Exercise on kernel-PCA

November 19, 2021

From the link `https://github.com/alexdepremia/Unsupervised-Learning-Datasets.git` download the files data_kPCA.txt and labels_kPCA.txt. The first file contains 10 columns with the variables (real numbers) describing the data, while the second one contains a single column with a class (integer number) associated with these variables.

1. Preprocess the data in the first file by centering the variables and divide them by their standard deviation.
2. Divide your data set into two. One (with the first 1000 data points and labels) would be employed as test set, while the other (the last 9000 data points) would be employed as learning set.
3. Use the program from the previous exercise for computing the Principal Components Analysis on the learning data set. Obtain and plot the eigenvalue spectrum using logscale for the y-axis. Project the data in the two first PCs and color it by label.
4. For an increasing number of principal components (from 1 to 10):
   (a) Apply a multinomial logistic regression (you can program it yourself or use an external library) for learning a model in the learning data set.
   (b) Transform the coordinates of the test data set with the matrix learned from the learning data set and make a prediction based on the logistic model. Quantify the quality of the prediction by computing the Mutual Information between the ground truth classification and the predicted labels.
5. Repeat points 3. and 4. but using the kernel-PCA with a Gaussian kernel. Use as width, for instance, the average distance from the fifth nearest neighbor for each data point.
6. [Optional] Implement ISOMAP, project the data set into 2D and color by label.
7. [Optional] Repeat the exercise with the Anuran data set from the first exercise.

Notes & clues

1. On `https://web.stanford.edu/~hastie/Papers/glmnet.pdf` you can find details on Multinomial Logistic Regression and how to implement it as a Maximum Likelihood problem. Otherwise, you can use, for instance, the scikit-learn implementation.

2. The Mutual information between two set of discrete labels is computed as

$$I = \sum_{y} \sum_{x} p\left(x, y\right) \log \left(\frac{p\left(x, y\right)}{p\left(x\right) p\left(y\right)} \right)$$