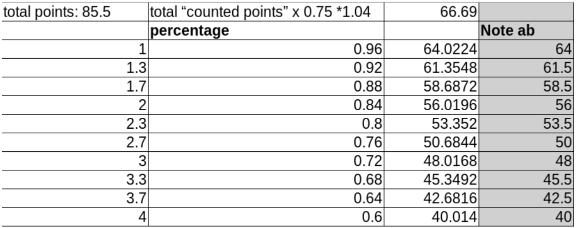
12 Final Projects

12.1 On Grading
12.2 Basic Considerations
12.3 Sample Projects
12.4 Pick one!

Updates

12.1 On Grading

- contrary to my first slides, I'm not allowed to round the $\frac{2}{3}$ regular homeworks (code+deliverables) + $\frac{1}{3}$ written report grade to thirds; the average will now be cut after the first decimal (generally did result in better grades for last years grades!)
- for the homework, the following grading table will apply



total points: 85.5	total “counted points” x 0.75 *1.04	66.69	
	percentage		**Note ab**
1	0.96	64.0224	64
1.3	0.92	61.3548	61.5
1.7	0.88	58.6872	58.5
2	0.84	56.0196	56
2.3	0.8	53.352	53.5
2.7	0.76	50.6844	50
3	0.72	48.0168	48
3.3	0.68	45.3492	45.5
3.7	0.64	42.6816	42.5
4	0.6	40.014	40

Requirements I

12.2 Basic Considerations

The basic idea of the project is to give you the possibility of applying the methods you learned on a bigger problem without the framework of: do that, get that! Thus, we expect you to:

- introduce the problem area and why/how Python/ML can solve the problem. State a research question you want to answer!
- describe all the methods (if not being part of the labs) and data you are using. Use a suitable, scientific citation style
- make graphics, which support your line of argumentation in answering the question.

Requirements II

12.2 Basic Considerations

- scientifically discuss your results and give appropriate references for any non-apparent insight.

So, a report could look like (you don't have to number the parts or take the headlines literally)

- Introduction: introduce the problem area, and define the question
- Methods: outline which kinds of methods you are employing and which implementations you use (e.g. the major libraries you already got introduced to: tensorflow, keras, sklearn or new/other things - most of them offer a scientific citation!)

Requirements III

12.2 Basic Considerations

- Data description: explain which data you use (cite the source appropriately!). Also take some time to describe properties of features with histograms/images/....
- Build a model/... working with the data. Explain in detail the challenges you've faced and how you built the model (cross-validation, neural network layer architecture, loss, regularization, ...)
- Make a conclusion on your results. Compare with other published works and detail any shortcomings of your work!

Requirements IV

12.2 Basic Considerations

Formal requirements are very straighforward:

- **no more than 5 A4 pages in the report** (each extra page will reduce the grade). A group should aim for 4 (so we can see your equal contributions!), a single person can do 2-3 if everything is there! Also you can do an appendix for graphics (but properly reference it in the text and put exemplary pictures in the core) which is not counted to the 5 page limit

Requirements V

12.2 Basic Considerations

- you can use an IEEE-template. We don't mandate any fixed style, but please *be consistent*. Just keep sane default settings for margins and Co., then tune it to fit, if you run into the length problem. If you have a lot of references, decide if you really read them (...) and if it's still taking more than half a page, reduce the font size suitably (here, small fonts are ok - just don't leave information out (but you can switch to a less verbose style)).
- on each reports first page, we want a title "block": the title (of your research question. Not "Report"), your names, your studies, your matriculation numbers and the location (TU Munich).

Requirements VI

12.2 Basic Considerations

- hand in all the code you created for your analysis. Also include the data. If the data is not easily available or too big, include a script, which fetches the data to a local directory (and which does not include any additional functionality). If the data is behind a sign-in-wall, upload it to sync-and-share and give a link.
- for nice figures, pdf-export is best, but with non-TEX-tools, it's probably hard to get that. Capture all figures with a number and a concise description. **Don't forget to make the legends and axis labels legible!**

If things go wrong

12.2 Basic Considerations

so the deadline approaches and your model is not working as it should.
first: don't panic, then:

- search for any possible faults/improvements in your model. If you find them: describe why this is the problem and how it could be fixed (if you don't have time for it any more)!
- do a *really thorough* literature research: did other people do similar things, or could it be that the data is not as good as advertised
- also try to describe your data really well - check if it's diverse by doing suitable histograms/visualization of the feature space

It won't be perfect this way, but it can still be very good!

Make a dashboard

12.3 Sample Projects

Create a dashboard for data, allowing different views and applying some simple (regression/clustering/...?) analysis tools on demand for a selection of data. The report will be relatively short here!

recommended tools plotly's dash; see the gallery (check it out for inspiration!)

skills Python, very technical. working into a new library (which also has a different plotting mechanism)

data I'd suggest some Covid-Data (easy to come by,for example). You can take any other dataset of suitable complexity (global warming, ...) - probably geodata is also interesting. Or a weather app with DWD open data

Image classification with neural networks I

12.3 Sample Projects

Training a state of the art image classification network from scratch is very expensive (think thousand of dollars for energy + a good chunck of labeled data), nowadays one uses prespecified models (ResNet, VGG16 ...) trained on certain datasets (Imagenet,...), then fixes the layers as a sort of “black-box-classification”-engine and modifies only input/output-layers to fit the problem before training.

recommended tools keras, a pretrained image classification network. GPU!

skills adapting code, time-management (you’ll probably need access to (our) GPU systems for training eventually, unless you have one yourself)

Image classification with neural networks II

12.3 Sample Projects

data shoulder implant X-ray data (look at the associated paper), basically every other non-large-scale dataset (breast cancer comes to mind)

Credit scoring the right way? I

12.3 Sample Projects

The logistic classification model we built for the credit scoring dataset was not too great. Here, you will research the state of the art to do it and demonstrate your research on a sample dataset.

recommended tools sklearn, LightGBM, XGBoost, literature

skills literature research, applying libraries

data
- the SBA credit dataset from the classification exercise - there's a bigger dataset available in the supplemental info

Credit scoring the right way? II

12.3 Sample Projects

- the ML-repositories data on Taiwan credit card defaults (the corresponding paper is bad)
- there's a book with signup and datasets (if you google, you'll find direct links) on the internet
- a small set from Germany long ago
- a really big dataset on kaggle (maybe choose a suitable subset)

Sentiment analysis

12.3 Sample Projects

Social networks enabled the collection of a mass of “people”-utterances. For purposes of power or business, it’s helpful to know, what these people seem to think in their mass of communication. That’s where sentiment analysis comes in, to enable classification of textual data according to their “positivity”.

recommended tools sklearn, maybe keras if you’re ambitious

skills data processing/feature engineering (text to usable feature vectors), new ML methods

data
- tweets on US airlines with sentiment labels
- a large IMDB-dataset
- collect your own and use with a pretrained model. Theoretically you can relatively easy grab tweets from twitter, but you don’t have labels (I’ve tried this once, so contact me, if you want to do this)

Customer/Market segmentation

12.3 Sample Projects

Nowadays, a business can collect a wide variety of information on each customer transaction. A important part of applied ML in this context is to use clustering techniques to identify relevant subgroups.

recommended tools sklearn

skills exploratory data analysis, learn how to visualize relatively raw data, new clustering methods

data

- a Instacart Market Basket dataset - the competition was for time-series prediction, but some people have done visualizations for clusters as well!
- British online retail dataset - probably relatively hard, but there's a paper
- ill people are customers too. No corresponding clustering paper (only statistical model building)

Customer "classification"

12.3 Sample Projects

Another big topic in ML for online businesses is building models for customer "churn" (customers leaving) or other events. Prediciting which customers are about to leave (or might be interested in buying more) can be thought of as a classification task

recommended tools keras, sklearn

skills building large scale classification models, reading code

data
- the KDD-cup of 1997 - predicting whether a marketing campaign is successful (in the 1998 competition, there's also a regression problem, which uses the same data) (you can of course also do the 1998-problem)
- the KDD-cup of 2009 - predicting 3 classes for customer behavior on a large dataset get the data here (probably the small version)

Predictive Maintenance

12.3 Sample Projects

Predictive maintenance is also a topic, which is relatively “hot” in the industry context. The idea is to use different measurable properties to predict failures of components or devices

recommended tools sklearn, keras

skills exploratory data analysis, time series prediction

data I guess businesses think, that this data is too important, but NASA has a very nice amount of data available (the most popular seems to be the turbine timeseries data); there’s alsos a paper having a nice overview over the problems. There’s also a dataset on elevators from Huawei and on water pumps in Africa.

Autoencoders for face recognition

12.3 Sample Projects

A very simple procedure used in face detection is is to compute a PCA of the image, yielding a so called eigenface.
Naturally, one can try to use an autoencoder as done in this paper or in this student project

recommended tools keras, GPU!

skills handling image data, augmenting your data, building a neural network

data I think the coolest think would be to actually predict the faces of friends and family? (in addition to some training data)

Activity Classification: Random Forests I

12.3 Sample Projects

Random Forests are a very important ensemble method for classification. Here you should apply it to a dataset, which was already classified with random forests and ideally other methods. Try to find if you can match the published results and compare your approach. Try to approach it from a cleanroom perspective, so don't only use their results to build your model.

recommended tools sklearn, you can also decide to try another method (tell us!)!

skills reproducing a research paper, using new methods

Activity Classification: Random Forests II

12.3 Sample Projects

data for example you can the use the Bar Crawl dataset, which has an associated conference paper; find out how creepy companies will get with your smartwatch data (or by a German company or SUPER CREEPY); also for "your" smart home

Gaussian process regression I

12.3 Sample Projects

Gaussian-Process based regression (which allows the prediction of an "uncertainty") is being heavily applied in the geostatistics community (as kriging) as well as in the engineering community for building "surrogate" model for FEM simulation

recommended tools sklearn or one of the more specialized GP-frameworks

skills understanding new learning models, understanding the use of learning for varying physical problems

data
- do a classic kriging analysis with geospatial data, for example found here: a wiki on different datasets or the page to a book on Geostatistics (especially Brenda.dat; site injects data into all files... YAAAAHOOOO)

Gaussian process regression II

12.3 Sample Projects

- do your very own surrogate model building. Simple FEM-problems seem to be found in some benchmarks (according to this paper doing a GP model for FEM simulations).
  YOU NEED TO BE FAMILIAR WITH FEM IF YOU WANT TO DO THIS (e.g. Comsol, Calculix, CodeASTER, Nastran) and we can support you with computing power, but not with technical details

Wildcard project I

12.3 Sample Projects

if you have your very own project idea, you can follow this as well:

- make a short summary like we did for all the others - listing relevant paper(s), methods and the data you have available
- if you want to do this in parallel to some job/thesis with overlapping data: tell your supervisor to get in contact with us. Tell us your supervisor

Send an e-Mail next week and we can setup some individual discussion on this!

Ideas for data:

- Forest cover types
- (US) census data (or any other public data)

Wildcard project II

12.3 Sample Projects

- Video characteristics and encoding resource requirements
- College Scorecard-Data (for clustering)
- Usage data of the ICE Wifi (clustering or geo visualization!)
- Flinkster booking data
- Oktoberfest foods classification: an free dataset by a ML group at TUM

Choose a project I

12.4 Pick one!

- make a shortlist, check out the paper abstracts and dataset descriptions, then decide for one!
- do so in a email **until 2021-07-11**. Name the topic you've choosen and your partner/group. Put your partner in CC.

If you don't pick a topic, we'll pick it for you (single person groups)!
NOTE for changes: you are free to change the scope of your project a little bit over time as that's what happens if you really work on a problem. That's, why we have some "topic" and "method"-focused problems. In general it's fine to change the data-area or the method, for example:

Choose a project II

12.4 Pick one!

- market-segmentation data: we don't want you to run clustering on astronomical data - but you can use any market/customer-data you find more interesting with any method you want.
- random-forest-method: stick to the random forests. If you find some more interesting data, use it (but of course, you can use a chosen dataset for trying other classification methods!)

So basically, we want the report to still focus on what's given as the broad topic!

BUT: If you change your data, talk to us first!

Organisation I

12.4 Pick one!

- **(PRELIMINARY) DEADLINE: 2021-09-15 @ 2330 Munich time.** (will check with the examination office for “Notenschluss” this week)
- we’ll make a break now, **please evaluate the topics on your own!**
- we’ll be available for hands-on consultation for all groups during the assigned hours in calendar week 27/30/33 and 36: ask questions on the project or how to prepare the data! (we’ll be available as long as you want there, but please write an e-mail before if you need a lot of time on a specific date)

Organisation II

12.4 Pick one!

- also: on 2021-07-26/27 do a project proposal presentation if you want to be allowed to the exam, but didn't present sufficient homework
- if you want another topic, this is your chance as well. If you show up and describe your problems you can **pick a new one until 2021-07-31**
- until the deadline: we'll try to answer the chat/email-questions at least once a week.

Organisation III

12.4 Pick one!

- you can also send in a draft anytime and if there are glaring errors or missed, but not properly communicated requirements, we will make an announcement to all of you! (subject to a high latency, so if you change stuff don't worry and bombard us with the updated things – our inbox doesn't mind and we don't waste time critizing old work)