Learning from Data 1, Tutorial Sheet for week 3

School of Informatics, University of Edinburgh

Instructor: Dr David Barber

1. Consider the following 3-dimensional datapoints:

$(1.3, 1.6, 2.8)(4.3, -1.4, 5.8)(-0.6, 3.7, 0.7)(-0.4, 3.2, 5.8)(3.3, -0.4, 4.3)(-0.4, 3.1, 0.9)$

Perform Principal Components Analysis by:

- Calculating the mean, $\boldsymbol{m}$ of the data.
- Calculating the covariance matrix $S = \frac{1}{5}\sum_{\mu=1,\ldots,6}(\boldsymbol{x}^\mu - \boldsymbol{m})(\boldsymbol{x}^\mu - \boldsymbol{m})^T$ of the data.
- Finding the eigenvalues and eigenvectors $\boldsymbol{e}^i$ of the covariance matrix. (This is easy to do in MATLAB using the `eig` command.)

You should find that only two eigenvalues are large, and therefore that the data can be well represented using two components only. Let $\boldsymbol{e}^1$ and $\boldsymbol{e}^2$ be the two eigenvectors with largest eigenvalues.

- Calculate the two dimensional representation of each datapoint $(\boldsymbol{e}^1 \cdot (\boldsymbol{x}^\mu - \boldsymbol{m}), \boldsymbol{e}^2 \cdot (\boldsymbol{x}^\mu - \boldsymbol{m})), \mu = 1, \ldots, 6$.
- Calculate the reconstruction of each datapoint $\boldsymbol{m} + (\boldsymbol{e}^1 \cdot (\boldsymbol{x}^\mu - \boldsymbol{m}))\boldsymbol{e}^1 + (\boldsymbol{e}^2 \cdot (\boldsymbol{x}^\mu - \boldsymbol{m}))\boldsymbol{e}^2$, $\mu = 1, \ldots, 6$.

2. Consider a set of N-dimensional data $\boldsymbol{x}^\mu, \mu = 1, \ldots, P$. PCA is performed on this dataset and the eigenvectors $\boldsymbol{e}^1, \ldots, \boldsymbol{e}^M$ are used to represent the data, together with the mean $\boldsymbol{m}$. That is, each original N dimensional datapoint $\boldsymbol{x}^\mu$ is represented in the form

$$\boldsymbol{x}^\mu \approx \boldsymbol{m} + \sum_{i=1}^{M} a_i \boldsymbol{e}^i \tag{1}$$

- What are the optimal coefficients a_i if we are to minimise the square length of the residual vector? That is, determine a_i that minimises $(\boldsymbol{x}^\mu - \boldsymbol{m} - \sum_{i=1}^M a_i \boldsymbol{e}^i)^2$. Hint: differentiate.
- Consider two vectors $\boldsymbol{x}^a$ and $\boldsymbol{x}^b$ and their corresponding PCA approximations $\boldsymbol{m} + \sum_{i=1}^M a_i \boldsymbol{e}^i$ and $\boldsymbol{m} + \sum_{i=1}^M b_i \boldsymbol{e}^i$. Approximate $(\boldsymbol{x}^a - \boldsymbol{x}^b)^2$ by using the PCA representations of the data, and show that this is equal to $(\boldsymbol{a} - \boldsymbol{b})^2$.
- Explain how the above result enables us to rapidly compare distances between PCA representations of datapoints, and how therefore in the Eigenfaces experiment, we can perform nearest neighbour classification rapidly, once the largest eigenvectors have been calculated.