STATG019 – Selected Topics in Statistics 2015

Lecture 5

Unsupervised Kernel Methods

Dr Franz J. Király

Response	Average	Total
String kernels: combinatorial kernels for text mining, document classification and genome analysis	12%	4
Graph kernels: combinatorial kernels on graphs and between graphs for learning molecules, or biological and social networks	9%	3
Kernel quantile regression: predicting the median and other quantiles in non-linear distributional data, e.g. population analysis	3%	1
Kernel CCA: finding highly correlating non-linear features in high-dimensional data, e.g. time series	15%	5
Kernel k-means: non-linear clustering with kernels	9%	3
More on novelty and outlier detection with kernels	6%	2
Vapnik-Chervonenkis learning theory; the VC inequality and the main ideas behind its proof	9%	3
Cross-validation techniques in general and for kernels in particular	9%	3
Kernel on-line learning: how to modify kernel methods to cope with sequential data; algorithmic techniques and learning guarantees	12%	4
Kernels for big data: how to cope with huge data sets; kernel Hebbian, Nyström approximation, sub-sampling, inducing variables	12%	4

Today

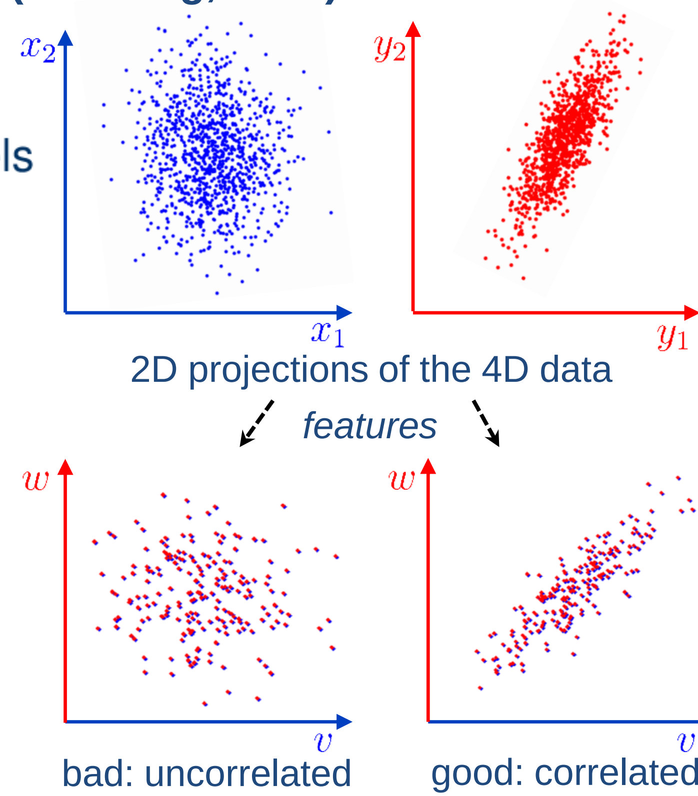
Kernel Canonical Correlation Analysis

Canonical Correlation Analysis (Hotelling, 1936)

Input: data points $(x_1, y_1), \dots, (x_N, y_N) \in \mathbb{R}^n \times \mathbb{R}^m$

unsupervised: neither x_i nor y_i are interpreted as labels but as *two equitable classes* of covariates

(for readability assume centered data, i.e. $\sum_{i=1}^N x_i = \sum_{i=1}^N y_i = 0$)



Output: Linear features from both covariate classes with high correlation between each other

Mathematically: coordinates $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$ maximizing correlation between $\langle v, x_i \rangle$ and $\langle w, y_i \rangle$

$$v, w = \underset{v,w}{\operatorname{argmax}} \, |\operatorname{corr}(Xv, Yw)| = \underset{v,w}{\operatorname{argmax}} \frac{(v^\top X^\top Y w)^2}{v^\top X^\top X v \cdot w^\top Y^\top Y w}$$

Remarks: optimal v, w are non-unique

for maximizers v, w and $\alpha, \beta \in \mathbb{R}$, scaled directions $\alpha v, \beta w$ are also maximizers

Good idea: posit $\|v\| = \|w\| = 1$ **Better idea:** posit $\|Xv\| = \|Yw\| = 1$

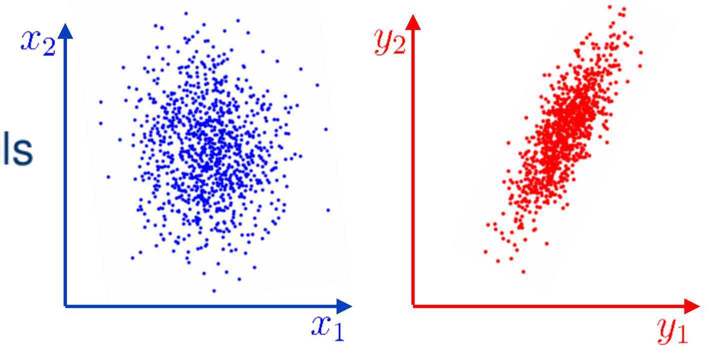
Yields **quadratic program** (quadratically constrained):

$$v, w = \underset{v,w}{\operatorname{argmax}} \, v^\top X^\top Y w = \underset{v,w}{\operatorname{argmax}} \, (v^\top X^\top Y w)^2 \quad \text{s.t.} \quad \begin{aligned} v^\top X^\top X v &= 1 \\ w^\top Y^\top Y w &= 1 \end{aligned}$$

Canonical Correlation Analysis (Hotelling, 1936)

Input: data points $(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^n \times \mathbb{R}^m$

unsupervised: neither x_i nor y_i are interpreted as labels but as *two equitable classes* of covariates

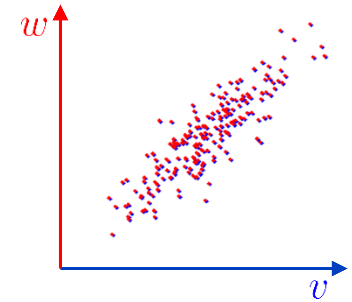


2D projections of the 4D data

features

Output: coordinates $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$

maximizing correlation between $\langle v, x_i \rangle$ and $\langle w, y_i \rangle$



good: correlated

Quadratic program (quadratically constrained):

$$v, w = \operatorname*{argmax}_{v,w} v^\top X^\top Y w = \operatorname*{argmax}_{v,w} \left(v^\top X^\top Y w\right)^2 \quad \text{s.t.} \quad \begin{array}{l} v^\top X^\top X v = 1 \\ w^\top Y^\top Y w = 1 \end{array}$$

Solution by the Lagrangian approach:

$$L(\lambda_v, \lambda_w, v, w) = v^\top X^\top Y w - \frac{\lambda_v}{2}(Xv)^\top(Xv) - \frac{\lambda_w}{2}(Yw)^\top(Yw) + \frac{\lambda_v + \lambda_w}{2}$$

$$\left.\begin{array}{l} \dfrac{\partial L}{\partial v} = X^\top Y w - \lambda_v X^\top X v \qquad \dfrac{\partial L}{\partial w} = Y^\top X v - \lambda_w Y^\top Y w \\ v^\top \dfrac{\partial L}{\partial v} - w^\top \dfrac{\partial L}{\partial w} = \lambda_w w Y^\top Y w - \lambda_v v X^\top X v = \lambda_w - \lambda_v \end{array}\right\} \overset{!}{=} 0 \quad \text{for extremum}$$

(program is smooth, so no boundary cases)

computation implies: $v^\top X^\top Y w = \lambda_v = \lambda_w =: \lambda$, and maximizer v, w must satisfy

$$(X^\top X)^{-1} X^\top Y (Y^\top Y)^{-1} Y^\top X \cdot v = \lambda^2 v$$
$$(Y^\top Y)^{-1} Y^\top X (X^\top X)^{-1} X^\top Y \cdot w = \lambda^2 w$$

generalized eigenvalue problem

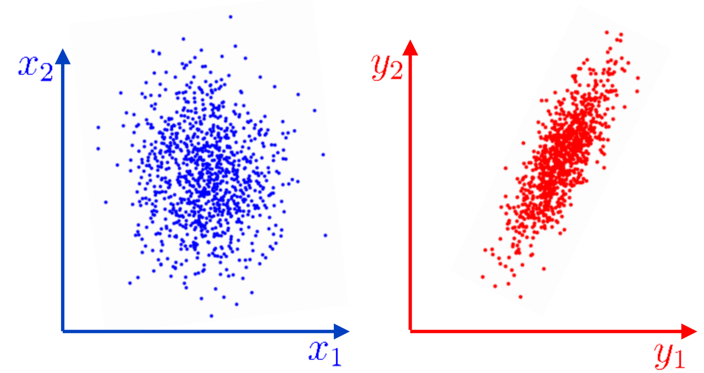
can be efficiently solved

Canonical Correlation Analysis

Input: data points $(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^n \times \mathbb{R}^m$

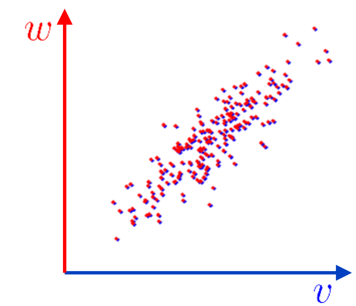
Output: coordinates $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$

maximizing correlation between $\langle v, x_i \rangle$ and $\langle w, y_i \rangle$



2D projections of the 4D data

features



good: correlated

Generalized eigenvalue problem

$$(X^\top X)^{-1} X^\top Y (Y^\top Y)^{-1} Y^\top X \cdot v = \lambda^2 v$$
$$(Y^\top Y)^{-1} Y^\top X (X^\top X)^{-1} X^\top Y \cdot w = \lambda^2 w$$

Observe: maximizer v, w must satisfy: $v \in$ rowspan X, $w \in$ rowspan Y

writing $a = Xv$ and $b = Yw$, one obtains:

$$X(X^\top X)^{-1} X^\top Y (Y^\top Y)^{-1} Y^\top \cdot a = \lambda^2 a \quad = \mathcal{P}_X \mathcal{P}_Y \cdot a$$
$$Y(Y^\top Y)^{-1} Y^\top X (X^\top X)^{-1} X^\top \cdot b = \lambda^2 b \quad = \mathcal{P}_Y \mathcal{P}_X \cdot b$$

where $\mathcal{P}_A$ denotes projection on colspan A (not rowspan A !)

so a, b, λ^2 are *leading* left and right singular vector and value to $\mathcal{P}_X \mathcal{P}_Y$

other left/right singular vectors: ***"canonical components"***

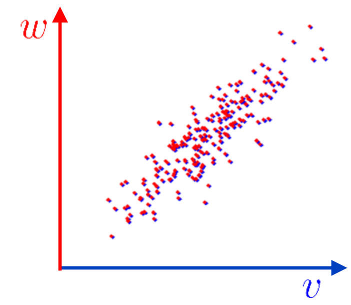
Kernelization: from properties of the pseudo-inverse (see lecture 4):

$$\mathcal{P}_X = XX^\top (XX^\top XX^\top)^+ XX^\top$$
$$= K_{XX} K_{XX}^{-2} K_{XX} \quad = I \text{ for Gauss kernel}$$

… does not work since assumption $a = Xv, b = Yw$ does not kernelize well

Kernel Canonical Correlation Analysis

(Akaho, 2001)
(Fyfe, Lai, 2001)



Input: data points $(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^n \times \mathbb{R}^m$

Output: coordinates $v \in \mathcal{F}$ and $w \in \mathcal{F}$

maximizing correlation between $\langle v, \phi(x_i) \rangle$ and $\langle w, \phi(y_i) \rangle$

Kernelization: use that $v = X^\top \alpha,\ w = Y^\top \beta$ (representer thm)

thus $X(X^\top X)^{-1} X^\top Y (Y^\top Y)^{-1} Y^\top X X^\top \alpha = \lambda^2 X X^\top \alpha$

$Y(Y^\top Y)^{-1} Y^\top X (X^\top X)^{-1} X^\top Y Y^\top \beta = \lambda^2 Y Y^\top \beta$

from properties of pseudo-inverse: $X(X^\top X)^{-1} X^\top = X X^\top (X X^\top X X^\top)^{+} X X^\top$

yields: $$\begin{pmatrix} 0 & K_{XX} K_{YY} \\ K_{YY} K_{XX} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda^2 \begin{pmatrix} K_{XX}^2 & 0 \\ 0 & K_{YY}^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Shrinkage regularization:

replace K_{AA}^{2+} by $(K_{AA}^2 + \gamma_A I)^{-1}$ this maximizes $\dfrac{\langle u, v \rangle^2}{(\|u\|^2 + \gamma \|\alpha\|)(\|v\|^2 + \gamma \|\beta\|)}$

eigenvalue problem:

$$\begin{pmatrix} 0 & K_{XX} K_{YY} \\ K_{YY} K_{XX} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda^2 \begin{pmatrix} K_{XX}^2 + \gamma_X K_{XX} & 0 \\ 0 & K_{YY}^2 + \gamma_Y K_{YY} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Temporal kernel CCA: multidimensional time series $x(t), y(t)$

(Bießmann et al, 2009) rows of Y are $y(t)$ rows of X are $[x(t), x(t+1), x(t+2), \ldots]$

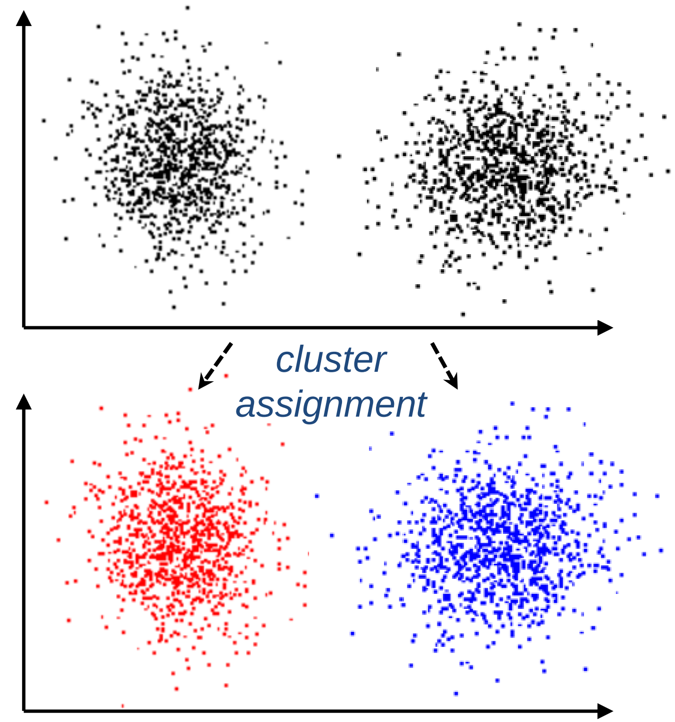
$\alpha(\tau)$ = part of α $\lambda = \lambda(\tau)$ "canonical correlogram" $\tau^* = \underset{\tau \geq 0}{\operatorname{argmax}}\, \lambda(\tau)$

Kernel k-means

K-means clustering **(Steinhaus, 1957)**

Input: data points $x_1, \ldots, x_N \in \mathbb{R}^n$ *(unlabelled)*

Output: cluster labels $y_1, \ldots, y_N \in \{c_1, \ldots, c_K\}$ (this is the "K")



Main idea: cluster label = "color" of closest cluster mean

Algorithmic idea: double iteration (EM-type)

1. cluster labels $y_1, \ldots, y_N \leftarrow$ closest cluster mean color
2. recompute cluster means $\mu(c_1), \ldots, \mu(c_K)$

(plus various initialization strategies)

Good news: converges, since every step decreases

non-negative loss $D(y) = \sum_{i=1}^{N} \|x_i - \mu(y_i)\|^2$

Bad news: in general to a local minimum

Sort-of-good news: **(Aloise, 2009)**

Doing notably better is NP-hard

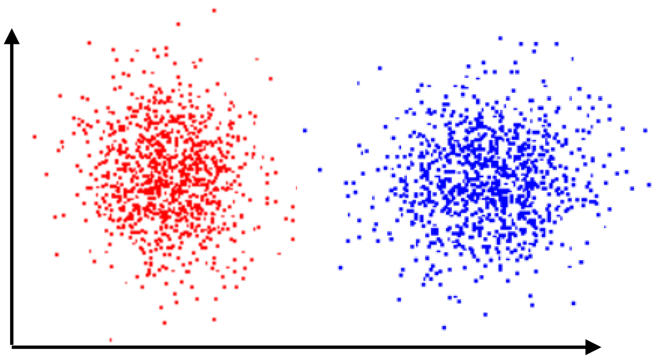
Reformulation as single-step iteration:

$$\|x - \mu(c_i)\|^2 = x^\top x - \frac{2}{\#C_i} \sum_{z' \in C_i} x^\top z' + \frac{1}{\#C_i^2} \sum_{z,z' \in C_i} z^\top z'$$ where C_i is cluster i

allows **(1.)** without explicit computation of means **(2.)**

Directly Kernelizable

Spectral relaxation (Dhillon et al, 2004)



Input: data matrix $X \in \mathbb{R}^{N\times n}$ rows = pts $x_1, \ldots, x_N$

Output: cluster labels $y_1, \ldots, y_N \in \{c_1, \ldots, c_K\}$

K-means loss: $D(y) = \sum_{i=1}^{N} \|x_i - \mu(y_i)\|^2$

if there was only one cluster:

$$D(y) = \left\| X^\top \left(I - \frac{\mathbb{1}\mathbb{1}^\top}{N} \right) \right\|_F^2 = \mathrm{Tr}(XX^\top) - \frac{\mathbb{1}^\top}{\sqrt{N}} XX^\top \frac{\mathbb{1}}{\sqrt{N}}$$ where $\mathbb{1}$ is the vector of ones

in general,write C_i for the i-th cluster, let $U \in \mathbb{R}^{K\times N}$, $U_{ij} := \begin{cases} \frac{1}{\sqrt{\#C_i}}, & \text{if } x_j \in C_i \\ 0 & \text{otherwise} \end{cases}$

$$D(y) = \|X\|_F^2 - \|ZX\|_F^2 = \mathrm{Tr}(XX^\top) - \mathrm{Tr}\left(U(XX^\top)U^\top\right) = \mathrm{Tr}(K_{XX}) - \mathrm{Tr}\left(UK_{XX}U^\top\right)$$

Observation: $U^\top U = I$ and U enters only in the second term

Relaxation: consider *all* orthogonal U, not only those of special form (as defined above)

Then $\operatorname*{argmin}_U D(U) = \operatorname*{argmax}_U \mathrm{Tr}(UK_{XX}U) =$ first K eigenvectors of K_{XX}

Relation to other clustering/unsupervised learning algorithms: replace K_{XX} by

$W^{1/2}K_{XX}W^{1/2}$ W weights
weighted spectral K-means

$D^{1/2}AD^{1/2}$ A adjacency/similarity matrix, $D = \mathrm{diag}(A\cdot\mathbb{1})$
normalized cut/spectral clustering

THE END

(of kernels)