Homework 3

Jean-Philippe Vert

Due March 2, 2009

1 Kernel PCA

Let $S_{train} = (x_1, \ldots, x_n)$ and $S_{test} = (x_{n+1}, \ldots, x_{n+p})$ be two sets of points in a space endowed with a positive definite kernel K. Propose an algorithm to project the set S_{test} onto the first principal directions obtained by kernel PCA on the set S_{train}

2 Kernel ridge regression

Given a training set $(x_i, y_i)_{i=1,\ldots,n}$ where x_i is a point of a space endowed with a positive definite kernel K and y_i is a real number, the *kernel ridge regression with offset* algorithm solves:

$$\min_{f \in H_K, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) + b - y_i\right)^2 + \lambda ||f||_{H_K}^2 ,$$

where H_K is the RKHS of the kernel K. Propose an algorithm to find f and b.

3 Application

Download the data `xtrain.txt, xtest.txt, ytrain.txt, ytest.txt` from the course homepage. For a few kernels (e.g., linear and Gaussian with different bandwith):

- Visualize the training and testing sets by projecting them on the first 2 kernel principal directions (plot them on the same picture with a different sign or symbol for the training and test points).

- Train a kernel ridge regression model with offset on the training set, and compute the performance of the model on the training set and on the test set (using mean square error, MSE). Plot the training MSE and the testing MSE as a function of λ. Explain the shape of the curves you obtain.