Machine Intelligence II **Exercise Sheet 3**
SS2012, Prof. Obermayer due: 10.05.2010

Principal Component Analysis (PCA)

This exercise requires to compute the principal components (PCs) for different datasets. The necessary data and image files are contained in `datafiles.zip` as provided on ISIS.

3.1 PCA: Toy Data (3 points)

(a) Load the dataset `toypca/pca_data.dat` and make a *scatter plot* of the centered data.

(b) Calculate the PCs and make another scatter plot of the same data with the PCs as coordinates.

(c) Plot the reconstruction of the data in the original coordinate system. Make two additional plots using only the first and the second PC for reconstruction, respectively.

3.2 PCA: Image Data (4 points)

The directory `imgpca` contains two categories of training images: *nature* (prefix `n`) and *buildings* (prefix `b`). For both categories do the following:

(a) Sample a total of at least N=5000 patches (e.g. 500 per image) of 8x8 pixels from this set of images and assemble them in a big Nx64 matrix.

(b) Calculate the PCs of these image patches and show the first 24 as 8x8 image patches.

Question: Are there differences between the PCs for buildings vs. nature? How many PCs should you keep for each of the image groups?

3.3 Kernel PCA: Toy Data (3 points)

(a) Create a toy dataset of 2-dimensional data points $\boldsymbol{x}^{(\alpha)} = (x_1^{(\alpha)}, x_2^{(\alpha)})$, $\alpha = 1, \dots 90$. The points represent iid samples of 30 points from 3 different distributions with uncorrelated and normally distributed (sd=0.1) coordinate values. The first sample should be centered on $\langle \boldsymbol{x}^{(\alpha)} \rangle_1 = (-0.5, -0.2)$, the second on $\langle \boldsymbol{x}^{(\alpha)} \rangle_2 = (0, 0.6)$, and the third $\langle \boldsymbol{x}^{(\alpha)} \rangle_3 = (0.5, 0)$.

(b) Do a KPCA using the RBF kernel (see below) and calculate the coefficients for the representation of the eigenvectors (PCs) in the space spanned by the transformed data points.

(c) Visualize the first 8 PCs in the 2-dimensional input space in the following way: Use equally spaced "test" gridpoints and determine their PC values by projecting onto the first 8 eigenvectors in feature space. For example, plot contour lines indicating points that yield the same projection onto the PCs. You may also use a heat-map or pseudo-color plot (e.g. `pcolor` in Matlab) to distinguish the different regions. How do you interpret the results?

$$\text{RBF kernel:} \qquad k(x^{(\alpha)}, x^{(\beta)}) = \exp\left(-\frac{\|x^{(\alpha)} - x^{(\beta)}\|^2}{2\sigma^2}\right)$$

Total points: 10