Machine Learning I: Assessments and hints for the group work

Dr. Luisa Barbanti, Daniel Meister and Dr. Meta-Lina Spohn

HS 2023 @ HSLU Luzern

Admin

In order to obtain the credits for the course "Machine Learning I" students must deliver a group work and take part to the written exam. The written exam will take place during the exam session at the end of the semester (the exact exam date is not known yet).

Repeating Students

Repeating students must enter on the dedicated ILIAS file, whether they will do the group work, take the exam or both. This information must also be communicated to Martina Gander. Deadline is the deadline for the repetition modules ("Anmeldefrist der Repetitionsmodule").

Written exam

- It is planned to take place in Luzern (physical presence), but depending to the pandemic situation it may take place on-line.
- It lasts for 60 minutes.
- It will account for 60% of the final grade.
- Examination language is English.
- It does not imply the use of computers. Note, however, that you may have to complete some **R**-codes or to write some pseudo-code.
- It is "closed book".
- Note, however, that you are allowed to take a summary wit you.
  - The summary must be hand-written on a A4 sheet (both sides allowed).
  - NB: it must be written by yourself!
  - Summaries that are too similar/identical will be confiscated and the exam grade penalised by 0.5 points.
  - Indeed, the aim of the summary is to have you learn while producing it.
  - Simply taking a good summary from someone else... does not bring much to the learning process.

Group work

Administrative information

- It will account for 40% of the final grade.
- It represents a complete analysis of a data set where you use the machine learning methods seen in class.
- In particular, you must use:
  - a Linear Model
  - a Generalised Linear Model with family set to Poisson
  - a Generalised Linear Model with family set to Binomial

  - a Generalised Additive Model
  - a Neural Network
  - a Support Vector Machine
  - solve an optimisation problem
- It must be uploaded on ILIAS in a dedicated delivery folder.
- **the deadline to deliver the group work is on Friday two weeks (for FS) and resp. three weeks (for HS) after the semester end at 5pm (i.e. in HS23 the deadline is on January 12th 2023).**
- NB: the deadline is hard. Late submissions will be penalised by a point (i.e. maximum grade achievable is then 5).
- Email submission are not accepted. The work must be submitted to ILIAS (even in case of late submissions)
- It is done by groups of **three** students (no exceptions are being made to this point).
- The group composition (names) communicated to us by entering them in the corresponding file on ILIAS (folder "Group Work" then click on "Names group work" and enter your names. Deadline for communicating the group composition is week 1.
- The choice of the dataset must be definitive by the end of week 2.
- NB: it is not possible to have mixed groups composed by repeating and "first attempt" students.
- NB: if you fail to submit together as a group, everyone in the group will be penalised.

Data set choice

- The group work represents a complete analysis of a real data set.
- The choice of the data set to be analysed is free. Nevertheless, the data set, must fulfill the following requirements:
  - Moderate size (N = [10^3, 10^5], 10-20 predictors).
  - Please do not use data sets with more than 10^5 observations.
  - Real data.
  - Must contain both: continuous and categorical variables.
  - At least one categorical variables must have more than two levels.
  - Data must be chosen by end of week 2 at the latest.
- Note that kaggle is just one of the many sources of data sets. You may want to consider websites like
  - https://zenodo.org
  - www.opendata.swiss
  - https://opendata.swisscom.com
  - https://datasetsearch.research.google.com/
  - http://puenktlichkeit.ch/
  - https://data.world/uci/
  - www.mldata.io
  - https://www.pxweb.bfs.admin.ch/
  - https://www.ostluft.ch/
  - http://www.agrometeo.ch/de/meteorology/datas
  - https://data.stadt-zuerich.ch/
  - https://data.gov.sg/dataset, (just to mention a few examples...).
- **R** itself comes with a few hundreds data sets (type "data()" to see the list), the vast majority of **R** add-on packages also comes with data sets.
- Note also that if you found a cool data set that does not fulfill all the requirements, you may complement it with other data sets (e.g. by adding coordinates, weather data or similar things).
- Note that the data set used by the lecturers cannot be reused for your work.
- You can also use data from your employer. If you do so, please make sure that you are allowed to share this data with us and your colleagues!

Deliverables

- The core deliverable piece of work is a pdf/html file created with Rmarkdown.
- Note, however, that we want you to deliver more than just the pdf/html file
- In particular, you are expected to upload a zipped file that contains:
  - The raw data.
  - A pdf/html file that documents the data preparation, graphical analysis, modelling and comparison of the models.
  - The corresponding **R**-code (and .Rmd file).
  - The files must be structured in folders properly documented via a "ReadMe" file.
- The work must be uploaded on ILIAS as a .zip file with the following properties:
  - name your .zip file as `FamilyNameStudent1_FamilyNameStudent2...` (e.g. `Barbanti_Meister_Spohn.zip`).
- New since HS 1: you should also add the matriculation number of one of the students
  - So, the naming should look like `FamilyNameStudent1_FamilyNameStudent2...MatriculationNumber` (e.g. `Barbanti_Meister_Spohn_010923.zip`).
  - We need someone's matriculation number because we want to give you feedback in anonymous way. So, at the end of the evaluation of the GW we will upload a single file with all comments for all groups.
- Make sure that the work is uploaded by the deadline. NB: you can submit a pre-final version if you like. If you do so, make sure that you clearly name your files according to the following scheme: "Barbanti_Meister_Spohn_PreFinal.zip".
  - **note however that we will only be able to download the last version you uploaded on ILIAS.**

Rmarkdown and collaboration

- **Do compile often to avoid last minute issues with files that do not compile properly.** Every year there are students who don't hand in the required html/pdf file because they don't compiled their files often enough and then had problems at the very end of the project.
- Do find an appropriate and efficient way to collaborate with the other group members (e.g. git, github, dropbox, ...).

Analysis structure and content

- The analysis must contain an analysis of the chosen data set with all methods seen in class.
- Note, however, that the total length of the document cannot exceed 30 pages. If you are submitting an html document, check its length!
- Cross validation or other methods for model comparing must be used on one single methods (e.g. you use Cross Validation to compare 2-3 SVM models).
- Students are free to choose a measure of fit that they find more appropriate.
- In case students cannot find an appropriate measure of fit, they can use the Root Mean Squared Error (RMSE).
- **Do** comment on the results of all phases.
- Also write a short "conclusions" paragraph.

Personal contribution

- All three group members must contribute in "similar/same" way to the group work.
- Each of the three group members must take the lead on a a given model. This must be indicated in the pdf file (eg. write that Luisa Barbanti took the lead on the the Linear Model section).
- Note that this does not mean that the person will need to work on that chapter on their own. On the contrary, as the name implies, the group work is supposed to be a team work.
- As there are more than three methods to use in the group work, only three of them need to have a clear lead from one of you.
- Note also that you should fill the "team agreement" and upload it into the dedicated folder on ILIAS.

Assessment criteria for the group work

- We deliberately choose not to provide a set of assessment criteria
- Indeed, we don't want you to focus on 4-5 possible criteria and maximise them
- We rather want you to produce a nice html of pdf document of the analysis of a data set
- **Your final aim should be satisfying a potential client**
- So, act as you were producing this document for a client
- The main point is that, given a data set to analyse, there is no "best solution". There are many different ways to produce an excellent analysis
- Some analyses, may focus more on the model interpretation, another analysis may focus on visual aspects of the graphs, yet another analysis may focus on a sophisticated pdf or html document...
- In other words, there are many ways on how you could get a 6 for the group work
- Be creative and try to produce something that would satisfy a client

Hints for the group work

Compile often

- Avoid compiling your Rmd file the day before the deadline.
- If you don't compile often, the .Rmd file is very likely to not compile properly.
- Do compile often.
- You can compile every time you add a few lines of code.
- If you compile your Rmd file after adding a few lines of code and the Rmd does not compile... you know where to search for the mistake.
- If you changed very many things in your code and you get troubles in compiling, then it is much more difficult to understand what new line of code is creating the issue
- If you get compilation troubles and can't find the mistake, you can start commenting out portions your code. If the then file compiles properly... you know that the line that creates issues is among those ones that you just commented out
- Finally, note that compiling documents may take some time... especially when complex models are fitted
- In these case you can use the argument option `cache = TRUE` such that a chunk is newly evaluated only if it was modified since the last compilation. If the chunk remained unchanged, then the old results are used
- Alternatively, you can save fitted models as .RDS files (see the *saveRDS()* function) such that the model does not need to be re-fitted at each compilation. You can then load the fitted model with the "readRDS()" function.
- Google for examples to best understand the hints about caching and using .RDS files

Write a nice story

- This analysis must a story that you want to tell to a client
- There are many ways to tell a nice story
- Be creative
- make use of your strengths and interests:
  - you are interested in parameter interpretation or bootstrap confidence intervals? $\Rightarrow$ then put emphasis on this
  - you are very good at creating fancy html documents? $\Rightarrow$ than spend time on this
  - you want to deepen your knowledge of SVMs? $\Rightarrow$ than expand on that topic and add things we haven't seen in class
- Remember, nevertheless, that the overarching goal is to "produce an analysis of your data set and use all methods"

Use methods in a sensible way

- Don't forget that some methods are suited to almost all situations, others are not

- Indeed, you can fit a SVM to count data as well at binary data
- However, it makes no sense to fit a plain Linear Model to count or binary data. You should use GLMs instead (with an appropriate family)
- So, fit an LM only to a response variable which is continuous
- If all the numerical variables in your data set are count data
  - it means that you didn't read the rules on how to choose the data set
  - nevertheless, you can still transform the count data with the square-root function and fit an Linear Model (`lm(sqrt(some.counts) ~ ...)`)

Compare methods in a fair way

- The final aim of your analysis strongly depends on the story that you created
- In particular, if you tell a story where a predictive model is needed (which is often the case here) than obviously prediction is the performance to be assessed
- Note that you may want to fit models to make good predictions, but also to better "understand the data". In other words, prediction is often only a part of the job
- However, when comparing predictive performance remember to compare models in a fair way
- A very common mistake is to fit a very simple linear model, with no non-linearities, with no interactions and then to compare it to a complex non-parametric methods such as NN or SVN
- Most likely the very simple model will have a predictive performance that is well below the other methods
- So, if you want to compare LM with other techniques in a fair way use complex models that contain interactions, non-linearities and even better both of them
- Note that you may still want to fit a simple Linear Model that is easy interpret alongside more complex ones that are more suited for prediction

Log-transform "amounts"

- The vast majority of the data sets analysed for the group work contain continuous variables that are amounts (see course material if you forgot what amounts are)
- Typical examples of amounts seen in the group works are:
  - weight
  - income
  - costs
  - concentrations
- Most of the time these response variables are modelled with a Linear Model, which is a correct choice for continuous variables
- **NOTE, however, virtually always "amounts" need to be log-transformed before being analysed**
- So, do fit model like `lm(log(costs) ~ ...)` when dealing with amounts

Length of the document

- As mentioned above, brevity is important
- Indeed, you are not allowed to hand in anything longer than 30 pages
- If you are not producing a pdf, but a html file you can
  - to render the html as a pdf to see whether is exceeds 30 pages
  - or simply print on paper to see whether is exceeds 30 pages (or looking at the print preview and see how many pages this is!)
  - Note that working with a html file as an output is fully ok!
  - If you produce an html file, submit is as such together with a "check" that it does not exceed the 30 pages
  - We have absolutely no preferences between pdf and html
- So make use of simple tricks to shorten your document

  - e.g. omit messages such as those created when loading add-on packages (see the "message" chunk option)
  - if the output of a model is too long... omit it or cut it (e.g. with the `capture.output()` function)
- In general is bad practice to omit code though
- Indeed, without code... it is difficult to understand what was done...
- In html documents you can add a button to show/hide code

There are plenty of data sets

- Avoid working on the same data sets as other groups!
- There is plenty of data sets freely available, so really avoid this
- Kaggle is a popular source of data sets for this group work. Note, however, that often kaggle data sets come with existing analyses, which in turn questions your actual contribution and creativity. Using a non-kaggle data set is taken as a sign of engagement and genuine work from your side.

Complementary information and project structure

Please structure your submission by complementing it with the following information:

- a ReadMe file to describe what all files in the submitted .zip file are
- The source of the data set (e.g. a website or a company that provided the data)
- A brief "contextual" description of the project (e.g. where the data comes from and so on)

If you want, you can complement the ReadMe file with information about who did what in your group.

Groups of three people

If your group is not composed by three people as required, you should explain why. In case some students left the masters you should state who where the team members at the beginning. Teams that are not composed by three students who don't provide an explanation will be penalised.

Write conclusions

- Do write conclusions
- This implies commenting and interpreting the results of your analyses as well as writing a short paragraph with conclusions at the end of your analysis
- note that brevity is essential here
- don't write very long interpretations for any single plot, just what is needed
- keep in mind that a client will be reading your report. So, you want to be informative, but not boring

FAQs

- Do we need to use all methods in our analysis? ⇒ Yes, do use all methods seen in class (LM, GLMs, ...)
- We realised our data set is not so well-suited as we believed, can we change it even though the deadline for choosing a data set is overdue? ⇒ In principle you should avoid changing data set during the group work. No data set is perfect and it can well happen that none of the methods finds any structure in the data (welcome to real data science).
- Why is there no example of such an analysis, we struggle in understanding what is required! ⇒ There is no example because then all students would stick to it. We want you to be creative and come up with interesting stories. You can use the Labs as inspiration though.
- Our data set is very large and models take ages to fit? ⇒ you can work on a subsample
- In your labs you fit Linear Models with one predictor only, should we do that too? ⇒ NO, NO, NO! Please do not fit linear Models with one predictor only. This is only done for didactic purposes. In reality, you never fit a Linear Model with one predictor only. **Do fit models were all selected predictors are considered**
- The residuals of our linear model don't seems to be normally distributed and the their variances does not seem to be stable, shall use another dataset? ⇒ No, residuals diagnostics is not the main focus in this course. So, there is no need to change your dataset. **Note, however, that in the vast majority of cases you simply need to log-transform the response variable when fitting a Linear Model and the model assumptions are better fulfilled.** In particular, if the response variable are "amounts" (continuous variables that can only take positive values such as costs, weight, length...) then you must log-transform them!
- Shall we add a regression line on the exploratory graphs? ⇒ No, please always add smoothers to exploratory graphs (i.e. graphs where the response variable is plotted against a predictor).
- We found a very interesting data set, but it as too many predictors (e.g. 150), shall we find another data set? ⇒ No, you can work on a subsample of all the predictors. Just 10-20 predictors that you find most interesting.
- We found a very interesting data set, but it as too many observations (e.g. 350'00), shall we find another data set? ⇒ No, you can work on a subsample of all the observations. Just take a random subsample of 1-10k observations.
- Why didn't we get feedback about our group work? ⇒ Simply because you didn't follow the submission rule. Remember to name you .zip file appropriately and to upload it on ILIAS on time, and you will get your feedback. If you didn't do so, we are sorry, but we cant provide any feedback!
- Can we use the data set used for the R-bootcamp? ⇒ Yes, you can. Make sure that the overlapping between the two reports is minimal though.