DataVaders - Projektterv

Csiszár András, Mészáros Zsolt, Szakál Mátyás, Vass Máté

Gépi tanulási módszerek gyakorlat – 2022

# Formai követelmények

Minden szekció maximum **1** oldal hosszú lehet. Betűméret: **12**. Ahol lehet, folyó szövegben fogalmazzunk listák és táblázatok helyett, hacsak nem tudunk azokkal sok helyet megspórolni. **Maximális pontszám** eléréséhez az összes felsorolt kérdésre válaszolni kell, valamint saját ötleteket is tartalmaznia kell a leírásnak, ezzel bizonyítva, hogy a csapat mélyen megvizsgálta a feladatot.

## Dokumentáció beadása

Az elvárt formátum: pdf, ezt CooSpace-en kell beadni. Minden beadás tartalmazza a korábbi szekciókat is, például a 2. Mérföldkő minden előtte lévő Mérföldkövet és szekciót tartalmaz.

| **Mérföldkő neve** | **Pontszám** | **Mérföldkő határideje** | **Dokumentáció beadásának határideje** |
| --- | --- | --- | --- |
| Feladatleírás | +5 | - | 2022-02-28 |
| Adatfeldolgozás (MK - I) | 20 | 2022-03-07 | 2022-02-28 |
| Modellezés (MK - II) | 20 | 2022-03-28 | 2022-03-21 |
| További fejlesztések (MK - III) | 20 | 2022-05-02 | 2022-04-25 |
| Projekt prezentálása | 5 | 2022-05-09 | - |
| Egyéni feladatok | 35 | 2022-05-02 | 2021-04-25 |

# Feladat leírása

* Mi a feladat?
  + A projekt során egy felügyelt gépi tanulási problémát oldunk meg egy általunk tetszőlegesen választott adathalmazon. A munkához egy banki adatbázist választottunk, amiben a bank az ügyfeleiről tárol el különféle adatokat. Ezek alapján azt szeretnénk megjósolni, hogy egy új ügyfél a jövőben várhatóan fog-e betéti számlát nyitni.
* Feladat típusa
  + Osztályozás.
* Motiváció
  + A meglévő adatok alapján megpróbálhatjuk megjósolni, hogy egy új ügyfél tervez-e megtakarítást elhelyezni a bankban.
  + Miért hasznos ez a feladat?
    1. Így céltudatosan tud a bank hirdetéseket és telefonos ajánlatokat tenni bizonyos ügyfeleknek.
* Adathalmaz leírása
  + Adat formátuma
    1. .csv
  + Adatsorok jellemzői
    1. Rendezett adatsoraink vannak
    2. Jellemzők:
       1. *Kliens adatai*: életkor, munka fajtája, kapcsolati státusz, tanulmányok, van-e hiteltartozása, átlagos éves egyenleg, van-e lakáshitele, van-e személyi hitele
       2. *Legutóbbi kapcsolatfelvétel jellemzői:* kapcsolatfelvétel típusa, kapcsolatfelvétel hónapja, kapcsolatfelvétel napja, kapcsolatfelvétel időtartama
       3. *Egyéb jellemzők:* kapcsolatfelvételek száma, utolsó kapcsolatfelvétel óta eltelt napok száma, eddigi összes kapcsolatfelvétel száma, előző kapcsolatfelvétel kimenetele
    3. Címke:
       1. Az ügyfélnek van-e a banknál betéti számlája?
  + Prediktálási cél
    1. F-e az ügyfél a jövőben betéti számlát nyitni?
  + Adathalmaz mérete
    1. 45211 felcímkézett rekord
    2. 16 jellemző vektor
    3. 1 bináris (igen/nem) címke
* Használt környezet és eszközök:
  + Programozási nyelv
    1. Python (Google Colaboratory Notebook)
  + Gépi tanulást megvalósító könyvtárak
    1. NumPy, pandas, scikit-learn
  + Verziókövetés
    1. Github repository: https://github.com/material1999/gepitan\_project

# Adatfeldolgozás (MK - I)

* Adatfelosztási módszer
  + Mivel nagyon nagy mennyiségű felcímkézett példa áll rendelkezésünkre, így terveink szerint 70%-15%-15% felosztást fogunk alkalmazni (70% train, 15% dev, 15% test)
  + Így nagyjából 30 ezer adaton tudunk majd betanítani, továbbá 7-7 ezer adaton fejleszteni, valamint tesztelni.
* Tanuláshoz felhasznált jellemzők
  + A tanuláshoz terveink szerint az adathalmazunk jellemzésénél felsoroltak közül az *“1. Kliensek jellemzői”* attribútumokat szeretnénk felhasználni. Ezeket gondoltuk a legfontosabb információknak a feladatunk szempontjából, a többi jellemző kihagyásával elkerülhetjük az esetleges túltanulást.
  + Összehasonlításképp tervezzük a modellünk tanítását egy bővített jellemző halmazon is elvégezni, és az így kapott eredményeinket különböző metrikák szerint összehasonlítani a szűkebb halmazon kapottakkal.
* Mi fog történni az üres cellákat tartalmazó rekordokkal, ha vannak?
  + Mivel azoknak a rekordoknak a száma, amikben van üres cella, nem jelentős, ezért egyszerűen el fogjuk dobni őket (körülbelül 2 ezer rekord a 45 ezerből). Azért is döntöttünk így, mert 1-1 hiányzó adatot nem egyértelmű, mely másik rekordok alapján lehet hatékonyan prediktálni.
* Előfeldolgozási lépések
  + A tanításhoz használt jellemző vektorok közül a szöveges információk csak adott, igen kis méretű halmazokból vehetnek fel értékeket. Így ezeket először egy OrdinalEncoder segítségével egyenletes eloszlású számsorrá alakíthatjuk, majd egy MinMaxScaler segítségével lenormálhatjuk a [0,1] intervallumba.
  + Ezek után a többi, számszerű adatot tartalmazó jellemző vektort egy StandardScaler segítségével fogjuk transzformálni, ezzel a kapott oszlopokban az átlagunk 0, a szórásunk pedig 1 lesz.
  + A két transzformáció eredményét összesítve készen is áll az adathalmazunk a később definiált modellünk betanítására.
* A futtatásokhoz jelenleg nem tervezünk külső szkript-eket írni, a futtatási paramétereket külön kód blokkokba szervezzük ki a egyes szekciók elején, és minden programkód a Google Colaboratory Notebook-ban lesz megtalálható.

# 

# Modellezés (MK - 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”`

# 

# További fejlesztések (MK - 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

# Egyéni feladatok (Csiszár András)

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

# Egyéni feladatok (Mészáros Zsolt)

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

# Egyéni feladatok (Szakál Mátyás)

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

...

# Egyéni feladatok (Vass Máté)

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

...

# Egyéni feladatok (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.