MVA "Kernel methods"
Homework 2

Jean-Philippe Vert

Due January 25, 2012

Exercice 1.
Given two sets of real numbers $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the covariance between X and Y is defined as

$$cov_n(X, Y) = \mathbf{E}_n(XY) - \mathbf{E}_n(X)\mathbf{E}_n(Y)\,,$$

where $\mathbf{E}_n(U) = (\sum_{i=1}^n u_i)/n$. The covariance is useful to detect linear relationships between X and Y. In order to extend this measure to potential nonlinear relationships between X and Y, we consider the following criterion:

$$C_n^K(X, Y) = \max_{f,g \in \mathcal{B}_K} cov_n(f(X), g(Y))\,,$$

where K is a positive definite kernel on $\mathbb{R}$, $\mathcal{B}_K$ is the unit ball of the RKHS of K, and $f(U) = (f(u_1), \ldots, f(u_n))$ for a vector $U = (u_1, \ldots, u_n)$.
1. Express simply $C_n^K(X, Y)$ for the linear kernel $K(a, b) = ab$.
2. For a general kernel K, express $C_n^K(X, Y)$ in terms of the Gram matrices of X and Y.

Exercice 2.
In order to cluster a set of vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ into K groups, we consider the minimization of:

$$C(z, \mu) = \sum_{i=1}^n \| x_i - \mu_{z_i} \|^2$$

over the cluster assignment variable z_i (taking values in $1, \ldots, K$ for all $i = 1, \ldots, n$) and over the cluster means $\mu_i \in \mathbb{R}^p, i = 1, \ldots, K$.

1. Starting from an initial assignment z^0, we can try to minimize $C(z, \mu)$ by iterating:

$$\mu^i = \underset{\mu}{\operatorname{argmin}}\ C(z^i, \mu)\,, \qquad z^{i+1} = \underset{z}{\operatorname{argmin}}\ C(z, \mu^i)\,.$$

Explicit how both minimization can be carried out (note: this method is called k-means).

2. Propose a similar iterative algorithm to perform k-means in the RKHS $\mathcal{H}$ of a p.d. kernel K over $\mathbb{R}^p$, i.e., to minimize:

$$C_K(z, \mu) = \sum_{i=1}^{n} \| \Phi(x_i) - \mu_{z_i} \|^2\,,$$

where $\Phi : \mathbb{R}^p \to \mathcal{H}$ satisfies $\Phi(x)^\top \Phi(x') = K(x, x')$.

3. Let Z be the $n \times K$ assignment matrix with values $Z_{ij} = 1$ if x_i is assigned to cluster j, 0 otherwise. Let $N_j = \sum_{i=1}^n Z_{ij}$ be the number of points assigned to cluster j, and L be the $K \times K$ diagonal matrix with entries $L_{ii} = 1/N_i$. Show that minimizing $C_K(z, \mu)$ is equivalent to maximizing over the assignment matrix Z the trace of $L^{1/2} Z^\top K Z L^{1/2}$.

4. Let $H = Z L^{1/2}$. What can we say about $H^\top H$? Do you see a connection between kernel k-means and kernel PCA? Propose an algorithm to estimate Z from the solution of kernel PCA.

5. Implement the two variants of kernel k-means (Questions **2** and **4**). Test them with different kernels (linear, Gaussian) on the *Libras Movement Data Set*[1] ($n = 360, p = 90, K = 15$). Visualize the data mapped to the first two principal components for different kernels, and check how well clustering recovers the 15 classes. (note: only use the first 90 attributes for clustering, the $91st$ one is the class label).

[1] `http://archive.ics.uci.edu/ml/datasets/Libras+Movement`