Machine Learning

Applied Computational Intelligence, High Performance Computing
(sem. 1)

Prof. dr. Czibula Gabriela
Contact: gabriela.czibula[at]ubbcluj.ro

I. Aims of the activity

1. To introduce the fundamental principles, techniques, and applications of Machine Learning.
2. To cover the principles, design and implementation of learning programs which improve their performance on some set of tasks by experience.
3. To offer a broad understanding of fundamental machine learning algorithms and their use in data-driven knowledge discovery.
4. To get familiarized with the methodology of research in the field of Machine Learning.
5. To offer an understanding of the current state of the art in machine learning to conduct original research in machine learning.

II. Specific competencies acquired

Professional competencies

1. Understanding the concepts, methods and models used in Machine Learning.
2. Understanding the principles, design, implementation and validation of learning systems.
3. Learning to conduct incipient original research in machine learning.

Transversal competencies

1. The ability to apply machine learning techniques in solving real world problems.
2. Responsible execution of lab assignments, research and practical reports.
3. Application of efficient and rigorous working rules.
4. Manifest responsible attitudes toward the scientific and didactic fields.
5. Respecting the professional and ethical principles.

III. Course content

1. Introduction in Machine Learning. Statistical Foundations.
2. Supervised learning.
    2.1. Tree-based learning.
    2.2. Neural networks-based learning.
    2.3. Support Vector Machines.
    2.4. Bayesian learning.
    2.5. Instance based learning.
3. Unsupervised learning.
4. Reinforcement learning. Other learning models and paradigms.

IV. ML activities

All activities require physical participation.

- The course materials (lecture notes, books, bibliographic material) will be available in **Files/Class materials** before the lecture hours.
- The lecture hours on Weeks 9-14 and lab hours on Weeks 8-12 will be allocated for research reports presentations (face-to-face), according to the planning available in **Files/Class materials/TimePlanningReport.docx.**
- The lab assignments must be submitted on MSTeams (through the corresponding Assignments) before the lab class.
- The research reports must be submitted on MSTeams (through the corresponding Assignments) at least 48h in advance of the presentation date.

Planning

Lectures

Weeks 9-14 – student presentations (research reports)

Labs

Lab 2 (Weeks 3/4) - Establishing the 2-person teams for the software project
Labs 3 – 4 Discussions on the lab assignments, research reports, other ML topics
Lab 5, 6 (Weeks 9-12) – Research reports presentations (face-to-face)
Lab 7 (Weeks 13/14) - Software projects presentation (face-to-face)

V. Bibliography

1. Mitchell, T., *Machine Learning*, McGraw Hill, 1997 (available at www.cs.ubbcluj.ro/~gabis/ml/ml-books)

2. Nillson, N., *Introduction to Machine Learning*, Stanford University, 1996 (available at www.cs.ubbcluj.ro/~gabis/ml/ml-books)

3. Sutton, R.S., Barto, A.G., *Reinforcement learning*, The MIT Press Cambridge, Massachusetts, London, England, 1998 (http://incompleteideas.net/book/the-book.html)

4. Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning*, MIT Press, 2016 (online edition at http://www.deeplearningbook.org/)

5. Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola, *Dive into Deep Learning*, 2020 (http://d2l.ai/)

6. Li Deng and Dong Yu, *Deep Learning*. *Methods and Applications*, Foundations and Trends® in Signal Processing, Volume 7 Issues 3-4, 2014 (https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DeepLearning-NowPublishing-Vol7-SIG-039.pdf)

7. Cristiani, N., *Support Vector and Kernel Machines*, BIOwulf Technologies, 2001

Journals

1. *Journal of Machine Learning Research* is a freely available WoS journal - http://jmlr.csail.mit.edu/
2. *Machine Learning* - http://www.springer.com/computer/artificial/journal/10994

3. *Expert Systems with Applications* - https://www.journals.elsevier.com/expert-systems-with-applications
4. *Neural Computation* - http://www.mitpressjournals.org/loi/neco
5. *Neural networks* - http://www.elsevier.com/wps/find/journaldescription.cws_home/841/description#description
6. *IEEE Transactions on Pattern Analysis and Machine Intelligence* - *http://www2.computer.org/portal/web/tpami/*
7. *Bioinformatics* is a journal focusing on analysing biological data - http://bioinformatics.oxfordjournals.org/
8. others

….

Conferences

International Conference on Machine Learning, Neural Information Processing Systems, Conference on Computational Learning, International Conference on Knowledge Discovery and Data Mining, European Conference on Machine Learning, and others.

Resources

- ML repositories
  http://archive.ics.uci.edu/ml/
  https://www.kaggle.com/datasets
- RL repository
  http://www-all.cs.umass.edu/rlr/
- RL resources
  - https://aikorea.org/awesome-rl/
  - http://busoniu.net/repository.php
- SBSE repositories
  - http://promise.site.uottawa.ca/SERepository/
  - http://openscience.us/repo/
- Object recognition data sets
  - https://cvgl.stanford.edu/teaching/cs231a_winter1314/lectures/datasets.pdf
- Computer vision image datasets
  - http://www.cs.utexas.edu/~grauman/courses/spring2008/datasets.htm
- Other data sets
  https://vincentarelbundock.github.io/Rdatasets/datasets.html

VI. Activity

Two components are required: a **research report** (**Section VII**) and a **software project** (**Section VIII**).

VII. Research report

A **research report** on an ML-based topic/technique, based on some recent research papers that will be **prepared by each student individually**.

a) a written paper of 9-10 pages using the LaTeX template available in the folder Files/Class Materials/LaTeX template (research paper)

- The use of the LaTeX template is compulsory.
- Overleaf - LaTeX editor (online)
- TeX Live - https://www.tug.org/texlive/ (offline)

b) an oral presentation

- ppt/tex + Q&A
- a one-page **outline** of the presentation

Requirements

The research report will present a survey of some recent research papers (about 5-10 titles) on the chosen topic.

The report should fulfill the requirements of a research paper:

- suggestive title corresponding to the content;
- about 10-15 lines abstract;
- introductory section, introducing the topic and detailing the purpose and structure of the paper;
- a section presenting the importance and relevance of the chosen topic from a theoretical and practical perspective (application domains)
- main sections presenting recent approaches and experimental results on the chosen topic
- a discussion section which analyses the literature approaches reviewed in the previous sections from a critical perspective (evaluation of methodology, data sources, results and findings)
- concluding remarks and further work section
- bibliography of more than 5 titles; the bibliography entries should be written **correctly and completely**; all the bibliography items must be cited in the text.

E.g.,

- Manevitz, L., Yousef, M., 2007. One-class document classification via neural networks. Neurocomputing 70, 1466–1481 (**@article** bibtex entry)
- Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014, pp. 1188–1196 (**@inproceedings** bibtex entry)

- Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press (**@book** bibtex entry)
- Google. Google Code Jam Competition. https://codingcompetitions.withgoogle.com/codejam. Online; accessed 15 September 2021 (**@misc** bibtex entry)

Written paper

– The written paper will have to be submitted at least **48h** in advance of the presentation date.
  - The research reports will be checked against plagiarism.
  - The maximum allowed similarity score is 25%.

Use of generative AI and AI-assisted technologies

Elsevier’s AI author policy

- “Authors are allowed to use generative AI and AI-assisted technologies in the writing process, but **only to improve the language and readability of their paper** and with the appropriate **disclosure.”**
- **“Authors must disclose the use of generative AI and AI-assisted technologies in the writing process by adding** a statement at the end of their manuscript in the core manuscript file, before the References list. The statement should be placed in a new section entitled ‘Declaration of Generative AI and AI-assisted technologies in the writing process’.

  Statement: During the preparation of this work the author(s) used [NAME TOOL / SERVICE] in order to [REASON]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Oral presentation

- The oral presentation of the research report is not compulsory.
- The scheduling for the research report presentations can be found in **Files/Class materials/TimePlanningReport.docx**.
  - the one-page outline of the presentation must correspond both to the written text and to the oral presentation, and should be self-explanatory;
  - the one-page outline is graded only if the oral presentation will be delivered.
- **10 minutes** will be allocated for the oral presentation + about **5 minutes** for questions and discussions.
- The structure of the presentation (Powerpoint, LaTeX) should reflect the structure and content of the research report.

The scheduling for the oral presentations should be respected. If a presentation is not delivered on the due date, it cannot be rescheduled. In this case, the grade for the theoretical report will include only the grading of the written article.

Deadline

Week 4 – a one page **presentation document** containing:

- a brief description of the approached topic and its importance (relevance) (3-4 paragraphs);
- the title of the research report;
- bibliographical references (at least 5);
- the **presentation date**
  - before selecting your presentation date please make sure that there are free slots available (in the document **Files/TimePlanningReport.docx**) and include your **name** and the **title of the report** on the available slot in the document

Grading

- Identical reports are **NOT** allowed.
- The **score** for a **research report** is composed as follows:
  - **5%** the grade for the **presentation document**;
  - **95%** the grade for the written paper and the oral presentation (if the case) computed as follows:
    - 0.5 p = one-page outline (graded only if the oral presentation will be delivered);
    - 0.75 p = abstract, introduction and title (corresponding to the content);
    - 2.5 p = main sections (including experimental results);
    - 1 p = discussion section;
    - 0.5 p = conclusions and future work;
    - 0.75 p = bibliography accuracy;
    - 2.0 p = quality of oral presentation
    - 2.0 p = accuracy of question answering
- The delay of **one week** in providing the document with the theoretical report topic, title and presentation date will be penalized with **1 point**.
- The failure to deliver the presentation at the due date will lead to a grade penalty of **1 point** for each delayed week.

VIII. Software project

The **software project** will have to demonstrate the use of **one** or **two** **traditional**/**conventional** ML techniques on (one or two) specific learning task(s) and will be developed in **teams of 2 students** (**the teams will be formed at your choice**).

Requirements

- A **labeled tabular data set** (from existing data repositories) with instances characterized by predefined/handcrafted features will be chosen.
- **One** or **two** learning tasks (**at your choice**) will be **defined** on the chosen data (e.g., supervised classification and/or supervised regression, supervised classification/regression and/or unsupervised classification, supervised classification/regression vs reinforcement learning)
- For **each** learning task, a **classical** ML model will be chosen, and **two** implementations are required:
    - an implementation **from scratch**, without using existing ML environments
        - the ML model (learning algorithm, hyperparameters optimization, etc) will be **fully implemented**
        - existing libraries are allowed only for **data visualisations** or **numerical computations** (e.g., numpy)
    - an implementation using an existing **ML software**
        - Python libraries (Scikit-learn, Keras, etc)
        - WEKA http://www.cs.waikato.ac.nz/ml/weka/
        - RapidMiner http://rapid-i.com
        - Orange http://www.ailab.si/orange/
        - ROCKIT http://xray.bsd.uchicago.edu/krl/KRL_ROC/software_index.htm
        - SVM software http://www.support-vector-machines.org/SVM_soft.html
        - MATLAB
        - others
- In the case of **two** learning tasks, the ML models should be chosen from different classes of learning methods (e.g., network-based, tree-based, instance learning-based, etc).

Project components

All documents required for the software project should be provided as .pdf files generated in LaTeX using the template available in the folder Files/Class Materials/LaTeX template (lab assignment)

- **the use of the LaTeX template is compulsory;**
- **for each of the specific requirements (from each component) you should include sections/subsections in your LaTeX document to increase its readability.**

The following **components** are required for the software project:

(1) Definition of the **learning task(s)** (doc)

- problem(s) definition (the statement in natural language, **what** should be solved);

  - problem(s) specification (input data and preconditions, output and postconditions);
  - specification of the learning task(s) (**Task**, **Performance**, **Experience**);

(2) Data analysis (doc)

  - analysis of the features used in learning
    - correlation;
    - independence;
    - feature importance;
  - data statistics
  - data visualization and interpretation

(3) Used **ML technique(s)** (doc)

  - brief description of the employed ML technique(s)
  - design of each of the learning tasks
    - target function to be learned (formal definition)
    - representation of the learned function
    - learning algorithm
    - learning hypothesis

(4) Related work summary (doc)

  - a summary of the literature approaches addressing the same learning task(s) and/or employing the same data set, including their performances (obtained results)

(5) Experimental results and discussion (doc);

  - details about the ML model(s) implemented from scratch
    - architectures, specific details
    - hyperparameters setting (a hyperparameter optimization should be used)
  - experimental results - summary (in a tabular form) of the performances of the ML model(s) implemented from scratch
    - for performance evaluation of all ML models
      - cross-validation must be used;
      - a statistical analysis of the results must be provided (e.g., standard deviation, confidence intervals).
    - evaluation measures
      - **for supervised classification**: accuracy, precision, recall, sensitivity, f-measure, AUC (Area under the ROC curve), AUPRC;
      - **for regression**: MAE (Mean of Absolute Errors), RMSE (Root Mean Squared Error), NRMSE (Normalized Root Mean Squared Error), R2
      - **for unsupervised classification (clustering):** both **internal** and **external** evaluation measures
  - discussion
    - for each of the ML model(s) implemented from scratch
      - analysis/interpretation of the obtained performance
      - explainability/interpretability of the models (using a selected algorithm – LIME, SHAP, etc)

        - comparison to related work
            - summarizing (in a tabular form) the values for the performance metrics for:
                - the model(s) implemented from scratch
                - the model(s) implemented using a library
                - the results from literature (related work)
            - details about the implementations from existing libraries should be given (these models should be tested using the same testing methodology as for the from scratch implementations
    - comparative analysis and interpretation of the performances obtained for the two selected learning tasks (if it is the case)

(6) the electronic version of the source code (.zip archive submitted through the corresponding assignment)

(7) **QA** session regarding the source code and the project implementation (face-to-face during Lab7) - ONLY if all the required documentations (components 1-6) were delivered.

Deadlines

Lab 2 - Establishing the 2 student-teams (the teams will be included in the documents below):

- **HPC group** - Files/Lab/HPC Lab - Teams software project.xls
- **ICA1 group** - Files/Lab/ICA1 Lab - Teams software project.xls
- **ICA2 group** - Files/Lab/ICA2 Lab - Teams software project.xls

Lab 3 – component (1)
Lab 4 - component (2)
Lab 5 – component (3)
Lab 6 – component (4)
Lab 7 – components (5) + (6) +(7)

Grading

- Identical projects are **NOT** allowed.
- The **maximum** grade
    - for projects that include **two** learning tasks is **10**
    - for projects that include **one** learning task is **6**
- The grade for the software project is computed as a weighted **average** of the grades received for the required components (10% components 1 and 4, 15% components 2 and 3 , 25% component 5 and 25% components 6+7).
- The delay of **one week** in completing a lab assignment will be penalized with **1 point**.

IX. Grading

The **final grade** is computed as follows:

10%	Class attendance and activity
45%	Research report
45%	Software project

The activities (components of the final grade) **are not compulsory**.

If **components (5) + (6) + (7)** for the software project will not be delivered at **Lab7**, they may be delivered in the **examination session**, with a penalty of **2 points** for the delay**.**

- A minimal final grade of 5 is required to pass the course.

X. Semester and final grades

The grades received during the semester, as well as the final grades for the **Machine learning** course are available at the following link

https://docs.google.com/spreadsheets/d/e/2PACX-1vTfhb55FcjUHHgjlppBC3cYjR5gNc0nYdcbJRXW88CPt23rwXyPoanz19rnaI7ztobW24CKK8XQfBJg/pubhtml?gid=1675021479&single=true

The code used for displaying the grades is the **unique identification code** (from Academic Info).

XI. Retake session

– The activities (research report/software project) can be graded in the retake session, excepting the oral presentations, but **ONLY** if at least one activity (research report or software project) has been (partially) completed during the semester.
  - The maximum grade for the **research report** in the retake session is **5.5**.
  - The maximum grade for a **software project** in the retake session is **6**.

– **No** report or software project can be resubmitted for grade increase in the retake session.

XII. Possible topics for the Theoretical Report (not an exhaustive list)

The topics below are suggestions only. Please feel free to choose other ML related topics.

1. One-class classification
2. Recurrent neural networks
3. Time delay neural networks
4. Self organizing maps
5. Hebbian learning
6. Semi-supervised learning
7. Radial Basis Function networks
8. Decision Trees (fuzzy, lazy, etc)
9. Bayesian learning
10. Machine learning in bioinformatics

11. Machine learning in software engineering (*search-based software engineering*)
12. Instance based learning
13. Case based learning
14. Boosting algorithms
15. Bagging algorithms
16. (Deep) Q-learning
17. Adaptive clustering
18. (Deep) Support vector machines
19. Kernel methods in machine learning
20. Association rule mining
21. Hidden Markov Models
22. Generative earning
23. Few-shot learning
24. Deep Reinforcement Learning
25. Autoencoders
26. Belief network learning
27. Generative models
28. Siamese networks
29. Contrastive learning
30. Representation learning

a.s.o