Kernel-based Methods for Unsupervised Learning

LEAR project-team, INRIA

Zaid Harchaoui

Lyon, Janvier 2011

Outline

Outline

Machine learning : a tentative big picture

Unsupervised learning (*learning without a teacher*)

- Find structure of $\mathbf{x} \in \mathcal{X}$, given observations $\mathbf{x}_i, \ i = 1, ..., n$

Supervised learning (*learning with a teacher*)

- Predict $y \in \mathcal{Y}$ from $\mathbf{x} \in \mathcal{X}$, given observations $(\mathbf{x}_i, y_i), \ i = 1, ..., n$

Machine learning : a tentative big picture

Applications in many fields

- *Computer vision*
- Bioinformatics
- Audio/speech processing
- Text mining
- Computational astronomy
- etc.

Interplays

- interplay between statistics and optimization, with a look towards AI
- interplay between theory, algorithms, and real applications

Unsupervised learning

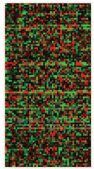
Dimension reduction

face images

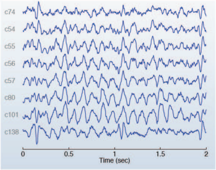
Zambian President Levy Mwanawasa has won a second term in office in an election his challenger Michael Sata accused him of rigging, official results showed on Monday.

According to media reports, a pair of hackers said on Saturday that the Firefox Web browser, commonly perceived as the safer and more customizable alternative to market leader Internet Explorer, is critically flawed. A presentation on the flaw was shown during the ToorCon hacker conference in San Diego.

documents



gene expression data



MEG readings

Unsupervised learning

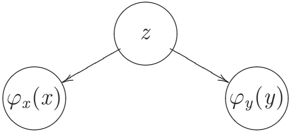
Dimension reduction

- Computational efficiency : space and time savings
- Statistical performance : fewer dimensions $\rightarrow$ regularization
- Visualization : discover underlying structure of the data

$\rightarrow$ PCA and KPCA

Unsupervised learning

Feature extraction



x: 

y: “A view from Idyllwild, California, with pine trees and snow capped Marion Mountain under a blue sky.”

Unsupervised learning

Feature extraction

- Multimodality : leverage the correlation between the modalities
- Statistical performance : take advantage of both views of the data
- Putting in relation : discover underlying relations between the modalities

$\rightarrow$ CCA and KCCA

Unsupervised learning

Clustering

Unsupervised learning

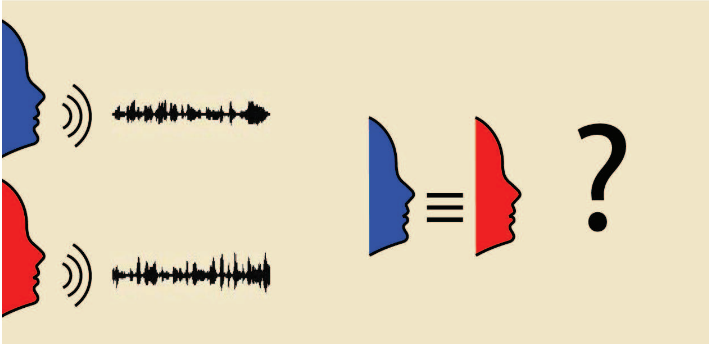
Clustering

- Semantics : grouping datapoints in meaningful clusters
- Statistical performance : intrinsic degrees of freedom of the data
- Visualization : discover groupings between datapoints

$\rightarrow$ spectral clustering and temporal segmentation

Unsupervised learning

Detection problems



Unsupervised learning

Detection problems

- Balance risks : control detection rate with a guaranteed false alarm probability
- Power : detect differences not only in mean or covariance

$\rightarrow$ homogeneity testing, change detection

Outline

Kernel methods

Machine Learning methods taking $\mathbf{K} = [k(X_i, X_j)]_{i,j=1,\dots,n}$ (Gram matrix as input for processing a sample $\{X_1, \dots, X_n\}$, where $k(x, y)$ is a similarity measure between x and y defining a positive definite kernel.

Strengths of Kernel Methods

- Minimal assumptions on data types (vectors, strings, trees, graphs, etc.)
- Interpretation of $k(x, y)$ as a dot product $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ in a reproducing kernel Hilbert space $\mathcal{H}$ where the observations are mapped *via* $[\phi : \mathcal{X} \to \mathcal{H}]$ the feature map $\phi(\bullet) = k(\bullet, \cdot)$

Kernel methods

Positive-definite kernel

- definition : given a set of objects $\mathcal{X}$, a positive definite kernel is a symmetric function $k(x, x')$ such that for all finite sequences of $x_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$,

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \ .$$

- Aronszajn theorem : k is a positive-definite kernel iif there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi(\cdot) : \mathcal{X} \to \mathcal{H}$ such that for any $x, x' \in \mathcal{X}$

$$k(x, x') = \left\langle \Phi(x), \Phi(x') \right\rangle_{\mathcal{H}} \ .$$

Kernel methods

Reproducing kernel Hilbert space

- Assume k is a positive definite kernel on $\mathcal{X} \times \mathcal{X}$
- Aronszajn theorem : k is a positive-definite kernel iif there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi(\cdot) : \mathcal{X} \to \mathcal{H}$ such that for any $x, x' \in \mathcal{X}$

$$k(x, x') = \left\langle \Phi(x), \Phi(x') \right\rangle_{\mathcal{H}} \ .$$

- Lexicon : $\mathcal{X} =$ Input space, $\mathcal{H} =$ Feature space, $\Phi(\cdot) =$ Feature map

Reproducing kernel Hilbert space

- Feature map is the Aronszajn map $\Phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$
- Function evaluation $f(\mathbf{x}) = \langle f, \Phi(\mathbf{x}) \rangle_{\mathcal{H}}$
- Reproducing property $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$

How does the feature space look like ?

Example : space of shapes of birds

How does the feature space look like ?

Feature map ?

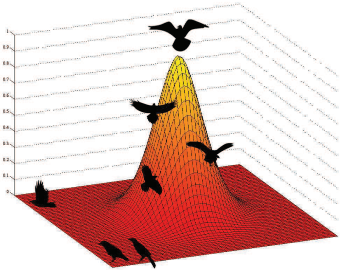
How does the feature map look like ?

$k\Big($  $, \cdot \Big)$

How does the feature space look like ?

Feature map ?

The feature map is a function whose values span the whole range of shapes with varying magnitudes.



Examples of Kernels

Kernels on vectors

$$
\begin{aligned}
\text{Polynomial} \quad & k(\mathbf{x}, \mathbf{y}) = (c + (\mathbf{x}, \mathbf{y}))^d \\
\text{Laplace} \quad & k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_1 / \sigma) \\
\text{RBF} \quad & k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_2^2 / \sigma^2)
\end{aligned}
$$

Examples of Kernels

Kernels on histograms

Kernels built on top of divergence between probability distributions

$$\psi_{JD}(\theta,\theta') = h\left(\frac{\theta+\theta'}{2}\right) - \frac{h(\theta)+h(\theta')}{2},$$

$$\psi_{\chi^2}(\theta,\theta') = \sum_i \frac{(\theta_i-\theta'_i)^2}{\theta_i+\theta'_i}, \quad \psi_{TV}(\theta,\theta') = \sum_i |\theta_i-\theta'_i|,$$

$$\psi_{H_2}(\theta,\theta') = \sum_i |\sqrt{\theta_i}-\sqrt{\theta'_i}|^2, \quad \psi_{H_1}(\theta,\theta') = \sum_i |\sqrt{\theta_i}-\sqrt{\theta'_i}|.$$

$$k(\theta,\theta') = \exp(-\psi(\theta,\theta')/\sigma^2)\ .$$

The kernel jungle

Kernels on histograms

- Pyramid match kernels (Grauman and Darrell, 2005)
- Multiresolution (nested histograms) kernels (Cuturi, 2006)
- Walk and tree-walk kernels (Ramon & Gaertner, 2004 ; Harchaoui & Bach, 2007 ; Mahe et al., 2007)

Kernels from statistical generative models

- Mutual Information Kernels (Seeger, 2002)
- Fisher kernels (see Shawe-Taylor & Cristianini, 2004)

Other kernels

- Kernels of shapes and point coulds (Bach, 2007)
- Kernels on time series (Cuturi, 2007)

How does the feature space look like ?

Classical kernel trick

- Describes what happens to pairs of examples
- Focuses on the *pointwise* effect of the feature map on an example

"Remixed" kernel trick

- Describes what happens to a random sample from a probability distribution
- Focuses on the *global* effect of the feature map on a sample

Outline

Coordinate-free definitions of mean and covariance

Usual definitions

- need explicit basis to define quantities
  $\rightarrow$ tricky in high-dimensional/∞-dimensional feature spaces

Coordinate-free definitions

- define quantities through their projections along any direction
  $\rightarrow$ allow direct application of the *reproducing property*

Mean vector and mean element

Empirical mean element

Empirical mean vector $\hat{\mu}$ of $X_1, \ldots, X_m \sim \mathbb{P}$

$$\forall \mathbf{w} \in \mathcal{X},$$

$$(\hat{\mu}, \mathbf{w}) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{x}_\ell, \mathbf{w})$$

Empirical mean element $\hat{\mu}$ of $X_1, \ldots, X_m \sim \mathbb{P}$

$$\forall f \in \mathcal{H},$$

$$\langle \hat{\mu}, f \rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} \langle \phi(\mathbf{x}_\ell), f \rangle_{\mathcal{H}}$$

Mean vector and mean element

Empirical mean element

Empirical *mean element* $\hat{\mu}$ of $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathbb{P}$

$$\forall f \in \mathcal{H},$$

$$\langle \hat{\mu}, f \rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} \langle \phi(\mathbf{x}_\ell), f \rangle_{\mathcal{H}}$$

$$\langle \hat{\mu}, f \rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} \langle k(\mathbf{x}_\ell, \cdot), f \rangle_{\mathcal{H}}$$

$$\stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} f(\mathbf{x}_\ell) \text{ (reproducing property)}$$

$$\stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} \sum_{j=1}^{n} \alpha_j k(\mathbf{x}_j, \mathbf{x}_\ell) \text{ , if } f(\cdot) = \sum_{j=1}^{n} \alpha_j k(\mathbf{x}_j, \cdot)$$

Centering in feature space

Gram matrix

$\mathbf{K} = [k(X_i, X_j)]_{i,j=1,\ldots,n}$ of all evaluations of the kernel $k(\cdot,\cdot)$ on the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Centering in feature space

To center all $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)$ *simultaneously*, do

$$\mathbf{K} \leftarrow \tilde{\mathbf{K}} = \Pi \mathbf{K} \Pi \ ,$$

where

$$\Pi = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T \ .$$

Covariance matrix and covariance operator

Empirical covariance operator

Empirical covariance matrix $\hat{\Sigma}$ of $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathbb{P}$

$$\forall \mathbf{w}, \mathbf{v} \in \mathcal{X},$$

$$(\mathbf{w}, \hat{\Sigma}\mathbf{v}) = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{w}, \tilde{\mathbf{x}}_\ell)(\tilde{\mathbf{x}}_\ell, \mathbf{v})$$

$$\tilde{\mathbf{x}}_\ell = \mathbf{x}_\ell - \hat{\mu}\ .$$

Empirical covariance operator $\hat{\Sigma}$ of $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathbb{P}$

$$\forall f, g \in \mathcal{H},$$

$$\left\langle f, \hat{\Sigma} g \right\rangle = \frac{1}{m} \sum_{\ell=1}^{m} \left\langle f, \tilde{\phi}(\mathbf{x}_\ell) \right\rangle \left\langle \tilde{\phi}(\mathbf{x}_\ell), g \right\rangle$$

$$\tilde{\phi}(\mathbf{x}_\ell) = \phi(\mathbf{x}_\ell) - \hat{\mu}\ .$$

Covariance matrix and covariance operator

Covariance operator

Empirical covariance operator $\hat{\Sigma}$ of $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathbb{P}$

$$
\begin{aligned}
&\forall f, g \in \mathcal{H}, \\
&\left\langle f, \hat{\Sigma} g \right\rangle = \frac{1}{m} \sum_{\ell=1}^{m} \left\langle f, \tilde{\phi}(\mathbf{x}_\ell) \right\rangle \left\langle \tilde{\phi}(\mathbf{x}_\ell), g \right\rangle \\
&\qquad\qquad = \frac{1}{m} \sum_{\ell=1}^{m} \{ f(\mathbf{x}_\ell) - \langle \hat{\mu}, f \rangle_{\mathcal{H}} \} \{ f(\mathbf{x}_\ell) - \langle \hat{\mu}, g \rangle_{\mathcal{H}} \} \; .
\end{aligned}
$$

Computing variance along a direction in feature space

Gram matrix

$\mathbf{K} = [k(X_i, X_j)]_{i,j=1,\ldots,n}$ of all evaluations of the kernel $k(\cdot,\cdot)$ on $x_1, \ldots, x_n$.

Covariance along two directions

$$\left\langle f, \hat{\Sigma} g \right\rangle = \frac{1}{m} \alpha^T \tilde{\mathbf{K}} \tilde{\mathbf{K}} \beta \; ,$$

where

$$f(\cdot) = \sum_{j=1}^{n} \alpha_j k(\mathbf{x}_j, \cdot) \; ,$$

$$g(\cdot) = \sum_{j=1}^{n} \beta_j k(\mathbf{x}_j, \cdot) \; .$$

Mean element and covariance operator

Population mean element and covariance operator

Population mean element μ and population covariance operator Σ of $\mathbf{x} \sim \mathbb{P}$

$$\langle \mu, f \rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \mathbb{E}[f(\mathbf{x})] \,, \quad \forall f \in \mathcal{H}$$

$$\langle f, \Sigma g \rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \mathrm{Cov}[f(\mathbf{x}), g(\mathbf{x})] \,, \quad \forall f, g \in \mathcal{H}$$

Empirical mean element and covariance operator

Empirical mean element $\hat{\mu}$ and empirical covariance operator $\hat{\Sigma}$ of $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathbb{P}$

$$\langle \hat{\mu}, f \rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} f(\mathbf{x}_\ell) \,, \quad \forall f \in \mathcal{H}$$

$$\left\langle f, \hat{\Sigma} g \right\rangle_{\mathcal{H}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{\ell=1}^{m} \{ f(\mathbf{x}_\ell) - \langle \hat{\mu}, f \rangle_{\mathcal{H}} \} \{ f(\mathbf{x}_\ell) - \langle \hat{\mu}, g \rangle_{\mathcal{H}} \} \quad \forall f, g \in \mathcal{H}$$

Some casual considerations before the real stuff

Supervised learning

- least-square regression, kernel ridge regression, multilayer-perceptron
  $\rightarrow$ tackled through (possibly a sequence of) linear of systems
- Operation \ in Matlab/Octave

Unsupervised learning

- (kernel) principal component analysis, (kernel) canonical correlation analysis, spectral clustering
  $\rightarrow$ tackled through (possibly a sequence of) eigenvalue problems
- Function eigs in Matlab/Octave

Outline

Kernel Principal Component Analysis (Schölkopf et al., 1998 ; Shawe-Taylor & Cristianini, 2004)

Principal Component Analysis (PCA)

A brief refresher

- Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ a dataset of points in $\mathbf{R}^d$
- PCA is a classical method in multivariate statistics to define a set of orthogonal directions, called *principal components*, that capture the maximum variance
- Projection along the first 2-3 principal components allows to visualize the dataset

Refresher on Principal Component Analysis

Computational aspects

- Maximum variance criterion corresponds to a Rayleigh quotient
- PCA boils down to an eigenvalue problem on the *centered* covariance matrix $\hat{\Sigma}$ of the dataset, *i.e.* the principal components $\mathbf{w}_1, \ldots, \mathbf{w}_d$ are the eigenvectors of $\hat{\Sigma}$ (assuming $n > d$)
- Computational complexity : $O(ndc)$ in time with a *Singular Value Decomposition* (SVD ; see `eigs` in Matlab/Octave), with n the number of points, d the dimension, c the number of principal components retained ; stochastic approximation version for nonstationary/large-scale datasets.

Variance along a direction and Rayleigh quotients

Variance along a direction

PCA seeks for directions $\mathbf{w}_1, \ldots, \mathbf{w}_c$ such that

$$
\begin{aligned}
\mathbf{w}_j &= \mathrm{argmax}_{\mathbf{w} \in \mathbb{R}^d; \mathbf{w}_j \perp \{\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}\}} \mathrm{Var}_{\mathsf{emp}} \frac{(\mathbf{w}, \mathbf{x})}{(\mathbf{w}, \mathbf{w})} \\
&= \mathrm{argmax}_{\mathbf{w} \in \mathbb{R}^d; \mathbf{w}_j \perp \{\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}\}} \frac{1}{m} \sum_{i=1}^{m} \frac{(\mathbf{w}, \mathbf{x}_i)^2}{(\mathbf{w}, \mathbf{w})} \\
&= \mathrm{argmax}_{\mathbf{w} \in \mathbb{R}^d; \mathbf{w}_j \perp \{\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}\}} \underbrace{\frac{(\mathbf{w}, \hat{\Sigma}\mathbf{w})}{(\mathbf{w}, \mathbf{w})}}_{\text{Rayleigh quotient}} \quad .
\end{aligned}
$$

Principal components $\mathbf{w}_1, \ldots, \mathbf{w}_c$ are the first c eigenvectors of $\hat{\Sigma}$.

Variance along a direction and Rayleigh quotients

Variance along a direction

KPCA seeks for directions $f_1, \ldots, f_c$ such that

$$f_j = \mathrm{argmax}_{f \in \mathcal{H}; f_j \perp \{f_1, \ldots, f_{j-1}\}} \mathrm{Var}_{\mathsf{emp}} \frac{\langle f, \phi(\mathbf{x}) \rangle}{\langle f, f \rangle}$$

$$= \mathrm{argmax}_{f \in \mathcal{H}; f_j \perp \{f_1, \ldots, f_{j-1}\}} \frac{1}{m} \sum_{i=1}^{m} \frac{\langle f, \phi(\mathbf{x}_i) \rangle^2}{\langle f, f \rangle}$$

$$= \mathrm{argmax}_{f \in \mathcal{H}; f_j \perp \{f_1, \ldots, f_{j-1}\}} \underbrace{\frac{\left\langle f, \hat{\Sigma} f \right\rangle}{\langle f, f \rangle}}_{\text{Rayleigh quotient}} .$$

Principal components $f_1, \ldots, f_c$ are the first c eigenvectors of $\hat{\Sigma}$. Is that it ?

Rescue theorems

Properties of covariance operators

RKHS Covariance operators are (Zwald et al., 2005, Harchaoui et al., 2008)

- self-adjoint (∞-dimensional counterpart of symmetric)
- positive
- trace-class

Consequence

The covariance operator $\hat{\Sigma}$ and the centered Gram matrix $\tilde{\mathbf{K}}$ share the same eigenvalues on the nonzero part of their spectra, and their eigenvectors are related by a simple relation.

Kernel Principal Component Analysis

KPCA algorithm

- Center the Gram matrix
- Performs an SVD on $\tilde{\mathbf{K}}$ to get the first c eigenvector/eigenvalue pairs $(e_j, \lambda_j)_{j=1,\dots,c}$.
- Normalize the eigenvector $\tilde{e}_j \leftarrow e_j/\lambda_j$
- Projections onto the j-th eigenvectors is given by $\tilde{\mathbf{K}}\tilde{e}_j$

Computational aspects of KPCA

Computational aspects

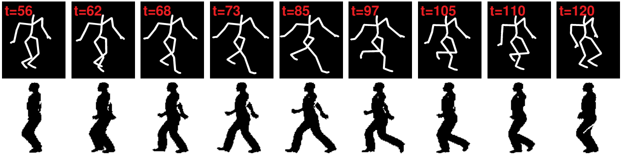
- Maximum variance in feature space corresponds to a Rayleigh quotient
- KPCA boils down to an eigenvalue problem involving the centered auto-covariance matrices $\tilde{\mathbf{K}}$
- Computational complexity : $O(cn^2)$ in time with a *Singular Value Decomposition* (SVD ; see `eigs` in Matlab/Octave), with n the number of points, c the number of principal components retained ; stochastic approximation version for nonstationary/large-scale datasets.

Low-dimensional representation with KPCA

Human body pose representation

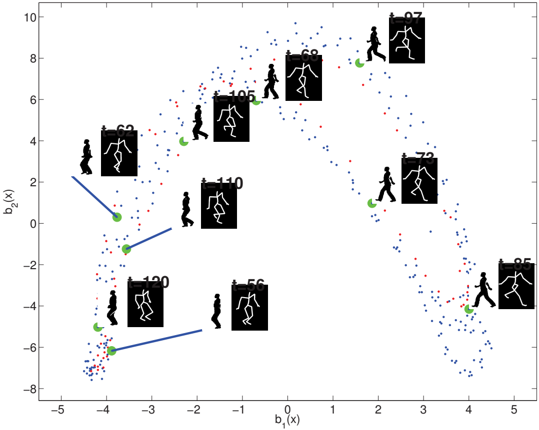
- Walking sequence of length 400 (containing about 3 walking cycles) obtained from the CMU Mocap database
- Data : silhouette images of size (160 100) taken at a side view

Human body pose representation (Kim & Pavlovic, 2008)



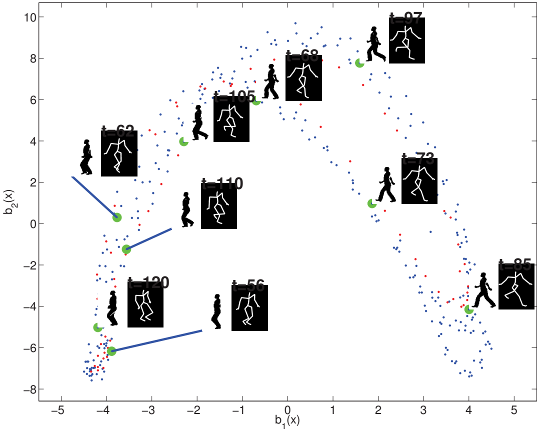
Low-dimensional representation with KPCA

Human body pose representation



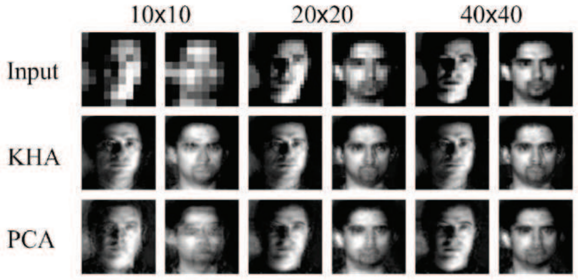
Low-dimensional representation with KPCA

Human body pose representation



Super-resoluton with KPCA (Kim et al., 2005)

Super-resolution



KPCA+n : unsupervised alignment (de la Torre & Nguyen, 2009)

Unsupervised alignment

KPCA + Rigid motion model

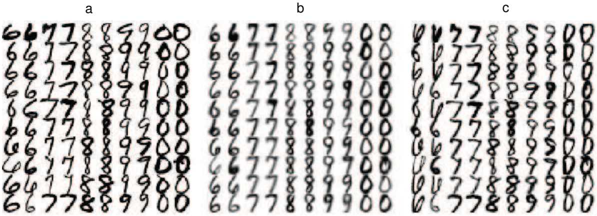


Figure: original USPS (a), Parametrized KPCA (b), congealing (c)

Applications

Popular

- Image denoising (digits, faces, etc.)
- Visualization of bioinformatics data (strings, proteins, etc.)
- Dimension-reduction of high-dimensional features (appearance, interest points, etc.)

Not so well-know property of KPCA

- Regularization in supervised learning can be enforced by projection $\rightarrow$ careful not to regularize twice!
- Useful in settings where ridge-regularization is impractical (Zwald et al., 2009; Harchaoui et al., 2009; Guillaumin et al., 2010)

Outline

Kernel Canonical Correlation Analysis (Shawe-Taylor & Cristianini, 2004)

Canonical Correlation Analysis (CCA)

A brief refresher

- Let $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ a dataset of points in $\mathbf{R}^d \times \mathbf{R}^p$, for which two *views* are available : the "x-view" and the "y-view"
- CCA is a classical method from multivariate statistics to define a set of pairs of orthogonal directions, called *canonical variates*, that capture the *maximum correlation* between the two views.
- Projection along the first 2-3 pairs of canonical variates resp. of "x-view" and the "y-view" allows to visualize the components dataset maximizing the correlation between the two views.

Refresher on Canonical Correlation Analysis

Computational aspects

- Maximum correlation criterion corresponds to a generalized Rayleigh quotient
- CCA boils down to a generalized eigenvalue problem involving the (centered) auto-covariance matrices $\hat{\Sigma}_{\mathbf{xx}}$ and $\hat{\Sigma}_{\mathbf{yy}}$ and on the (centered) cross-covariance matrix $\hat{\Sigma}_{\mathbf{xy}}$
- Computational complexity : $O(n(d+p)c)$ in time with a *Singular Value Decomposition* (SVD ; see `eigs` in Matlab/Octave), with n the number of points, d the dimension, c the number of canonical variates retained ; stochastic approximation version for nonstationary/large-scale datasets.

Cross-covariance matrix and cross-covariance operator

Empirical cross-covariance matrix

Empirical cross-covariance matrix $\hat{\Sigma}_{\mathbf{xy}}$ of $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathbb{P}_{\mathbf{x}}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_m \sim \mathbb{P}_{\mathbf{y}}$

$$\forall \mathbf{w}, \mathbf{v} \in \mathcal{X}, \mathcal{Y}$$

$$(\mathbf{w}, \hat{\Sigma}_{\mathbf{xy}} \mathbf{v}) = \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{w}, \tilde{\mathbf{x}}_\ell)(\tilde{\mathbf{y}}_\ell, \mathbf{v})$$

$$\tilde{\mathbf{x}}_\ell = \mathbf{x}_\ell - \hat{\mu}_{\mathbf{x}}$$

$$\tilde{\mathbf{y}}_\ell = \mathbf{y}_\ell - \hat{\mu}_{\mathbf{y}} \ .$$

Empirical cross-covariance operator $\hat{\Sigma}_{\mathbf{xy}}$ of $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mathbb{P}_{\mathbf{x}}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_m \sim \mathbb{P}_{\mathbf{y}}$

$$\forall f, g \in \mathcal{F}, \mathcal{H}$$

$$\left\langle f, \hat{\Sigma}_{\mathbf{xy}} g \right\rangle = \frac{1}{m} \sum_{\ell=1}^{m} \left\langle f, \tilde{\phi}(\mathbf{x}_\ell) \right\rangle \left\langle \tilde{\psi}(\mathbf{y}_\ell), g \right\rangle$$

$$\tilde{\phi}(\mathbf{x}_\ell) = \phi(\mathbf{x}_\ell) - \hat{\mu}_{\mathbf{x}}$$

$$\tilde{\psi}(\mathbf{y}_\ell) = \psi(\mathbf{y}_\ell) - \hat{\mu}_{\mathbf{y}} \ .$$

Covariance along two directions and generalized Rayleigh quotients

Covariance along two directions

CCA seeks for directions $(\mathbf{w}_1, \mathbf{v}_1)$ such that [1]

$$
\begin{aligned}
(\mathbf{w}_1, \mathbf{v}_1) &= \mathrm{argmax}_{(\mathbf{w},\mathbf{v})\in\mathbb{R}^d\times\mathbb{R}^p} \frac{\mathrm{Cov}((\mathbf{w},\mathbf{x}),(\mathbf{v},\mathbf{y}))}{\mathrm{Var}^{1/2}((\mathbf{w},\mathbf{x})\mathrm{Var}^{1/2}((\mathbf{v},\mathbf{y})} \\
&= \mathrm{argmax}_{(\mathbf{w},\mathbf{v})\in\mathbb{R}^d\times\mathbb{R}^p} \frac{(\mathbf{w},\hat{\Sigma}_{\mathbf{xy}}\mathbf{v})}{(\mathbf{w},\hat{\Sigma}_{\mathbf{xx}}\mathbf{w})^{1/2}(\mathbf{v},\hat{\Sigma}_{\mathbf{yy}}\mathbf{v})^{1/2}} .
\end{aligned}
$$

1. focus here on the first pair of canonical variates

Covariance along two directions and generalized Rayleigh quotients

Generalized Rayleigh quotient

Canonical variates $(\mathbf{w}_1, \mathbf{v}_1), \ldots, (\mathbf{w}_c, \mathbf{v}_c)$ are the first c pairs of vectors solutions of the generalized eigenvalue problem

$$\begin{bmatrix} \mathbf{0} & \hat{\Sigma}_{\mathbf{xy}} \\ \hat{\Sigma}_{\mathbf{xy}} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{v} \end{pmatrix} = \rho \begin{bmatrix} \hat{\Sigma}_{\mathbf{xx}} & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{\mathbf{yy}} \end{bmatrix} \begin{pmatrix} \mathbf{w} \\ \mathbf{v} \end{pmatrix} .$$

Covariance along two directions and generalized Rayleigh quotients

Covariance along two directions

Kernel CCA seeks for directions (f_1, g_1) such that [2]

$$
\begin{aligned}
(f_1, g_1) &= \mathrm{argmax}_{(f,g)\in\mathcal{H}\times\mathcal{H}} \frac{\mathrm{Cov}(\langle f, \phi(\mathbf{x})\rangle, \langle g, \psi(\mathbf{y})\rangle)}{\{\mathrm{Var}\,\langle f, \phi(x)\rangle + \epsilon\,\langle f, f\rangle\}^{1/2}\{\mathrm{Var}\,\langle g, \psi(x)\rangle + \epsilon\,\langle g, g\rangle\}^{1/2}} \\
&= \mathrm{argmax}_{(f,g)\in\mathcal{H}\times\mathcal{H}} \frac{\left\langle f, \hat{\Sigma}_{\mathbf{xy}} g\right\rangle}{\left\langle f, (\hat{\Sigma}_{\mathbf{xx}} + \frac{n\epsilon}{2})g\right\rangle^{1/2}\left\langle f, (\hat{\Sigma}_{\mathbf{yy}} + \frac{n\epsilon}{2})g\right\rangle^{1/2}}\,.
\end{aligned}
$$

2. focus here on the first pair of canonical variates

Correlation along two directions

Generalized eigenvalue problem

Coefficients of canonical variates $(\alpha_1, \beta_1), \ldots, (\alpha_c, \beta_c)$ are the first c pairs of vectors solutions of the generalized eigenvalue problem

$$\left[\begin{array}{cc} \mathbf{0} & \tilde{\mathbf{K}_x}\tilde{\mathbf{K}_y} \\ \tilde{\mathbf{K}_x}\tilde{\mathbf{K}_y} & \mathbf{0} \end{array}\right] \left(\begin{array}{c} \alpha \\ \beta \end{array}\right) = \rho \left[\begin{array}{cc} \tilde{\mathbf{K}_x}\tilde{\mathbf{K}_x} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{K}_y}\tilde{\mathbf{K}_y} \end{array}\right] \left(\begin{array}{c} \alpha \\ \beta \end{array}\right) .$$

Computational aspects of KCCA

Computational aspects

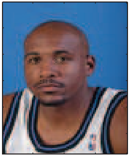
- Maximum correlation in feature space corresponds to a Rayleigh quotient
- KCCA boils down to a generalized eigenvalue problem involving the squared centered Gram matrices matrices $\tilde{\mathbf{K}_x}^2$ $\tilde{\mathbf{K}_y}^2$ and the product of the Gram matrices $\tilde{\mathbf{K}_x}\tilde{\mathbf{K}_y}$.
- Computational complexity : $O(cn^2)$ in time with a *Singular Value Decomposition* (SVD ; see `eigs` in Matlab/Octave), with n the number of points, c the number of principal components retained ; stochastic approximation version for nonstationary/large-scale datasets.

Multimedia content based image retrieval with KCCA

Multimedia

- Multimedia content → multi-view data
- images with text captions : text → “x”-view, image → “y”-view

Multimedia content based image retrieval (Hardoon et al, 2004)



I_1

I_2

I_3

Image	Label	Keywords
I_1	Sports	position college weight born lbs height guard
I_2	Aviation	na air convair wing
I_3	Paintball	check darkside force gog strike odt

Outline

Outline

Outline