Barbara Hammer, Philip Kenneweg, Bianca Schröder Universität Bielefeld

Maschinelles Lernen im Web (WiSe 24/25)
2. Sheet

Start: Monday, 28.10.2024.
End: The worksheets should be solved using Python, in groups of 2-3 people and will be presented in the Tutorials.
Discussion: Monday or Thursday, 11.11.2024 in the Tutorials.

Information

The worksheets will be made available in the Lernraum "392180 Maschinelles Lernen im Web (V)". Worksheets will usually be released every two weeks on Monday, and discussed during the tutorials two weeks later. In order to successfully finish the course, 50% of the available points have to be obtained by presenting the results in the tutorials.
The week in between the release and discussion of the sheet will be used to discuss the theoretical exercises, the implementation of various algorithms as presented in the lecture, as well as go deeper into the relevant material.

Exercise 1: Theory

(*0 Points*)

Which of the following statements is true? This will not be graded but is similar to the final exam questions.

(a) The eigenvalues of XX^t and X^tX are both the principal components of the data X.

On the other hand, while the eigenvectors of XX^t are the principal components, the one from X^tX are not, but they need to be linearly transformed using X.

(b) For whitened data, ICA and PCA yield the same results.

it targets dimensions which minimize Gaussianity.

(c) Given data from different classes, where each class can be represented as mixture of Gaussian. Then Fisher discriminant analysis aims for a linear projection which separates these classes best.

(d) Parametric dimensionality reduction methods provide direct out-of-sample mappings; this holds for ICA, PCA, UMAP.

(e) Isomap relies on MDS and thus offers closed form solutions via eigenvectors.

Afterwards, it uses MDS, where it could use metric one, or mode general solved e.g. by quasi-Newton.

(f) LLE has a quadratic effort for data described by pairwise distances.

(g) The following dimensionality reduction methods do not provide deterministic outputs: tSNE, UMAP, Laplacian Eigenmap.

outputs in different runs.

(h) The following methods can be used to project also larger data sets, since their complexity is not much affected by the number of data points (eg less than quadratically): Barnes hut t-SNE, ICA, MDS.

(i) The quality framework evaluates the neighborhood preservation of DR methods in a quantitative way and in linear time.

(j) The following methods take label information into account and thus can provide an informed projection: LDA, Fisher-t-SNE.

Exercise 2: Practical

(*10 Points*)

You can use all sources from the internet, but you must add a reference. Please solve and prepare a short presentation of the methods and results. Each group needs to prepare solutions for all of the tasks and needs to present them in the tutorial.

(a) *(5 Pts.)* Signal decomposition: Generate three time series of the form

$$f_1(t) = \sin(t) + 0.001\eta \tag{1}$$

$$f_2(t) = 2(t - \lfloor t \rfloor) + 0.002\eta \tag{2}$$

$$f_3(t) = 0.01t + 0.001\eta \tag{3}$$

where η is independent Gaussian noise per entry and t is the time. Generate a random 3×3 matrix A and compute the transformed values $\vec{x}(t) = A \cdot \vec{f}(t)$. Investigate the output of PCA and ICA on these data. Does PCA / ICA reconstruct the signals? Use a quantitative evaluation! Evaluate the sensitivity with respect to the noise.

(b) *(5 Pts.)* Take three high dimensional data sets, at least two of those real such as COIL or FashionMNIST with at least ten dimensions e.g. from scikit [1] or UCI KDD [2]. Compare the result of three different dimensionality reduction methods - inspect this visually but also compare a numeric evaluation.

[1] https://scikit-learn.org/stable/datasets/real_world.html

[2] https://kdd.ics.uci.edu/summary.data.type.html