Data Analysis and Model Classification
Exercise rules and regulations

Ruslan Aydarkhanov, Bastien Orset, Julien Rechenmann
Ricardo Chavarriaga, José del R. Millán

September 22, 2017

The EPFL course EE-516 Data Analysis and Model classification includes mandatory exercises which count **one third** towards of the final course grade. The exercise session will be held each Monday from 5:15pm to 6:45pm in the room INF1. Attendance in the exercise sessions is not mandatory but strongly encouraged.

1 Miniprojects

The evaluation of the exercises will be done in the form of **three miniprojects**. These projects will cover the topics for supervised and unsupervised machine learning strategies and require students to

- do the statistical evaluation of a given dataset
- formalize the problem
- find a suitable solution strategy (theoretical tools)
- implement this strategy in MATLAB
- report and discuss the results in a concise way

Detailed assignment sheets for each report will be issued at the start of each miniproject. The first miniproject will count 25% towards the exercise mark, the second 50%, and the third one 25%.

1.1 Guide sheets

Every week there will be an exercise guide sheet issued. These sheets contain small questions and implementation snippets linked with the contents of the lecture of the same week. Although they are **not mandatory**, we strongly advise you to take the time and work through them. They will incrementally guide you in your miniprojects and allow you to identify topics that you haven't fully understood.

1.2 Group work

Discussing the theoretical backgrounds and the implementation (problems) with your peers is a valuable way of deepening your personal understanding of the course contents. Therefore the miniprojects are designed to be done in groups of three people. **You are requested before October 2 to send one email per group** to *damc@listes.epfl.ch*, containing the names of the **three group members**. People unable to find a three-person group themselves are requested to send us an email with their names and we will help them finding a group.

1.3 Reports

At the end of every miniproject one report per group has to be handed in. These reports are meant to be **structured** analyses of the datasets and problems at hand. This includes the evaluation of the dataset, the description of the strategy and the results and findings from the implementation.

Not all data generated by your code is valuable, and not everything valuable is generated by your code! You should learn to focus on the main points of interests of your dataset and analysis, and give a **structured, clear and concise** discussion of your results and their interpretation.

Therefore the reports have a strict page limit of

- miniproject 1: **6 pages**
- miniproject 2: **12 pages**
- miniproject 3: **6 pages**

excluding a title page, and every additional page will be ignored!

1.4 Deadlines

Group formation	October 7th
Handing in the first miniproject	October 15th (23h59)
Handing in the second miniproject	November 19th (23h59)
Handing in the third miniproject	December 4th (23h59)

2 Asking questions

The exercise sessions are designed for you to deepen your understanding of the course contents, and therefore you **will** encounter questions or feel lost. We, the teaching assistants won't leave you hanging in such a situation, **but only if we can see that you tried hard to come up with a solution by yourself!**

2.1 Personal research

Talk to the members of your exercise group or to other students in the course. Consult the internet about your question, because most probably you are not the first one to come across that problem. And if it is a question about programming in MATLAB, **use the built in help function and the internet database** as well as the large MATLAB community online.

2.2 The moodle forum

All students should be inscribed to the course moodle (http://moodle.epfl.ch). There you will find the lecture slides, the weekly guide sheets and the forum. The moodle forum is meant to be **primarily** an exchange platform for you students to discuss problems and to share answers. So please do engage in a lively discussion and share your point of view whenever you can contribute a little bit to answer a question.

Also think about the **clearest way** how to formulate your question, and **add your thoughts** about how to (hypothetically) solve the problem, and why exactly you are stuck. **Then** post your well formulated question (see examples below) in the course moodle forum in the corresponding thread (we will open threads based on the course contents).

We are monitoring the forum to steer discussions in the right direction or to give hints how to find an answer to some of the tougher questions. But **be aware** that we are **not** commenting on any thread unless the posting is already older than two working days! In the meantime you should strive to **help each other**!

2.2.1 How to post a question

Like mentioned above, please come up with a clear and understandable question (use images and graphs to make your point!) which shows what you have already tried to solve your problem, and where exactly you are stuck. Here are some examples of do's and don'ts:

Good examples

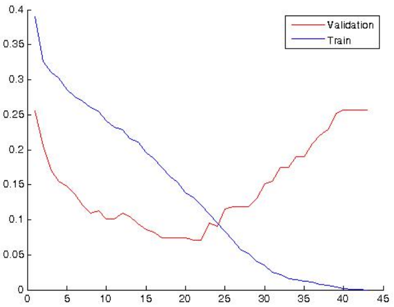
- Hello,

  I'm having some trouble with task 3 from Exercise 1. When I enter the following command (using EValX from part (c)) `bestValErrX, bestNumPCX ] = min(EValX);` BestValErrX returns EvalX and bestNumPCX returns 1...I don't understand why? EValX from part c is not the optimal (minimal) value, and we know from part e) that the associated ideal number of principal components is around 300, not 1. Is there something I understood wrong in how to execute that first command?

  Thank you in advance for your help

- In the attachment you find a plot for my validation and train errors for a quadratic model in function of the number of features used for the classifier building. Both train and validation error have a logical course. The training error keeps on decreasing towards zero whereas the validation error reaches an absolute minimum with increasing number of features. However, it actually seems quite counterintuitive that the validation error is lower than the training error for low numbers of features used as the model is trained based upon that very training set.

  I checked the whole code but couldn't find any bug. I used the following commands both for training and validation guesses and had training and validation set sizes of respectively 216 and 24 samples (ten fold cross validation on 90%-10% training-test split).The errors for both train and



validation are then calculated by comparing these guesses with the labels.

```
[guess]=classify(Validation,Training,TrainLabels,type);
[guess]=classify(Training,Training,TrainLabels,type);
```

Do you have any insight in what could be wrong here? Thanks for your answer!

Bad examples

- I am almost done with my script, however, I just need to know what to put in the last classification. Now that my number of principal components is decided I have a set called Data2 (Testing Set), and Data1, which is a separate set from which Data3 (Training set) and Data4 (Validation Set) were derived. I would like to know what do I have to compare Data2 with...I know it should be obvious but I got confused.

  Reply: You're not the only one who got confused... please state clearly what your different data sets are, and why and how you want to compare them.

- What is the difference between Fisher and Relief feature selection filters?

  Reply: This question is too general. The two methods have been discussed in the lecture, so please see the slides and try to identify the differences. It is advisable to address these general questions during the lecture hour. If still in doubt you can discuss with your classmates and TAs whether your hypothesis is correct.

- I just have a question on the classify function. The error that we get back is the one found on the validation test is that right? Or is it the training error? Thanks a lot

  Reply: Please see the MATLAB help for an answer.

2.3 Exercise hours

If you are still left unsatisfied with the hints or answers you could manage to get, please feel free to ask your question in the exercise session. Given the size of the class, please be aware that one teaching assistant only can be there for your group at most 10 minutes in total per exercise session. So **come prepared** and ask the most prominent questions first!

3 Important!

- The exercises are designed for group work and and are supposed to be doable without much help from teaching assistants.
- The exercise sessions are **not** there for walking you through the guidesheet.
- We are aware of the extremely short time for questions in the exercise sessions. Therefore you **need** to use the moodle forum to get all your questions answered!
- If you have general questions about the theory is better to ask them during the lecture hours.
- The mailing list *damc@listes.epfl.ch* is **only** there to communicate administrative questions, technical difficulties, etc. We will not answer questions related to the course content there! Please use the moodle forum for that purpose. This ensures that all students benefit from the answers.