Practicing Learning from Data

Preliminaries

Assignments

In this course 2 assignments need to be done:

- **Assignment 1 – Regression:** Your customer is a peer-to-peer credit marketplace in the US. It connects private lenders (people who hand out loans) with loan applicants. Your task is to provide a regression model that gives a recommendation to the lenders about the hight of a suitable interest rate for a given applicant (based on information that is available about the applicant).
- **Assignment 2 – Classificaton:** Provide a classification model for credit card customers using neuronal networks.

Main Task

Your main task is to provide a working model - one for each assignment. Every decision you took that lead you to your model must be documented and justified.

Groups

Both assignments are done in **groups of 4 students**. During both assignments the groups remain the same. We recommend that at least one member of the group has a background in programming.

Reality Check

In a real-world setting, your ultimate goal is to provide your customer with a model that makes good predictions (or gives good recommendations). Yet, since not all models will perform equally well for the data at hand, it is your task to try out different models, and to select the best one. To find out which of the models is the best, you will apply the validation set approach or the cross validation approach to the available historic data (that we, your customer, published for you on Moodle). After you will have found the best model, you will deploy it, so that it can be applied to new incoming data that you have never seen before.

In order to emulate the process of applying the model to new incoming data, we keep a part of the data set hidden from you (therefore also called “secret data”). After you have handed in your assignment, we will apply your best model to this “secret data” to see how well it performs in a real-world setting. This is the ultimate reality check for your model!

To do that, you need to provide us with a separate R file, in which (in this order)

1. we can specify the file name of the file that holds the “secret data”;
2. the file (i.e., the “secret data”) is loaded;
3. the new data is transformed in the same way as the published (historic) data; (This includes all transformations that you did before you trained/tested your model, including the splitting of the data into input and target data.)
4. your best trained model is loaded (functions are explained in the assignments);
5. your best trained model is evaluated with the new data.

You can test this file with the historical data published on Moodle!

Deliverables

At the end of the course your group hands in **2 deliverables – one for each assignment**. A deliverable for one assignment contains these two parts:

1. A PDF document containing a description and documentation of the solution. Specifically, the PDF document contains the following:
    a. An abstract with a summary of which methods and metrics you applied, and of your solution.
    b. An overview of all the important steps/cycles you performed. (These are the steps/cycles of the complete data science process discussed in the introductory lecture). i.e.
        i. which step/cycle you run,
        ii. what you changed in this cycle,
        iii. Very important: An explanation**/justification** of why you performed the steps/cycle in the way you did.
        iv. What you have learnt in this step/cycle
    c. A critical assessment of your solution. This includes its advantages and limitations.
    d. Answers to the questions raised in the main task description (see below).
    e. Lessons learnt (I.e., a reflection of the course assignment from your point of view.)
2. A ZIP file containing the code. Specifically, the ZIP file should contain the following:
    a. A saving of your best learnt model.
    b. The final version of your preprocessed historic data that you used for training and testing.
    c. The R code files of all trials/experiments you performed (including the preprocessing steps).
    d. An **executable** R-script ("main.R") that contains enough comments to guide a reader through your code. Please set the seed to one ("set.seed (1)"), so that we can reproduce your results.
    e. The file for the "reality check" described above that allows for loading and transforming the "secret data" set, and that evaluates it with your best trained model. **Without that working file you won't be able to get the points for the performance measurements (see below)**.

Deadline

Documents and R code files have to be delivered before **January 06, 2025, 12:00 (Swiss Time)**. Please upload them all in one zip file on Moodle. Please note: nobody stops you if you want to upload your solution earlier.

Assessment of the Assignments

We will evaluate each of your assignments in a 3-step process:

1. We assess the quality of your deliverables **(max. 80 points)**. Here we look at things like: Is your solution well documented? How much "brains" did you put into finding the best solution for your customer? Are your justifications logical and sensible? How much effort did you put into each of the steps of the "complete data science process" discussed in the introductory lecture? Etc.
2. We rate the performance of your best trained model on the historic data (the data published on Moodle) **(max. 20 points)**.
3. We rate the performance of your best trained model on the "secret" data set we use for the "reality check") **(max. 20 points)**.

This means you can achieve **max. 120 points per assignment**.

- To pass one of the 2 assignments, you need at least 45 points for this assignment.
- To achieve grade 6.0 for an assignment, you need at least 100 points for this assignment.
- If you achieve more than 100 points in one assignment, these points cannot be used to compensate for the other assignment or the exam.
- To compare model performance, we will apply the MSE in assignment 1 (regression), and accuracy in assignment 2 (classification).
- We rate the performance of your best trained model with the help of an assessment scheme. See the description for each assignment.
- Attention: Make sure your "Reality Check" file compiles! Otherwise you are assigned 0 points in steps 2 and 3.
- Notice that overfitting the model to the published historic data might give you the full 20 points in step 2, but very little points in step 1 and 3.

Practicing Learning from Data

Assignment 1 - Regression

Data Set

For the course assignment a version of the «Lending Club Loan Data» is used. The Lending Club (LC) operates an online peer-to peer credit marketplace in the US:

> "Lending club is a financial services company headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. At its height, LendingClub was the world's largest peer-to-peer lending platform. The company reported that $15.98 billion in loans had been originated through its platform up to December 31, 2015."
>
> (Wikipedia)

The data set published on Moodle contains real anonymized data describing **personal loans** issued through the Lending Club website. The data set contains historical data (i.e. loans from several years back) that can be used for training and testing. The original data set can be found on Kaggle. We use a modified version of the original data set.

The data dictionary provided on Moodle comes with the original data set. It provides rudimentary meta-data about the included features. In order to gain a sufficient business and data understanding (specifically, an understanding of the feature semantics), it is recommended to perform some research on the terminology used.

Main task

- Start with business understanding, data understanding and data exploration. According to your learnings in these steps, preprocess the data for model training and testing. (You may need to add preprocessing iterations later on.)
- Perform iterative experiments: Compute training error, test error and cross validation error of different regression models for the published data using the mean squared error and/or other appropriate metrics introduced in the lecture. Provide the R-script with all your experiments, and document your testing strategy, as well as your interpretations of the resulting evaluation scores in a pdf file (cf. delivery 1).
- Your experiments need to include at least one working regression model that gives a recommendation to lenders about a suitable interest rate (`int_rate`) for a given loan applicant, based on information that is available about the applicant.
- For your best model, provide an executable R script. Calculate its mean squared error ($MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ for n observations) on the full set of published data and document it in the pdf. Save the final version of your preprocessed data that you used for training and testing using `write.csv(dataset,"filename.csv")`.
- Additionally, save your best model using `saveRDS(model, file = "filename")` and provide the R script for reality check.

Assessment scheme

For assessing the performance of your model, we use the following assessment scheme:

MSE	Bonuspoints
$12 \leq MSE$	0
$11 \leq MSE < 12$	5
$10 \leq MSE < 11$	10
$9 \leq MSE < 10$	15
$MSE < 9$	20

That scheme will be applied for the published data (step 2) and the “secret” data (step 3).

Practicing Learning from Data

Assignment 2 - Classification

Data Set

For this assignment a version of the **«A Credit Card Dataset for Machine Learning»** is used. It contains real (anonymized) data describing information from customers applications together with their "pay-back behavior", i.e. how good did they pay back their credit card depts. The idea is that only customers are accepted in future which have a predicted good "pay back behavior". Here in the assignment, we focus on the prediction of the customer behaviour – *not* if we should accept or reject the credit card application! The original data set can be found at *kaggle* (https://www.kaggle.com/rikdifos/credit-card-approval-prediction). For the assignment we use modified versions of the data set. Please find them on Moodle.

Data Exploration & Preparation

- Our goal in the second assignment is to predict how good a (new) customer will pay back their credit card depts. In the data set application data from current customers (the first 18 attributes) together with their `status` (last attribute; target) are given.
- The attributes from the applications are

Attribute Name	Explanation	Remarks
ID	Client number	
CODE_GENDER	Gender	
FLAG_OWN_CAR	Is there a car	
FLAG_OWN_REALTY	Is there a property	
CNT_CHILDREN	Number of children	
AMT_INCOME_TOTAL	Annual income	
NAME_INCOME_TYPE	Income category	
NAME_EDUCATION_TYPE	Education level	
NAME_FAMILY_STATUS	Marital status	
NAME_HOUSING_TYPE	Way of living	
DAYS_BIRTH	Birthday	Count backwards from current day (0), -1 means yesterday
DAYS_EMPLOYED	Start date of employment	Count backwards from current day(0). If positive, it means the person unemployed.
FLAG_MOBIL	Is there a mobile phone	
FLAG_WORK_PHONE	Is there a work phone	
FLAG_PHONE	Is there a phone	
FLAG_EMAIL	Is there an email	
OCCUPATION_TYPE	Occupation	
CNT_FAM_MEMBERS	Family size	

- The last attribute `status` contains the "pay-back behavior", i.e. when did that customer pay back their depts:
  - 0: 1-29 days past due
  - 1: 30-59 days past due
  - 2: 60-89 days overdue
  - 3: 90-119 days overdue
  - 4: 120-149 days overdue
  - 5: Overdue or bad debts, write-offs for more than 150 days
  - C: paid off that month
  - X: No loan for the month

  Please note: We are learning only the pay-back behavior. The decision, i.e. if we accept a customer or not, is done in another process step – not here!

Main task

- Design your network. Why did you use a feed-forward network, or a convolutional or recursive network – and why not?
- Find a "reasonable" good model. Argue why that model is reasonable. If you are not able to find a reasonable good model, explain what you all did to find a good model and argue why you think that's not a good model.
- It's recommended to use k-fold validation (with *k = 10*) to find the best hyperparameters for your network.
- We will publish some historical data on Moodle. For the learning purpose you may split up the historical data into training, test, and validation data set. But for the assessment of the performance of your model we will use the whole data set, i.e. the accuracy of the model is determined on the whole historical data.
- Save your trained neural network with `save_model_hdf5`. Also save your data sets you used for training, testing and validation.

Assessment scheme

For assessing the performance of your model, we use the following assessment scheme:

Accuracy	Bonuspoints
< 80%	0
≥ 80% but < 85%	5
≥ 85% but < 90%	10
≥ 90% but < 95%	15
≥ 95%	20

That scheme will be applied for the published data (step 2) and the "secret" data (step 3).

Some hints

- Data preprocessing is easier here; no feature engineering is needed.
- You may be able to reuse parts of the exercises we used in our examples during lectures.
- All in- and output values need to be floating numbers (or integers in exceptions) in the range of [0,1].
- Please note that a neural network expects a R matrix or vector, not data frames. Transform your data (e.g. a data frame) into a matrix with `data.matrix` if needed.
- There are some models which show an accuracy higher than 90% (!) for training (and test) data – after learning more than 1000 epochs.