf{w}_y^\top C_{yy} \mathbf{w}_y)}} = \frac{\boldsymbol{\alpha}^\top X X^\top Y Y^\top \boldsymbol{\beta}}{\sqrt{(\boldsymbol{\alpha}^\top X X^\top X X^\top \boldsymbol{\alpha})(\boldsymbol{\beta}^\top Y Y^\top Y Y^\top \boldsymbol{\beta})}}.$$

引入两个映射函数$\phi_x(\mathbf{x})$和$\phi_y(\mathbf{y})$后，令K_{XX}表示为数据X的核矩阵，K_{YY}表示为数据Y的核矩阵，根据$K_{XX}=\phi_x(X)\phi_x(X)^\top, K_{YY}=\phi_y(Y)\phi_y(Y)^\top$可得核CCA的优化问题为

$$\arg\max_{\alpha,\beta}\frac{\alpha^\top K_{XX}K_{YY}\beta}{\sqrt{(\alpha^\top K_{XX}K_{XX}\alpha)(\beta^\top K_{YY}K_{YY}\beta)}}.$$

由于α和β的尺度对优化目标没有影响，因此可得等价的优化问题为

$$\arg\max_{\alpha,\beta}\alpha^\top K_{XX}K_{YY}\beta,$$

$$\text{s.t. } \alpha^\top K_{XX}K_{XX}\alpha=1,$$

$$\beta^\top K_{YY}K_{YY}\beta=1.$$

引入拉格朗日乘子，可以得到其等价的无约束优化问题为

$$(\alpha^*,\beta^*)=\arg\max_{\alpha,\beta}\alpha^\top K_{XX}K_{YY}\beta+\lambda_1(1-\alpha^\top K_{XX}K_{XX}\alpha)+\lambda_2(1-\beta^\top K_{YY}K_{YY}\beta).$$

关于α和β求导

$$K_{XX}K_{YY}\beta-2\lambda_1K_{XX}K_{XX}\alpha=\mathbf{0},$$

$$K_{YY}K_{XX}\alpha-2\lambda_2K_{YY}K_{YY}\beta=\mathbf{0}.$$

$$K_{XX}K_{YY}\boldsymbol{\beta}-2\lambda_1 K_{XX}K_{XX}\boldsymbol{\alpha}=\mathbf{0},$$
$$K_{YY}K_{XX}\boldsymbol{\alpha}-2\lambda_2 K_{YY}K_{YY}\boldsymbol{\beta}=\mathbf{0}.$$
$$\begin{aligned}\mathbf{0}&=\boldsymbol{\alpha}^{\top}K_{XX}K_{YY}\boldsymbol{\beta}-2\lambda_1\boldsymbol{\alpha}^{\top}K_{XX}K_{XX}\boldsymbol{\alpha}-\boldsymbol{\beta}^{\top}K_{YY}K_{XX}\boldsymbol{\alpha}+2\lambda_2\boldsymbol{\beta}^{\top}K_{YY}K_{YY}\boldsymbol{\beta}\\&=2\lambda_2\boldsymbol{\beta}^{\top}K_{YY}K_{YY}\boldsymbol{\beta}-2\lambda_1\boldsymbol{\alpha}^{\top}K_{XX}K_{XX}\boldsymbol{\alpha}.\end{aligned}$$

根据上式和约束条件可得$\lambda_1=\lambda_2$. 令$\lambda_1=\lambda_2=\lambda$，根据上式可得$\boldsymbol{\beta}$满足

$$\boldsymbol{\beta}=\frac{K_{YY}^{-1}K_{XX}\boldsymbol{\alpha}}{2\lambda}$$

代入 $K_{XX}K_{YY}\boldsymbol{\beta}-2\lambda_1 K_{XX}K_{XX}\boldsymbol{\alpha}=\mathbf{0}$，可得$\boldsymbol{\alpha}$满足

$$K_{XX}K_{XX}\boldsymbol{\alpha}=4\lambda^2 K_{XX}K_{XX}\boldsymbol{\alpha}$$
$$\mathbf{I}\boldsymbol{\alpha}=4\lambda^2\boldsymbol{\alpha}.$$

在得到矩阵A和B之后，映射后的数据$\phi_x(X)$和$\phi_y(Y)$在新的空间表示为

$$\phi_x(X)\mathbf{w}_x=\phi_x(X)\phi_x(X)^{\top}A=K_{XX}A,$$
$$\phi_y(Y)\mathbf{w}_y=\phi_y(Y)\phi_y(Y)^{\top}B=K_{YY}B.$$

核CCA带有正则化约束的优化问题表示为

$$\arg\max_{\alpha,\beta} \frac{\alpha^\top K_{XX} K_{YY} \beta}{\sqrt{(\alpha^\top K_{XX} K_{XX} \alpha + \kappa \| \mathbf{w}_x \|^2)(\beta^\top K_{YY} K_{YY} \beta + \kappa \| \mathbf{w}_y \|^2)}}$$

$$= \arg\max_{\alpha,\beta} \frac{\alpha^\top K_{XX} K_{YY} \beta}{\sqrt{(\alpha^\top K_{XX} K_{XX} \alpha + \kappa \alpha^\top K_{XX} \alpha)(\beta^\top K_{YY} K_{YY} \beta + \kappa \beta^\top K_{YY} \beta)}},$$

由于α和β的尺度对优化目标没有影响，因此可得等价的优化问题为

$$\arg\max_{\alpha,\beta} \alpha^\top K_{XX} K_{YY} \beta,$$

$$\text{s.t. } \alpha^\top K_{XX} K_{XX} \alpha + \kappa \alpha^\top K_{XX} \alpha = 1,$$

$$\beta^\top K_{YY} K_{YY} \beta + \kappa \beta^\top K_{YY} \beta = 1.$$

使用拉格朗日乘子法，可以得到与只等价的无约束的优化问题

$$\arg\max_{\alpha,\beta} \alpha^\top K_{XX} K_{YY} \beta + \lambda_1 (1 - \alpha^\top K_{XX} K_{XX} \alpha - \kappa \alpha^\top K_{XX} \alpha)$$

$$+ \lambda_2 (1 - \beta^\top K_{YY} K_{YY} \beta + \kappa \beta^\top K_{YY} \beta),$$

$$\arg\max_{\alpha,\beta} \alpha^\top K_{XX} K_{YY} \beta + \lambda_1 (1 - \alpha^\top K_{XX} K_{XX} \alpha - \kappa \alpha^\top K_{XX} \alpha)$$
$$+ \lambda_2 (1 - \beta^\top K_{YY} K_{YY} \beta + \kappa \beta^\top K_{YY} \beta),$$

关于α和β求导

$$K_{XX} K_{YY} \beta - 2\lambda_1 (K_{XX} K_{XX} \alpha + \kappa K_{XX} \alpha) = 0,$$
$$K_{YY} K_{XX} \alpha - 2\lambda_2 (K_{YY} K_{YY} \beta + \kappa K_{YY} \beta) = 0.$$

根据上式和约束条件可得$\lambda_1 = \lambda_2$. 令$\lambda_1 = \lambda_2 = \lambda$，根据上式可得$\beta$满足

$$\beta = \frac{(K_{YY} + \kappa \mathbf{I})^{-1} K_{XX} \alpha}{2\lambda}.$$

将上式代入 $K_{XX} K_{YY} \beta - 2\lambda_1 (K_{XX} K_{XX} \alpha + \kappa K_{XX} \alpha) = 0$，可得原优化问题等价于求解如下特征值问题

$$(K_{XX} + \kappa \mathbf{I})^{-1} K_{YY} (K_{YY} + \kappa \mathbf{I})^{-1} K_{XX} \alpha = 4\lambda^2 \alpha$$

参考文献

1. Hotelling H. Analysis of a Complex of Statistical Variables into Principal Components[J]. Journal of Educational Psychology, 1933, 24(6): 417-441.
2. Tipping M E, Bishop C M. Probabilistic Principal Component Analysis[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999, 61(3): 611-622.
3. Schölkopf B, Smola A, Müller K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem[J]. Neural Computation, 1998, 10(5): 1299-1319.
4. Harold H. Relations between Two Sets of Variates[J]. Biometrika, 1936, 28(3-4): 321-377.
5. Lai P L, Fyfe C. Kernel and Nonlinear Canonical Correlation Analysis[J]. International Journal of Neural Systems, 2000, 10(5): 365-377.