Machine Learning I
Practice Session I

1 Goals

The goal of this assignment is: (a) to learn how to use the Principal Component Analysis (PCA) technique presented during the class and (b) to get you acquainted with cases in which PCA can perform well and cases when it does not.

The document is divided into three parts:

- Part 1 – Getting started: Instructions on how to download and run MlDemos.
- Part 2 – How-to: Instructions on how to perform the objectives of the first practicals: importing a dataset, visualizing data, performing PCA
- Part 3 – Tasks and Questions: Set of tasks to be performed during the practical and questions you must answer.

For this first practical you will focus on loading various real-life datasets, visualizing the data distribution and then selecting good projections of the data through Principal Component Analysis. A good projection is such that it improves the separability of the data or reduces the dimensionality of the dataset, or do both. Throughout the practical session, you will be working on the following real data:

1. Wine cultivar
2. Autonomous driving
3. Fault detection of metal plates
4. Faces

You can download and find a description of the datasets on the moodle website for the class here.

1.1 Getting started

1.1.1 Using Virtual Machines

1. Login at the terminal with your GASPAR account
   If you are not already in a virtual machine:
   (a) Open the VMware horizon client and login with the GASPAR account
   (b) Select the STI Windows 10 Virtual Machine
2. Download and extract MLdemos under the Documents folder.
3. Launch mldemos.exe

1.1.2 Installation on other Windows machines

If you are not using the STI-Windows 10 virtual machine,you can download ML demos to your Windows machine, **download MLDemos**. All you need to do is to:

1. Get the software (downloadable on Moodle or at `http://lasa.epfl.ch/teaching/lectures/ML_Msc/MLDemos-Windows-Latest.zip`)
2. Unzip it in the folder of your choice[1].
3. Run the executable *mldemos.exe*

The software will provide a graphical interface for visualizing the data and algorithms you will use throughout this year.

Important: The only maintained and updated version of MLDemos is the Windows version. If your personal computer is a mac or linux, you have to use the virtual machine which can accessed remotely with any OS at `http://vdi.epfl.ch`).

2 How-to:

Numerical datasets

2.1 Import numerical dataset

The numerical datasets used for the practicals are csv files where each row represents a sample, each column an attribute and the last column is the class in which the sample belongs to. Also only the first row of the csv contains a short id tag for each attribute. Feel free to open the csv files with applications like Excel in order to familiarize yourself with the layout.

Follow these steps to import the datasets.

1. Launch MLDemos and select *File >Import > Data* from the menu and open the csv file you want to load. In this example we will use the wines' dataset.
2. The data loading interface will pop up (see Figure 1). At this dialogue box select the: "*First row as Header*" option and then click on the : "*Send data*" button. Now you should be able to see how the data are distributed at the first two dimensions where the different colors represent the classes (see Figure 2).

2.2 Visualize data

Real life datasets usually consist of many dimensions. Thus a 2D visualization may not be enough to determine if the dataset is easily classifiable and which dimensions are the most important. Thus, we can use different visualizations to interpret high-dimensional data. To achieve that change from 2D view to visualizations at the drop-down menu below the grid (see red box in Fig. 2)

We will focus on two types of visualizations:

- Individual plots which illustrate how the classes are distributed on each dimension using box-plots.
- Scatter plots which illustrate how the samples are distributed on 2D.

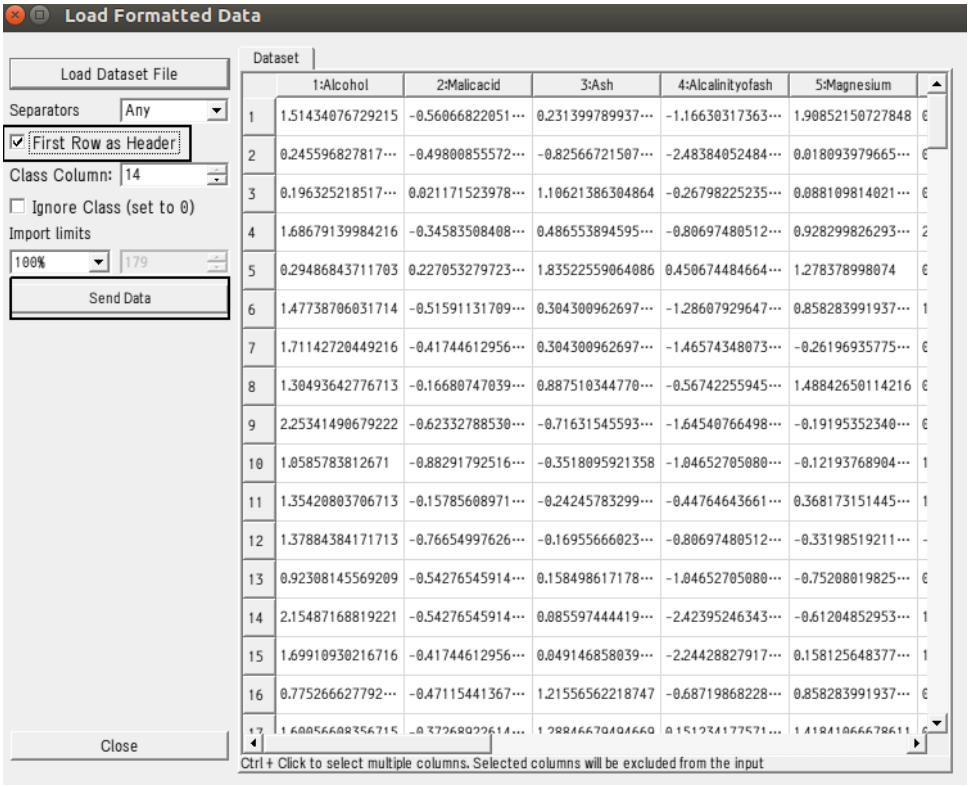

Figure 1: Data loading interface. Select the : "*First row as Header*" option and click on send data button

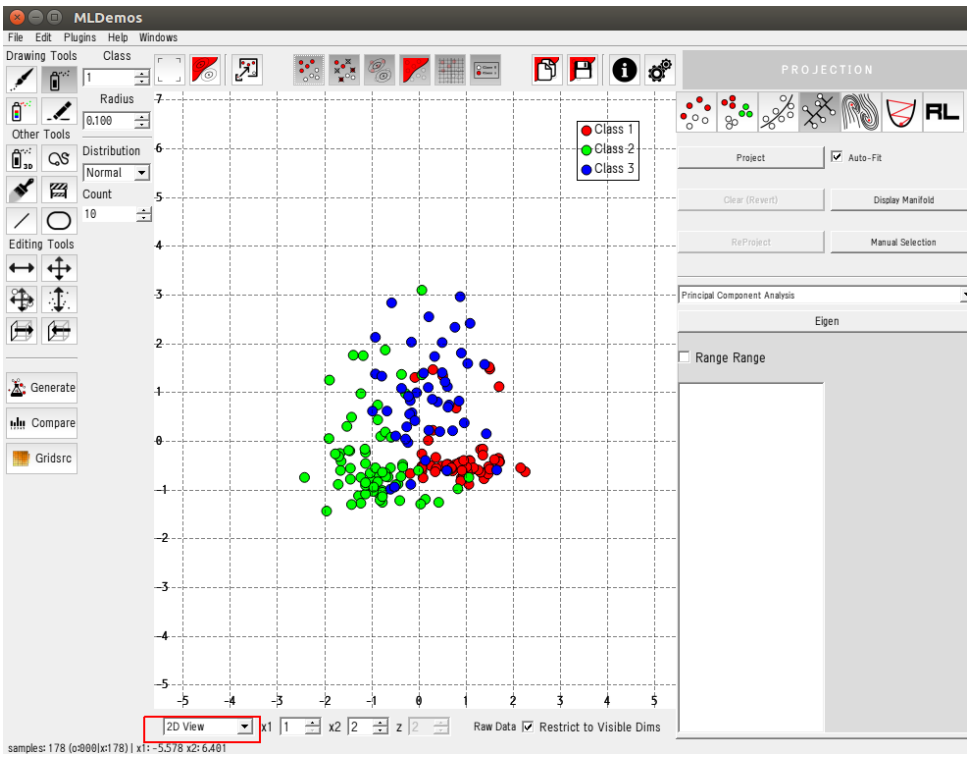

Figure 2: 2D visualization of the dataset.

(a) Individual plots

(b) Scatter plots

Figure 3: Data visualizations

You can select the type of plot from the drop-down menu at the top left (see Fig.3)

2.3 Perform PCA and dimensionality reduction

2.3.1 Performing PCA:

- Switch to 2D view (Box 1 in Fig. 4).
- From the top right toolbar select projection (Box 2 in Fig. 4)
- Select the Principal Components analysis from the drop-down menu (Box 3 in Fig. 4) and press the project button (Box 4 in Fig. 4)

This will project the original data to all the principal components. MlDemos also provides the percentage of explained variance for each principal component (Box 5 in Fig. 4) and the reconstruction error curve w.r.t. the principal components (Box 6 in Fig. 4)

Reconstruction Error Curve (REC): Projecting onto a subset of principal components results in an information loss which can be measured through the reconstruction error. The more components, the smallest the reconstruction error. The reconstruction error curve (REC) illustrates the rate at which the reconstruction error decreases as one includes more principal components. This creates the curve you can see in box 6 in Fig. 4.

[1]It is advised to decompress the MLdemos zip file in the desktop folder if you are using an EPFL computer to avoid folder/files path issues.

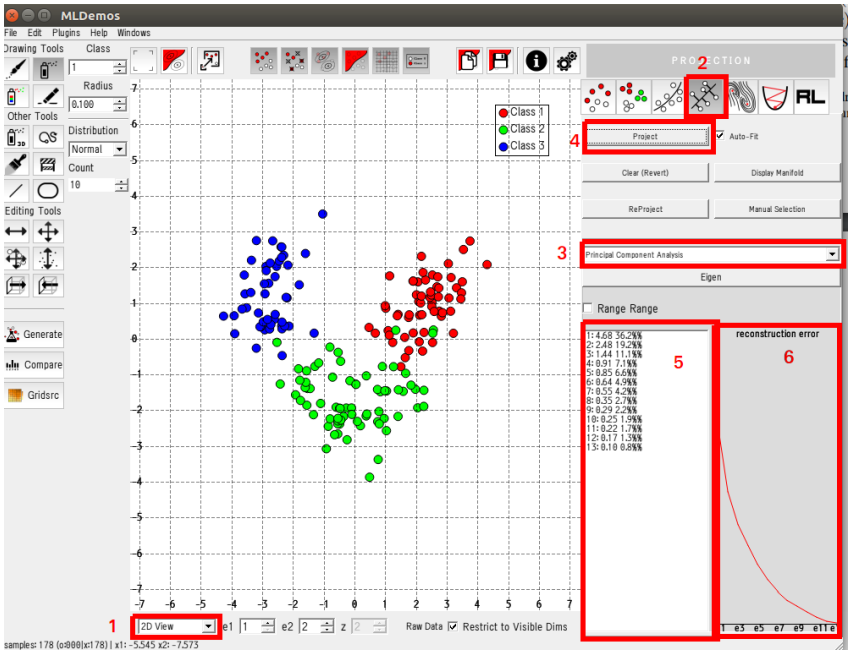

Figure 4: PCA interface.

2.3.2 Performing dimensionality reduction:

So far the original datataset is projected to its principal components which are of the same amount as the original dimensions. In order to reduce the dimensions of the data-set, you must select which projections will be kept. In order to do that:

- Tick the range button and select the range of projections which will be kept (Box 1 in Fig. 5)
- Press the ReProject button (Box 2 in Fig. 5).

Now only the projections which were specified by the range are kept and the others are discarded.

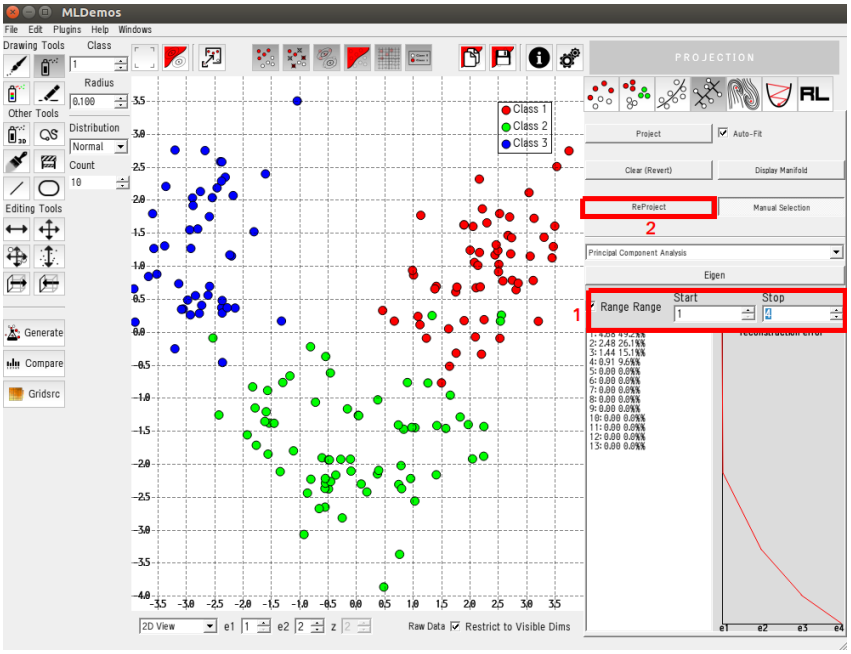

Figure 5: Dimensionality reduction.

Image Datasets

2.4 Importing and applying PCA on Image datasets

1. Select *Plugins* > *Input/Output* > *PCA faces*
2. From the *PCA faces* click the *Load Dataset* button (Box 1 in Fig. 6) and select the *faces.png* file

   This will load a set of images. The images have class labels, illustrated with a color box around the images. You can visualize the eigenvectors of the data by selecting the Eigenvectors button (Box 2 in Fig. 6).

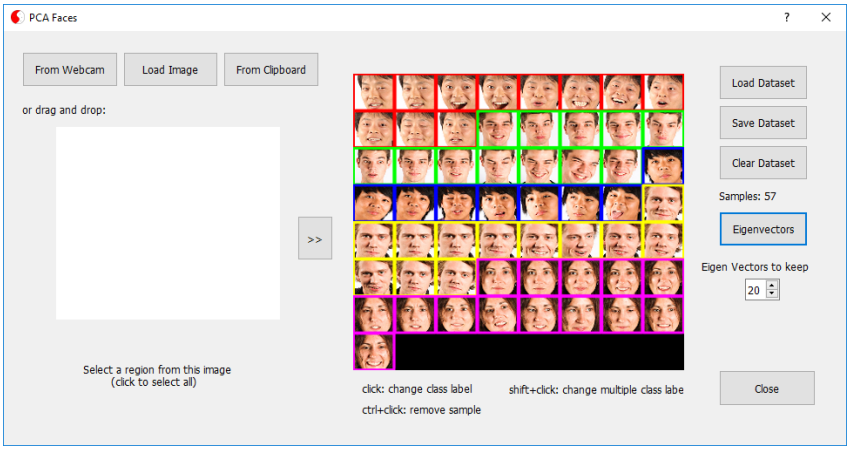

Figure 6: The PCA faces GUI.

Alongside with the visualization of the eigenvectors, you can get the reconstruction error curve and the percentage of explained variance per projection (Fig. 7). Based on those you can

select how many eigenvectors to keep (box 3 in Fig. 6). Just select the number and close the GUI.

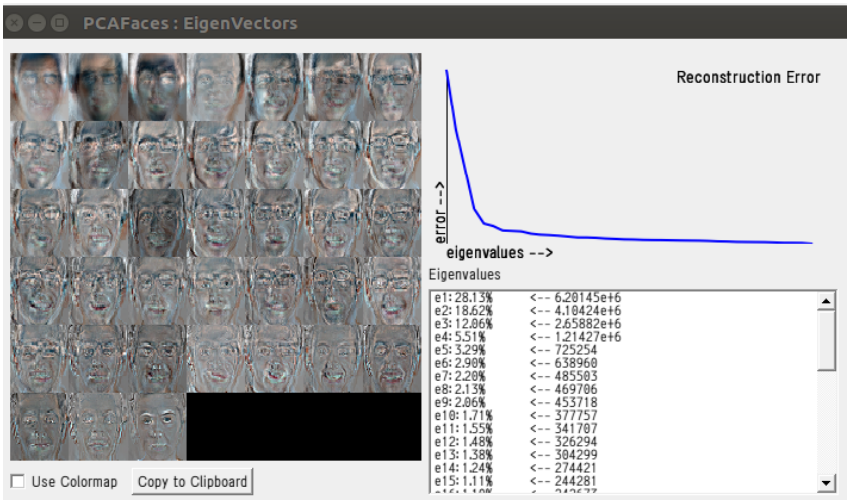

Figure 7: Visualization of eigenvectors as images (eigenfaces).

3 Questions

Make sure to answer in details to each of these questions.

Q1: Import the faces dataset and perform dimensionality reduction. How many and which projections are needed to linearly separate all the classes of the dataset? Which faces (samples) are projected at the highest values on the coordinates of the first three eigenvectors and which are canceled out (projected close to zero on the coordinates of the first three eigenvectors)? Examine the illustration of the eigenvectors (eigenfaces) and explain how they are related.

Q2: Import the *Fault detection of steel plates* dataset and visualize the data using individual plots. Which dimensions among the original dimensions are the best to separate the data? Why?

Q3: Apply PCA on the *Fault detection of steel plates* dataset and visualize the data using individual plots and 2D scatter plots of the first projections. Do not perform dimensionality reduction. Which projections appear to separate best the classes? Compare the number and quality (in terms of classes' separability) of the best principal projections and best original dimensions. What do you observe?

Q4: Import the *Autonomous driving* dataset and perform PCA. Select which components to keep based on the explained variance per component and the reconstruction error curve. Is the decision similar for both criteria? Why?

Q5: Select the principal components which allow to separate the classes. Help yourself with the visualization of individual plots. Are those the same as in Q4? Why?

Q6: Compare the class separability between the projections on the first three eigenvectors and the projections on the last three eigenvectors. What do you observe and why.