MACHINE LEARNING

Linear and Kernel
Canonical Correlation Analysis

Canonical Correlation Analysis (CCA)

GOAL:

Determine features in two (or more) separate descriptions of the dataset such that jointly these features represent well the dataset.

Applicable to datasets that are multimodal:

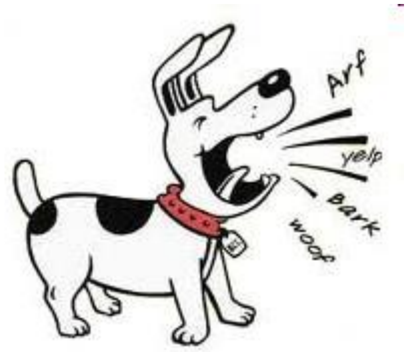
- audio & images/video
- biometric data (size, fingerprint, hair color, etc.)
- text and speech

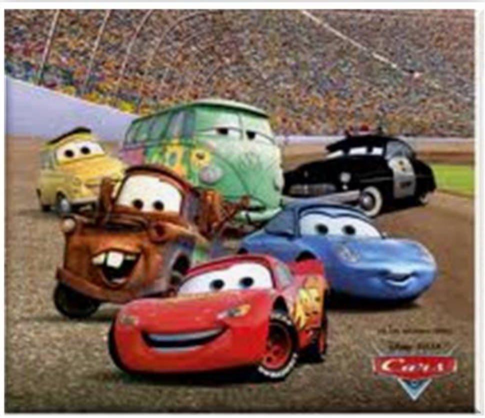
CCA is useful when the modalities have very different characteristics:

- different dimensions
- different features

CCA Principle

$x \in \mathbb{R}^{N_x}$ $\quad$ $y \in \mathbb{R}^{N_y}$





Video description $\quad$ **Audio description**

Extract hidden structure in each modality.

Search projections in X and Y.

$\{x^1, y^1\}$ $\quad$ $w_x \in \mathbb{R}^{N_x}$

$\{x^2, y^2\}$ $\quad$ $w_y \in \mathbb{R}^{N_y}$

$$\max_{w^x, w^y} corr\left(w_x^T x, w_y^T y\right)$$

CCA Derivation

Dataset if composed of M pairs of multidimensional variables

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

Search two projections w_x and w_y

$$\max_{w^x, w^y} corr\left(w_x^T X, w_y^T Y \right)$$

Crosscovariance matrix

C_{xy} is $N_x \times N_y$

Measure crosscorrelation between X and Y.

$$= \max_{w_x, w_y} \frac{w_x^T E\left\{ XY^T \right\} w_y}{\left\| w_x^T X \right\| \left\| w_y^T Y \right\|} = \max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}}$$

With X and Y zero mean, i.e. $E\{X\} = E\{Y\} = 0$

Covariance matrices

$C_{xx} = E\left\{ XX^T \right\}: N_x \times N_x$

$C_{yy} = E\left\{ YY^T \right\}: N_y \times N_y$

CCA Derivation

Correlation not affected by rescaling the norm of the vectors,

$\Rightarrow$ we can ask that $\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xx} w_x = \mathrm{w}_{\mathrm{y}}^{\mathrm{T}} C_{yy} w_y = 1$

$$\max \rho = \max_{w_x, w_y} \mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xy} \mathrm{w}_{\mathrm{y}}$$

u. c. $\mathrm{w}_{\mathrm{x}}^{\mathrm{T}} C_{xx} w_x = \mathrm{w}_{\mathrm{y}}^{\mathrm{T}} C_{yy} w_y = 1$

To determine the optimum (maximum) of ρ, solve by Lagrange:

$$L\left(w_x, w_y, \lambda_x, \lambda_y\right) = w_x^T C_{xy} w_y - \lambda_x \left(w_x^T C_{xx} w_x - 1\right) - \lambda_y \left(w_y^T C_{yy} w_y - 1\right)$$

Taking the partial derivatives over w_x, w_y

$$C_{xy} w_y = 2\lambda_x C_{xx} w_x$$

$$C_{yx} w_x = 2\lambda_y C_{yy} w_y$$

Multiply each equation by w_x and w_y respectively and substracting $\Rightarrow \lambda_x = \lambda_y := \lambda / 2$

CCA Solution

Replacing λ_x and λ_y by $\lambda / 2$, the partial derivatives become:

$$C_{xy} w_y = \lambda C_{xx} w_x$$

$$C_{yx} w_x = \lambda C_{yy} w_y$$

$\Rightarrow$ Which can be rewritten as

$$C_{xy} C_{yy}^{-1} C_{yx} w_x = \lambda^2 C_{xx} w_x$$

Generalized Eigenvalue Problem;
It can be reduced to a classical eigenvalue problem if C_{xx} is invertible

Solving for w_y gives:

$$C_{yx} C_{xx}^{-1} C_{xy} w_y = \lambda^2 C_{yy} w_y$$

If C_{yy} is invertible, it becomes an eigenvalue problem as for w_y.

These two eigenvalue problems yield a pair of q vectors $\left\{ w_x^i, w_y^i \right\}_{i=1..q}$, where $q = \min(N_x, N_y)$

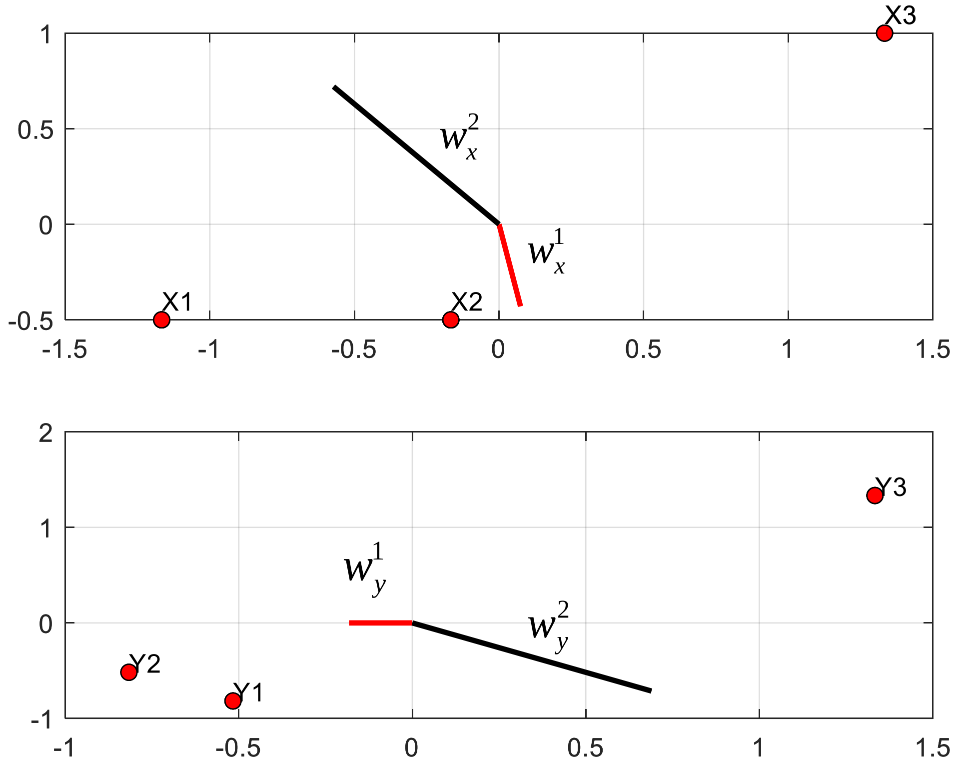
$w_x^i \in \mathbb{R}^{N_x}, w_y^i \in \mathbb{R}^{N_y}$

CCA Solution

The projection vectors can be visualized in original space.

If x and y are 2-dimensional spaces, we have at most 2 pairs of projections.

$\left\{w_x^1, w_y^1\right\}$ and $\left\{w_x^2, w_y^2\right\}$



Kernel Canonical Correlation Analysis

- CCA assumes linear projections in each space.
- Kernel CCA extends CCA to discover correlations in non-linear features.
- As for kPCA, kCCA will exploit the fact that CCA depends on computing inner product across datapoints, and replace these by the kernel function to apply linear CCA in feature space.

kCCA Principle

$x \in \mathbb{R}^{N_x}$

$y \in \mathbb{R}^{N_y}$

$\{x^1, y^1\}$

$$\max_{w^x, w^y} corr\left(w_x^T \phi_x(x), w_y^T \phi_y(y)\right)$$

$\{x^2, y^2\}$

Video description

Audio description

Assume two transformations

ϕ_x ϕ_y

And then perform correlation analysis in feature space across the two feature spaces.

kCCA derivation

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \in \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

Send into two separate feature spaces for data in *X* and in *Y.*

$$F_x = \left\{ \phi_x\left(x^i\right) \right\}_{i=1}^{M} \text{ and } F_y = \left\{ \phi_y\left(y^i\right) \right\}_{i=1}^{M}, \text{ with } E\left\{F_x\right\} = \sum_{i=1}^{M} \phi_x\left(x^i\right) = 0 \text{ and } E\left\{F_y\right\} = \sum_{i=1}^{M} \phi_y\left(y^i\right) = 0$$

Construct associated kernel matrices:

$K_x = F_x^T F_x, \; K_y = F_y^T F_y,$ columns of F_x, F_y are $\phi_x\left(x^i\right), \; \phi_y\left(y^i\right)$

kCCA derivation

In Linear CCA, we were solving for:

$$\max_{w_x, w_y} \mathrm{w}_x^T C_{xy} \mathrm{w}_y$$

$$\text{u.c.} \quad \mathrm{w}_x^T C_{xx} \underline{w_x} = \mathrm{w}_y^T C_{yy} \underline{w_y} = 1$$

In kernel CCA, we solve for:

$$\max_{w_x, w_y} \alpha_x^T \underbrace{F_x^T F_x}_{K_x} \underbrace{F_y^T F_y}_{K_y} \alpha_y$$

$$\text{u.c.} \quad \alpha_x^T \underbrace{F_x^T F_x F_x^T F_x}_{K_x} \alpha_x = \alpha_y^T \underbrace{F_y^T F_y F_y^T F_y}_{K_y} \alpha_y = 1$$

Express the projection vectors as a linear combination of images of datapoints in feature space (as in kPCA):

$w_x = F_x \alpha_x$ and $w_y = F_y \alpha_y$

$\Rightarrow w_x = \sum_{i=1}^{M} \alpha_{x,i} \phi_x\left(x^i\right)$ and $w_y = \sum_{i=1}^{M} \alpha_{y,i} \phi_y\left(y^i\right)$

Replace the covariance and crosscovariance matrices by the product of the projection vectors in feature space (as in kPCA):

$$C_{xx} = F_x F_x^T$$

$$C_{yy} = F_y F_y^T$$

$$C_{xy} = F_x F_y^T$$

kCCA Solution

$$\max_{w_x,w_y} \rho = \max_{\alpha_x,\alpha_y} \alpha_x^T K_x K_y \alpha_y$$

$$u.c. \left(\alpha_x^T K_x^2 \alpha_x\right) = \left(\alpha_y^T K_y^2 \alpha_y\right) = 1$$

Generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} K_x^2 & 0 \\ 0 & K_y^2 \end{pmatrix} \begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

This is again a generalized eigenvalue problem with α_x, α_y the dual eigenvectors (as dual eigenvectors in kPCA), see documentation in annexes for derivation.

kCCA Solution

If the intersection between the spaces spanned by $K_x\alpha_x,\ K_y\alpha_y$ is non-zero (with no centering), then the problem has a trivial solution, as $\rho \sim \cos\left(K_x\alpha_x, K_y\alpha_y\right)=1$ (see solution to the exercises).

Generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_xK_y \\ K_yK_x & 0 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix} = \lambda \begin{pmatrix} K_x^2 & 0 \\ 0 & K_y^2 \end{pmatrix}\begin{pmatrix} \alpha_x \\ \alpha_y \end{pmatrix}$$

Add a regularization term to increase the rank of the matrix and make it invertible (to avoid the trivial solution)

$$K_x^2 \rightarrow \left(K_x + \frac{M\kappa}{2}\mathrm{I}\right)^2 \quad , \quad \kappa > 0$$

kCCA for multiple modalities

$$X = \left\{ x^i \in \mathbb{R}^{N_x} \right\}_{i=1}^{M}, Y = \in \left\{ y^i \in \mathbb{R}^{N_y} \right\}_{i=1}^{M}$$

2-modalities

Can be extended to multiple modalities

L subdatasets: $X_1, ..., X_L$ with M observations each

Dimensions $N_1, ... N_L$:, i.e. $X_i : N_i \times M$

Applying L non-linear transformations ϕ_i, to $X_1, ... X_L$, resp.

$\rightarrow$ construct L Gram matrices: $K_1, ..., K_L$

$$\begin{pmatrix} 0 & K_1K_2 & .... & K_1K_L \\ K_2K_1 & 0 & .... & K_2K_L \\ . & & & \\ . & & & \\ K_LK_1 & K_LK_2 & .... & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_L \end{pmatrix} = \lambda \begin{pmatrix} \left(K_1 + \frac{M\kappa}{2} \mathrm{I} \right)^2 & & 0 \\ & .... & \\ 0 & & \left(K_L + \frac{M\kappa}{2} \mathrm{I} \right)^2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_L \end{pmatrix}$$

Interpretating the solution of kCCA

We cannot observe the projection vectors w_i.

But we can observe the projections of the datapoints on these vectors.

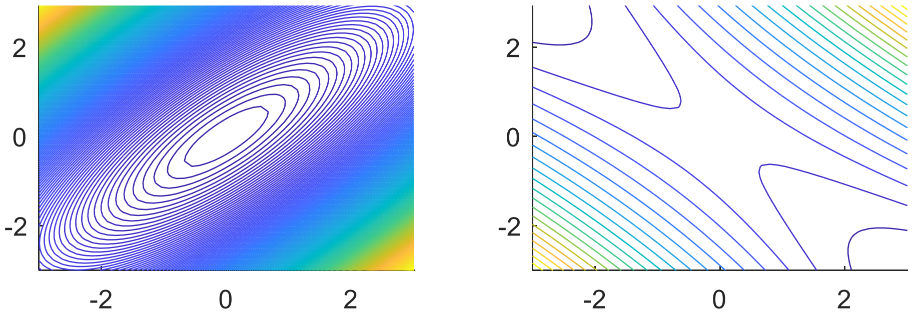
Recall that we have expressed the projection vectors as a linear combination of images of datapoints in feature space (as in kPCA):

$$w_x = \sum_{j=1}^{M} \alpha_{x,j} \phi_x\left(x^j\right)$$

$$\left\langle w_x, \phi(x) \right\rangle = \sum_{j=1}^{M} \alpha_{x,j} \underbrace{\left\langle \phi\left(x^j\right), \phi(x) \right\rangle}_{k\left(x^j, x\right)}$$

We can visualize the isolines solution:

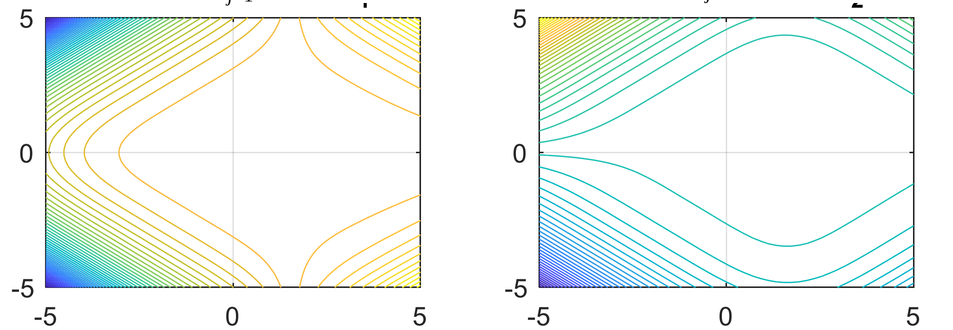
$$\left\langle w_x, \phi(x) \right\rangle = \sum_{j=1}^{M} \alpha_{x,j} k\left(x^j, x\right) = cst$$



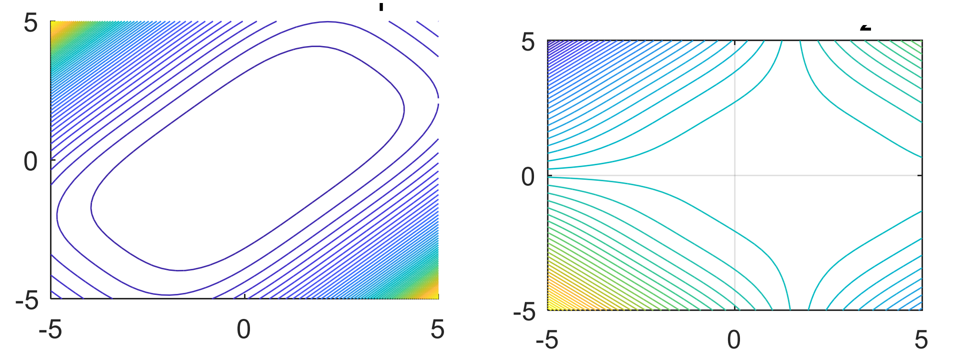
Homogeneous polynomial kernel $p = 2$

Example of Isolines in kCCA

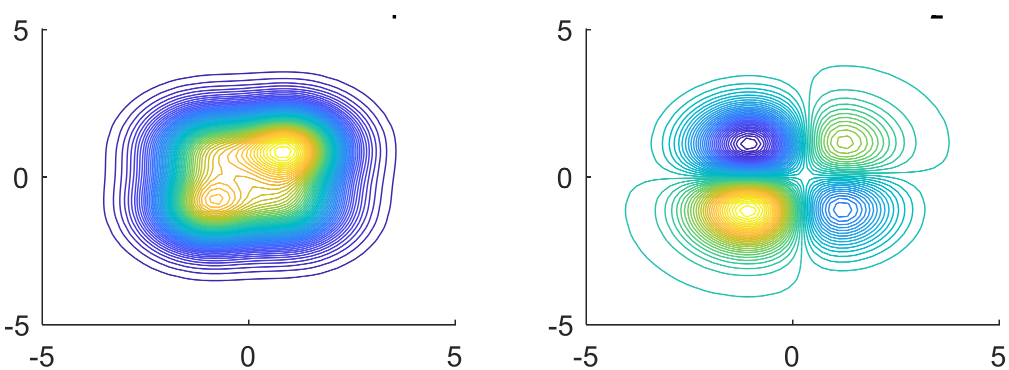
$$\left\langle w_x, \phi(x) \right\rangle = \sum_{j=1}^{M} \alpha_{x,j} k\left(x^j, x\right) = cst \qquad \left\langle w_y, \phi(y) \right\rangle = \sum_{j=1}^{M} \alpha_{y,j} k\left(y^j, y\right) = cst$$



Inhomogeneous polynomial kernel $p = 5,\ c = 1$



Inhomogeneous polynomial kernel $p = 4,\ c = 1$



RBF kernel

CCA and PCA

CCA is often thought of as a generalization of PCA.

- CCA resembles PCA in that it seeks to find correlations to reveal features. However, these are not the same correlations.
- CCA resembles PCA in that it can be solved in closed-form through an eigendecomposition of a matrix. But CCA and PCA have different matrices.
- CCA differs from PCA in that it finds different axes, in general.
- The axes found by PCA form an orthonormal basis of the space. This is not the case for CCA.
- The axes are not necessarily aligned with maximum variance in CCA.

CCA and kCCA: Summary

- CCA is an excellent mean to discover appropriate projections when your data is multi-modal.
- In each modality (separately), CCA finds projections that highlight features common to the datapoints as a whole.
- It generates projections that are different from performing PCA on each modality separately.
- The non-linear version of CCA, kernel CCA, generates sets of projections different from linear CCA and from kPCA.
- CCA and kCCA can be good pre-processing methods before performing more complex computation, such as clustering or classification.