Artificial neural networks

Assigment 4: PCA and SOM

Lood, Cédric
Master of Bioinformatics

June 19, 2016

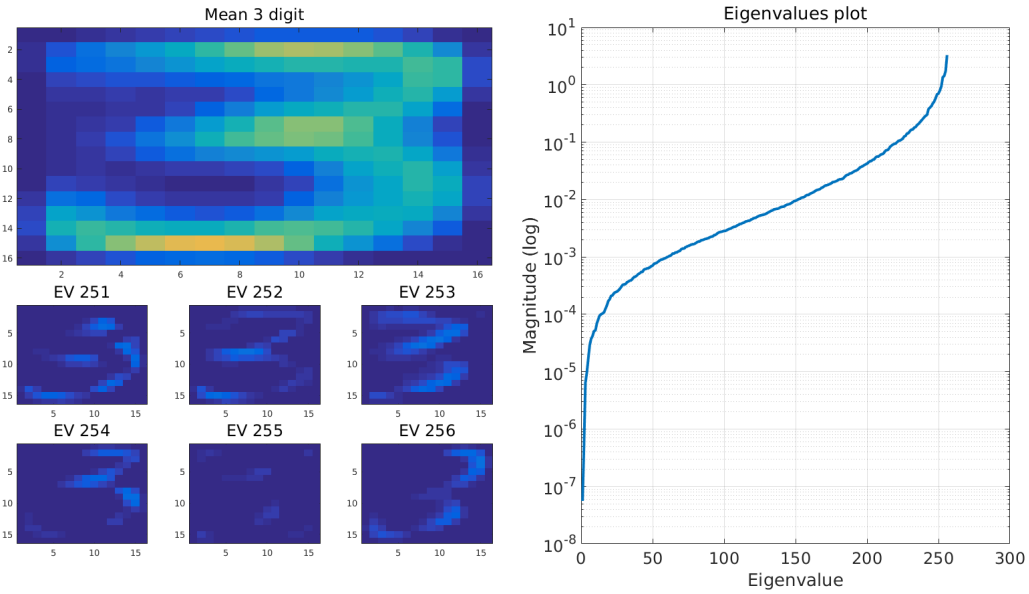
1 Context

In this exercise, I explored techniques of unsupervised learning using neural networks. Specifically Principal Component Analysis (PCA) and Self-Organizing Maps (SOM).

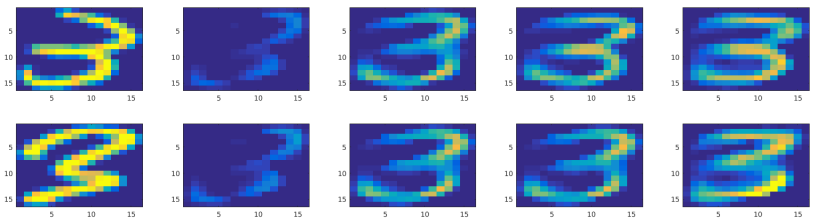
2 Principal component analysis

The idea behind PCA is to reduce the dimensionality of the data in the input space with the hope that the lower dimensional space captures most of the structure of the data. This is done by projecting the data onto the eigenvectors of the covariance matrix. Once the lower dimensionality space has been computed, another question one might try to solve is to see whether a correct reconstruction of the datapoints is possible.

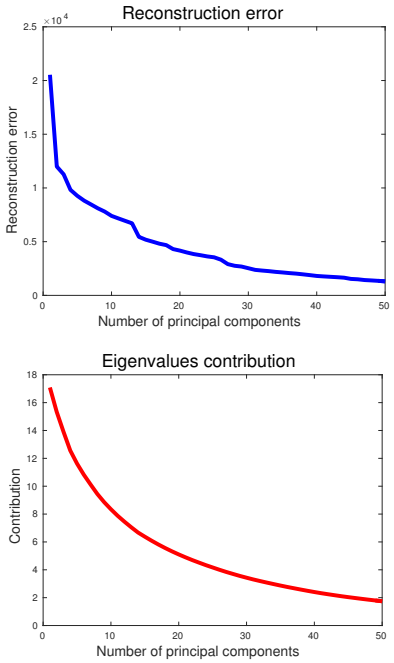
Here we will work with a subset of the US postal service database that consists of digitized versions of handwritten digits (16 by 16 pixels). In particular, we'll focus on the digit "3". On the top left-hand side of the figure below, we view the mean "3" over the 500 examples we have. The right-hand side plot shows the magnitude of the eigenvalues. The bottom left-hand side shows the plots of the eigenvectors associated with the highest eigenvalues. As can be seen on the "EV 256" plot, the curvy feature typical of a 3 seems to be associated with the highest eigenvalue.



In the next figure, you can observe the reconstruction process for 2 digits from the dataset. The rows start with the digit as it exists in the dataset, then the next are a reconstruction of the digit based on the first 4 principal components. As can be observed, the reconstructed digits are much more similar than the two original digits.



The graphic on the right-hand side reflects the decrease in the reconstruction error as the number of principal components is increased. One would expect that the error keeps decreasing, eventually reaching a value of 0, but after trying it out, I couldn't reach exactly 0. Visual inspection of the matrices containing the original dataset and the reconstructed one offered insights as to why that was. In the original dataset, many of the values are equal exactly to 0, which is not the case in the reconstructed dataset where the same positions are close to 0 but not exactly 0. I think this is due to numerical reasons. A comparison of the cumulative eigenvalues contributions and reconstruction error shows a very convincing correlation as can be observed on the bottom graph.



3 SOM

With self-organizing maps, the idea is to allow for insightful visualization in 2D or 3D of potentially very high dimensional prototype vectors (such as those created using vector quantization).

As can be seen on figure 2, a fixed topology for the prototype vectors is given before the training starts. The dataset is illustrated in the green points and consists of a hollow cylinder. After training, one can observe that the prototypes have moved closer to the data and captured the essence of the distribution. The grid topology seems to be more sparse than the other two alternatives.

On the right-hand side of the figure 2 you can see the results of the SOM in terms of the prototype vectors (red dots), each located within a cluster. Using the Rand Index for different trainings to verify the correctness of the simulated classification using the SOM vs the true labels, we obtain a maximum RI of 86% and an adjusted RI of about 72%. That we do not reach 100% correctness is expected as some of the clusters share considerable overlap.

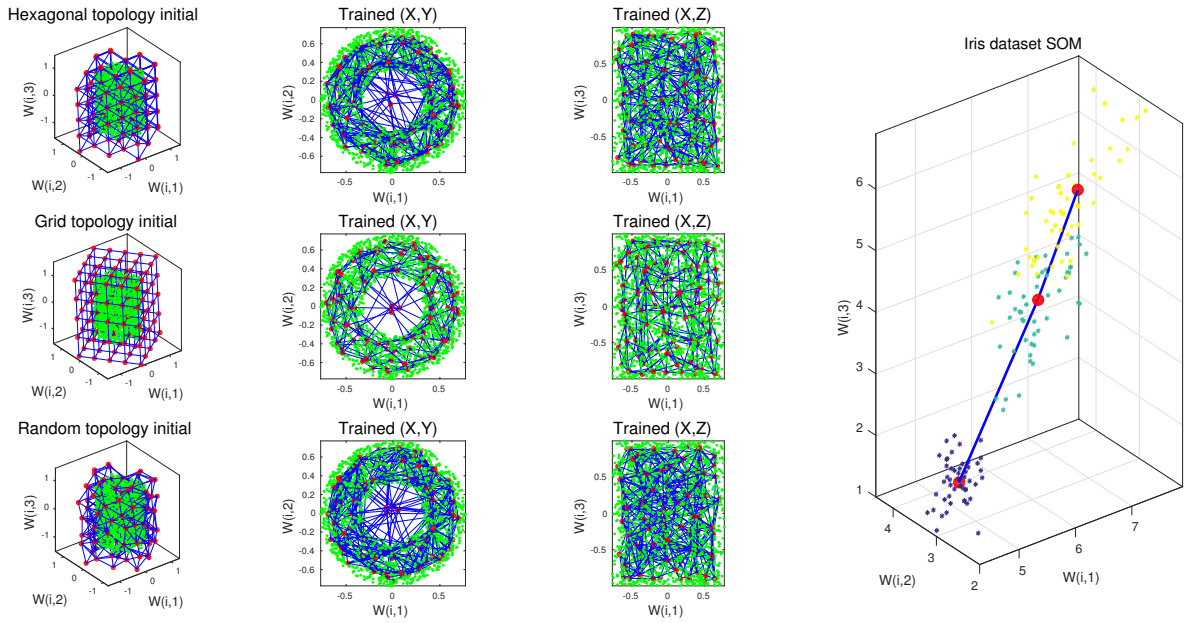


Figure 2: SOM