Machine Intelligence 2 **Exercise Sheet 1**
SS 2018, Obermayer/Augustin/Kashef due: 2018-05-02 at 23:55

Principal Component Analysis

Exercise 1.1: PCA: 2-dimensional Toy Data (2 points)

(a) Load the dataset `pca-data-2d.dat` and make a scatter plot of the *centered* data.

(b) Determine the Principal Component Directions (PCs) and make another scatter plot of the same data points in the coordinate system spanned by the 2 PCs.

(c) PCA can be used to compress data e.g. using only information contained in the first n out of N PCs. Plot the reconstruction of the data in the original coordinate system when using (i) only the first or (ii) only the second PC for reconstruction.

Exercise 1.2: PCA: 3-dimensional Toy Data (2 points)

(a) Load the dataset `pca-data-3d.txt`, center it, and show the scatter plot matrix.

(b) Determine the PCs and make the analogous scatter plot matrix for the 2d-coordinate systems spanned by the different pairs of PCs.

(c) Examine the 3d-reconstruction of the data in the original coordinate systems when using only (i) the first, (ii) the first two or (iii) all three PCs for reconstruction. Discuss how useful these directions (i.e., the PCs) are.

Exercise 1.3: Projections of a dynamical system (3 points)

Using the data from the file `expDat.txt`, we can interpret the data as describing the process of a "large" system with d=20 dimensions at each timepoint (cf. Exercise Sheet 1).

(a) Find the 20 Principal Components (PCs) of this dataset

(b) Plot the temporal evolution of the system projected onto the *first two PCs* by making (i) a scatter plot of the 100 datapoints in the 2d-coordinate system spanned by the first two PCs and (ii) a line plot of the 100 data point projections onto the first PC and onto the second PC (i.e., two lines in one plot where the x-axis shows the time index). Use color to indicate the time index in the scatter plot and use that color code also in the line plots to highlight the (temporal) relationship to the scatter plots.

(c) Create a new dataset by shuffling the data (i.e. reorder for *each* of the 20 columns the 100 data point components in a different random sequence).

(d) Plot the covariance matrices and scree plots for both the original and the scrambled data and interpret your results.

(e) What would be the result if shuffling the data points in the *same* sequence for all columns (that is randomizing the row order)? (To answer this question no programming is required.)

Exercise 1.4: Image data compression and reconstruction (3 points)

The file `imgpca.zip` contains different categories of training images. For pictures from *each* of the categories nature (prefix `n`) and buildings (prefix `b`) do the following:

(a) Sample (randomly) a total of at least N=5000 patches (e.g. 500 per image) of 16x16 pixels from this set of images and assemble them in a big Nx256 matrix.

(b) Calculate the PCs of these image patches and show the first 24 as 16x16 image patches. Are there differences between the PCs for buildings vs. nature?

(c) Answer using a scree plot: how many PCs should you keep for each of the two image groups? What are the resulting respective compression ratios?

(d) Reconstruct 3 arbitrary images by projecting all of its constituent (non-overlapping[1]) 16x16-patches onto the first n PCs of that image category for $n \in \{1, 2, 4, 8, 16, 100\}$.
→ Plot the projected image in original space.
→ Try both: the matching PCs (i.e. building PCs for building image reconstruction, nature PCs for nature image reconstruction) and also the inverse relation (building PCs for nature image reconstruction, nature PCs for building image reconstruction).
→ Choose a compact way of visualization, e.g., using a grid of subplots, where in the horizontal direction n and in the vertical direction the image is varied along the subplots.
→ Interpret the results.

Total 10 points.

[1] For boundary regions it might be required to overlap the patches such that no pixels of the image are neglected.