Applied Machine Learning
Practice Assignment

Professor: Aude Billard
Teaching Assistants: Laura Cohen, Denys Lamotte, Murali Karnam
contacts:
aude.billard@epfl.ch
laura.cohen@epfl.ch
denys.lamotte@epfl.ch
murali.karnam@epfl.ch

Winter Semester 2016

1 Introduction

In the practice part of the applied machine learning course, you will have the opportunity to apply each of the algorithms seen in the class on real datasets. At the end of the course, you will have to either give a written report or an oral presentation on your work. This document provides instructions to help you to prepare the material for your report or oral presentation.

1.1 Work to be done during the practice sessions

During the practice sessions, you must achieve the following:

- Build two datasets of your choice (dataset-1 and dataset-2). Dataset-1 will be used in the practice sessions on PCA, clustering and classification. Dataset-2 will be used for the practice session on regression.

  Choosing well the dataset is crucial to obtain good performances from machine learning algorithms. In the first practice session, the assistant will show you how to generate a dataset composed of images.

  Dataset-1: Generate a first dataset of minimum 50 images. The images must be composed of two main classes: for instance you could pick images of human faces and images that contain no human faces, and of several subclasses, e.g. people wearing glasses, people smiling or not, people with dark versus light hair, etc. **Be imaginative:** Generate a dataset that has your own choice of classes and subclasses. It does not have to be faces of humans.

  Dataset-2: Pick a subclass of dataset 1 and, for each image in this subclass, give a grade on a scale from 0 to 1 to the quality of the image. The quality of the image is measured according to a metric *you must design*. The metric should measure a feature of the subclass (e.g. if you have images of people with different hair color, you could grade how dark the hair is for the person shown in the image. You would then have a grade for each image which denotes the level of darkness of the hair.). To have enough statistics, increase the size of your dataset 2 by recording more images of this subclass so as to reach a total of 50

images for this subclass. Number your images and attach a grade to each image number. Store this as you will need this for practice session 3.

- Apply PCA on your dataset and choose the projections that allow you to best separate the classes and subclasses. Expect to have to use several projections in combination to separate the classes.
- Cluster your data using K-NN and K-means algorithms. Assess the quality of the clustering in both unsupervised and semi-supervised format.
- Classify your data using GMM + Bayes and SVM seen in class. Compare the performance of the classifiers and the sensitivity to choice of hyperparameters.
- Apply non-linear regression techniques to estimate the metric on your images. Compare the performance of the non-linear regression technique and the sensitivity of the techniques to choice of hyperparameters.

1.1.1 Content of the Report or Oral Presentation

The report or oral presentation must report on the following:

1. The dataset: number of datapoints, dimension of data, explanation of the classes and their features.
2. The most interesting PCA projections you found on this dataset.
3. A qualitative assessment of the results from applying K-means and KNN in unsupervised mode to your dataset. A quantitative assessment of applying the same methods tested in semi-supervised clustering mode. Specifically, make sure to report i) the values tested for the hyperparameters of the clustering technique, and ii) the ratios across labelled and unlabelled points tested in semi-supervised mode.
4. Classification performance on both training and testing sets after crossvalidation. Report on values tested for i) the hyperparameters, ii) the training/testing ratios, iii) number of folds for crossvalidation. Give performance results only for the most interesting choices of hyperparameters, training/testing ratios and number of folds.
5. Do as above but for regression performance.

Points 1 to 3 require a qualitative evaluation which must be accompanied with illustrative plots. Points 3 to 5 require a quantitative evaluation which must be accompanied with plots or tables of results. In both oral and written reports, make sure to **explain** your results. Make also sure to justify the choice of values you tested for the hyperparameters, training/testing ratios and number of folds.

1.2 Grading & Submission

If you have chosen the option to submit a written report, you must hand in the report no later than **December 16th 2016, 18h00**. Reports should be submitted online at the course webpage `http://lasa.epfl.ch/teaching/lectures/ML_Msc/#submission`. The submission form is located at the bottom of the page and indicates which submission is currently open. You should select your group and upload a .pdf file not more than 10 MB in size. You may upload multiple times, in which case, only the latest file will be graded.

Delays will be penalized: 1 point will be subtracted for each day of delay. The first day late counts starting one hour after submisson deadline. Practicals are conducted in teams of two. Unless told otherwise, we assume that the work has been shared equally by the members of the team and hence all members will be given the same grade. More information on the assignment and on the way the report should be written are given below.

The report or oral presentation is worth 25% of the overall grade. Grading scheme goes as follow: 80% of the grade is based on the content and is equally divided over the 5 points above. The last 20% will grade other aspects: clarity of the figures/tables for the written report, clarity of the speech and timeliness for the oral presentation.

1.2.1 Report

Write a report of maximum 10 pages (single column, 10pt minimum) in PDF format. **Pages beyond the tenth one will be ignored and material in these pages will not be graded.** The best way to write the report is to fill it in as you go during the practical sessions. Just jotting down some quick notes and adding some pictures while you experiment will save you hours once you work on the report itself. A qualitative evaluation should contain images (e.g. screenshots) which exemplify the concepts you want to explain (e.g. an image of a good projection and an image of a bad one). Make sure to plot only a subset of all the plots you may have visualized during the practical. Choose the ones that are the most representatives. Make sure that there is no redundancy in the information conveyed by the graphs and thus that each graph presents a different concept. Each graph/image should be accompanied with a caption that explains the content of the image. Bad captions are captions that contain solely the figure number! An example of good caption would typically read as follows: *Figure 2: The left plot shows the e1 and e2 projections of 10 images of human faces, typical of those shown in Figure 1.* In the main text, refer to *all* figures using their figure numbers. **Bad captions and lack of clear references to pictures in the text will be penalized.** A latex template for the report is provided on the course's website.

1.2.2 Oral Presentation

The oral presentation will last 10 minutes for a team of two people. Each person should speak for 5 minutes. Make sure to have a reasonable amount of slides. For 10 minutes, one usually targets around 10 slides, assuming that one speaks for around 1 minute on each slide. Rehearse your presentation before to make sure that the timing is good and that your speech is well coordinated among the two team members. We will clock the presentation and we will stop you after 10 minutes even if you are not done. A template for the slides is provided on the course's website.