Team5 project work plan

* Vékony Vivien (NY51MS)
* Galgóczki Nobert (LTPS80)
* Krezsó Krisztián (C39QFI)
* Ocskó Márk (CY95DD)
* Öhlsläger Tamás (AIQZV5)

Machine Learning practice – 2022

# Format requirements

The sections must not exceed **1** page including tables or figures as well. The font size must be **12**. Meaningful fluent text is preferred over plain lists or tables except when tables save space. For **maximum points** all the listed questions must be answered and also some additional information(**new ideas**) needed to show evidence that the team has investigated the task deeply.

## Document submission

The expected upload format is pdf. You can upload the document to coospace. Each submission should contain the previous sections, i.e. when submitting Milestone- II the document will include the Milestone - I as well.

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone name** | **Point** | **Milestone Deadline** | **Document submission deadline** |
| Task description | +5 | - | 2022-02-28 |
| Data processing (MS - I) | 20 | 2022-03-07 | 2022-02-28 |
| Modelling (MS - II) | 20 | 2022-03-28 | 2022-03-21 |
| Further improvements (MS - III) | 20 | 2022-05-02 | 2022-04-25 |
| Project presentation | 5 | 2022-05-09 | - |
| Individual tasks | 35 | 2022-05-02 | 2021-04-25 |

# Task description

* What is the task?
  + …
* task type
  + (classification, regression,...)?
* Motivation
  + Why did you choose this?
  + Why is it useful to solve the problem?
  + etc.
* Dataset description
  + data format
    1. csv, jpeg,..
  + what is the meaning of the features?
    1. raw data, features
  + prediction target
    1. what does it mean?
  + dataset size
    1. number of features
    2. training instances
  + etc.
* Environment and tools you plan to use:
  + Programming language
  + machine learning library
  + …
* Github repository link: <https://github.com/OcskoMark/GepiTan_Team5.git>
* other relevant information here...

# Data processing (MS - I)

* *Data splitting strategy*
  + *ratio of training, validation and test set*
  + *choose adequate sizes to support your experiments*
* *List some ideas for features*
  + *basic features given in the dataset(BASIC)*
  + *new features: feature A, feature B,...*
* *Mention what do you plan to do with missing values if there are any?*
  + *drop them(DROP), replace them by the average(AVG)...*
* *Name some preprocessing methods that you plan to apply*
  + *MinMaxscaling(MMS), StandardScaler(SS), PolynomialFeatures, OnehotEncoder,..*
* *other ideas here..*

Expected output: a set of scripts, which prepare the data for your models in the planned formats. An example script call, might be:

`python data\_processing.py --train\_size 10000 --val\_size 1000 --features “BASIC” --missing\_values “DROP” --preprocessing “MMS”`

# 

# Modeling (Milestone - II)

* Metric used for evaluation
  + accuracy, F1 score,..
  + what is the proper metric(s) for your problem?
* Baseline method
  + e.g. most frequent class classifier
  + a good baseline will help you to identify whether your models are learning or not.
* List some ideas for models
  + KNN, Logistic regression, SVM,..
* Hyper parameters you plan to optimize
  + k for KNN, l2 coefficient for Logistic regression, kernel\_type for SVM,...

Expected output: a set of scripts, which is able to train the listed models and evaluate them according to the metrics. Beside, the scripts are able to optimize some hyper parameters and save the best model. An example script call, might be:

`python knn\_training.py --dataset BASIC.csv --k\_values “1,2,3,4,5”`

`python knn\_training.py --dataset BASIC\_FEATURE\_A.csv --k\_values “1,2,3,4,5”`

`python svm\_training.py --dataset BASIC\_FEATURE\_A.csv --kernel\_types “poly,linear,..”`

`python evaluate.py --model baseline --metric “accuracy,f1”`

`python evaluate.py --model knn.model --metric “accuracy,f1”`

`python evaluate.py --model logreg.model --metric “accuracy,f1”`

# 

# Further improvements (Milestone - III)

For maximal points, improvements must be achieved in **two different directions** at least two times. That is, you need to select two directions, and improve the best solution two times according to each direction.

Possible directions:

* **reducing training time, while preserve model performance**
  + by using Feature selection techniques
  + Experimenting with the training data size, like 25,50,75%
  + dimensionality reduction
  + by some other ideas
* **improving performance**
  + by some ensemble techniques i.e. combining multiple models
  + by some advanced models not used during the course
  + by involving other data sources
  + by some other ideas
* **or might list a new direction**
  + your own suggestion1
  + your own suggestion2

# Individual tasks(Member-1)

The aims of the individual tasks could be:

* further performance improvements
* efficiency improvement by reducing training or prediction time
* investigating other data sources
* developing new features and evaluate their usefulness
* evaluate the best method on other related dataset
* your own idea

For maximal points, two aims should be selected.

## Task -1

## Task -2

# Individual tasks(Member-2)

The aims of the individual tasks could be:

* further performance improvements
* efficiency improvement by reducing training or prediction time
* investigating other data sources
* developing new features and evaluate their usefulness
* evaluate the best method on other related dataset
* your own idea

For maximal point, two aims should be selected.

## Task -1

## Task -2

# Individual tasks(Member-3)

The aims of the individual tasks could be:

* further performance improvements
* efficiency improvement by reducing training or prediction time
* investigating other data sources
* developing new features and evaluate their usefulness
* evaluate the best method on other related dataset
* your own idea

For maximal point, two aims should be selected.

## Task -1

...

## Task -2

...

# Individual tasks(Member-4)

The aims of the individual tasks could be:

* further performance improvements
* efficiency improvement by reducing training or prediction time
* investigating other data sources
* developing new features and evaluate their usefulness
* evaluate the best method on other related dataset
* your own idea

For maximal point, two aims should be selected.

## Task -1

...

## Task -2

...

# Individual tasks(Member-5)

The aims of the individual tasks could be:

* further performance improvements
* efficiency improvement by reducing training or prediction time
* investigating other data sources
* developing new features and evaluate their usefulness
* evaluate the best method on other related dataset
* your own idea

For maximal point, two aims should be selected.

## Task -1

...

## Task -2

...

# Suggestions for Project presentation (Milestone 3+1)

Unfortunately, too many presentations are of bad quality because of the slides. It is impossible to give detailed rules here on how to create good slides, because generic rules simply do not exist. But it is easy to give some "do" and "don't" tips...

* choose an adequate number of slides; for a 15 minutes presentation, around 10 slides (including titles) is OK.
* do not take several slides (and therefore several minutes...) to present concepts that everybody knows in the class (for example, do not explain what is an svm or a neural network. We have learned it during the course.)
* on the contrary, do not include too many details in your presentation; never forget that your talk will be appreciated only if the participants understand it (at least at 90%)!
* use a top-down approach to attract your audience
* Spend time on thinking about what you will present?
  + You will need to prioritize your results, for sure you won't have time to show everything what you have done during the semester
* What are the best graphs to present your results? A good graph worth is worth a thousand words
  + i.e. bar chart for comparing models
  + scatter plot for showing the relations of two variables
  + ...
* Spend time on motivating your work
* probe your talk in front of several people
* If you have some suggestions, write it to me. It might get included.