AI-MACHINE LEARNING FOUNDATIONS

Bachelor in Data and Business Analytics BDBA SEP-2024
AIMLF-DBA.2.M.A

Area Computer Science and AI

Number of sessions: 30

Academic year: 24-25

Degree course: SECOND

Number of credits: 6.0

Semester: 2º

Category: COMPULSORY

Language: English

Professor: **ALEJANDRO VACA SERRANO**

E-mail: avaca@faculty.ie.edu

CEO & Founder @ LenguajeNatural.AI

CEO and Founder - LenguajeNatural.AI (since Feb. 2024)

Data Scientist at Instituto de Ingeniería del Conocimiento (IIC) from 2019-2024

Deep Learning, Computer Vision & Natural Language Processor in different institutions since 2021.

? Google Scholar profile: shorturl.at/lnp58
??Best Poster Presentation Award at the NAACL 2022 LatinxAI Workshop.
??1st prize at the SomosNLP 2022 Hackathon for the BioMedIA project: https://huggingface.co/spaces/hackathon-pln-es/BioMedIA.
??Special award for project with the highest number of Likes for BioMedIA - Hackaton SomosNLP 2022.
??1st prize in Tasks 1 & 2 (2/2) at Exist22@Iberlef22, a challenge focused on detecting sexism in texts in English and Spanish.
???Main developer of RigoBERTa, the most capable Spanish language model, reaching a new State of the Art (best model in 10/13 tasks). Link to the paper: shorturl.at/ghjlW
??Best Data Scientist Special Award - Hackaton SpainAI 2021
??1st prize @ SpainAI 2021 Hackathon Time Series challenge.
??1st prize @ Computer Vision challenge of the SpainAI Hackathon 2021.
??3rd prize @ NLP challenge of the SpainAI Hackathon 2021.
??1st Prize @ Minsait Land Classification challenge - Cajamar UniversityHack2020.

Office Hours

Office hours will be on request. Please contact at:

LinkedIn: https://www.linkedin.com/in/alejandro-vaca-serrano/

e-mail: alejandro_vaca0@hotmail.com

SUBJECT DESCRIPTION

Machine Learning has been around for years but more recently Big Data is making it have increasingly more impact on our lives and day to day business activity. From email spam detection to assisting doctors identify life threatening illnesses, Machine Learning is a cross industry tool making it essential for any Data Scientist to have under their belt to succeed in their career path.

For a long time now, companies have gained a competitive edge through Business Intelligence, summarizing data from the marketplace and their internal organization. Dashboards have helped improve business decisions by looking to the past to see where they may have gone wrong or how they could improve. But what if a company could look into the future and make predictions? Or even better, what if they could look into the future, make predictions and then act to manipulate future events even happening at all?

Machine Learning is a subset of Artificial Intelligence. It can make predictions based on algorithms that learn from observational data. Humans do this all the time when making calculated expectations, we look to our past experiences and form a conclusion of what will probably happen. From wearables to self-driving cars, extracting meaningful and personalized insights from data is ever more in demand.

The general objective of this course is to understand the role of Machine Learning within the tasks of the data scientist or data analyst, and know how this is placed within the broader process of data mining. As well as a hands on approach to working with the algorithms you will also understand how an entire Machine Learning project works within a company setting and how a Machine Learning problem should be approached.

LEARNING OBJECTIVES

The goal of this course is to cover the fundamental techniques used in Machine Learning. These techniques can be classified broadly in supervised and unsupervised learning depending whether the algorithm is taught by looking at examples or not. Furthermore, the objective is not only to know the theory behind these techniques but, when and how to use them depending on a specific business question and the data available at hand.

In this course students will learn the end-to-end process of building a machine learning pipeline and will have a solid grasp of the theoretical and practical application of popular supervised and unsupervised machine learning algorithms. We will cover the following topics:

- Steps to follow to build an end-to-end Machine Learning project
- Exploratory Data Analysis: Handling, cleaning, and preparing data for ML algorithms
- Creating a predictive model by using labeled and unlabeled data
- Selecting and engineering features to make better predictive models
- Selecting a model and tuning hyperparameters using cross-validation
- Diagnose underfitting and overfitting (the bias/variance trade off)
- Most common supervised learning algorithms for regression and classification
- Reducing the dimensionality of datasets to fight the “curse of dimensionality”
- Other unsupervised learning techniques such as clustering or anomaly detection

TEACHING METHODOLOGY

IE University teaching method is defined by its collaborative, active, and applied nature. Students actively participate in the whole process to build their knowledge and sharpen their skills. Professor's main role is to lead and guide students to achieve the learning objectives of the course. This is done by engaging in a diverse range of teaching techniques and different types of learning activities such as the following:

Learning Activity	Weighting	Estimated time a student should dedicate to prepare for and participate in
Lectures	13.3 %	20.0 hours
Discussions	20.0 %	30.0 hours
Exercises in class, Asynchronous sessions, Field Work	13.3 %	20.0 hours
Group work	26.7 %	40.0 hours
Individual studying	26.7 %	40.0 hours
TOTAL	100.0 %	150.0 hours

AI POLICY

Generative artificial intelligence (GenAI) tools may be used in this course for research, code documentation, code generation and code fixing with appropriate acknowledgement. GenAI may not be used for exams, writing conclusions or extracting insights for the assignments. If a student is found to have used AI-generated content inappropriately, it will be considered academic misconduct, and the student might fail the respective assignment or the course.

If you are in doubt as to whether you are using GenAI tools appropriately in this course, I encourage you to discuss your situation with me.

Below, a suggested format to acknowledge the use of generative AI tools. Please note that acknowledging AI will not impact your grade.

I acknowledge the use of [AI systems link] to [specify how you used generative AI]. The prompts used include [list of prompts]. The output of these prompts was used to [explain how you used the outputs in your work]

If AI was permitted to use in your assignment, but you have chosen not to include any AI generated content, the following disclosure is recommended:

No content generated by AI technologies has been used in this assignment.

PROGRAM

The course kicks off with an introduction to machine learning and AI. Then, students will learn about supervised learning techniques, analyzing and implementing the main algorithms that solve regression and classification problems. This will be used to cover feature engineering, model validation and hyperparameter tuning.

The fundamentals of unsupervised learning will be covered analyzing and using different clustering techniques and their evaluation with a special focus on their application and interpretation. This will help the student to understand and apply the best technique based on the needs of the problem and the data available. Special emphasis is placed on the applications of each of these concepts to business analytics and data science bringing real world datasets and examples.

In this context, the course is divided into 5 modules, each module consists of 5 sessions.

- Module 1: Fundamentals of Machine Learning
- Module 2: Supervised learning: Regression
- Module 3: Supervised learning: Classification

- Module 4: Supervised learning: advanced modelling
- Module 5: Unsupervised learning

SUMMARY

Disclaimer: The following description of the material covered is tentative. An attempt will be made to cover all listed topics. However; the pace in the classes will depend on the group performance.

SESSION 1 (LIVE IN-PERSON)

Data Analytics, ML and AI:

In this session we will cover the fundamentals of machine learning, the difference between supervised and unsupervised learning, the data science cycle and we will make sure that our IDE is in place for the rest of the course.

SESSION 2 (LIVE IN-PERSON)

Linear Regression and beyond

During this session we will study the classical regression problem and see how we can improve it with more elaborated regression models beyond linear (polynomial regression, kernel smoothing, regression and smoothing splines, LOESS, Generalizes Linear Models and Generalized Additive Models).

SESSION 3 (LIVE IN-PERSON)

The concept of under and overfitting

During this session we will learn about two fundamental concepts in Machine learning, namely, underfitting and overfitting.

SESSION 4 (LIVE IN-PERSON)

Regression practice

A dataset will be given to the students to have it ready to apply regression and a prediction will be made applying all the content learned until now.

SESSION 5 (LIVE IN-PERSON)

Regularization techniques

The problem of overfitting can be curated using regularization techniques. In this session, we will study different regularization techniques. We will learn about ridge, lasso and elastic net regularization techniques, and they will be compared to the not regularized regression results.

SESSION 6 (LIVE IN-PERSON)

Regularization techniques practice

A dataset will be given to the students for them to do an end to end prediction project using regression. The purpose of this exercise is to check the understanding of the students on all the content covered up until now and as a wrapper of regression.

SESSION 7 (LIVE IN-PERSON)

Review/Questions/Inquiries.

SESSION 8 (LIVE IN-PERSON)

Introduction to classification

In this session classification problems will be introduced. We will introduce them by analyzing logistic regression for binary and multiclass classification.

SESSION 9 (LIVE IN-PERSON)

Training and evaluating classifiers

In this session, evaluation techniques will be explained and the main metrics applied in classification problems such as logit loss, accuracy, confusion matrix, sensitivity and specificity, AUC, ROC, etc. The way in which they are applied to different problems and the main constraints for each of them will be analysed.

SESSION 10 (LIVE IN-PERSON)

Classification Practice

A dataset will be given to students to solve a classification problem and evaluate the solution that has been achieved.

SESSION 11 (LIVE IN-PERSON)

Class imbalance

In this session, we will cover the class imbalance problem in detail and the potential solutions as well as when to apply each of them and how.

SESSION 12 (LIVE IN-PERSON)

Classification problem with class imbalance practice

A dataset will be given to the students for them to solve a classification problem with class imbalance.

SESSION 13 (LIVE IN-PERSON)

Review/Questions/Inquiries.

SESSION 14 (LIVE IN-PERSON)

Data Preprocessing, knowledge discovery process, data cleansing and pre-processing techniques

In this session, we will cover the fundamental techniques used to know the data, the knowledge discovery process, data cleaning, handling text and categorical variables, pre-processing techniques, transformers and pipelines.

SESSION 15 (LIVE IN-PERSON)

Data preparation and pre-processing pipeline exercise.

A dataset with raw data will be given to the students to work on it and go through the main operations required to have the data processed and ready to be used.

SESSION 16 (LIVE IN-PERSON)

Feature Engineering and selection

In this session, we will cover the fundamentals of model evaluation such as train/test splits, cross validation, resampling the variance bias trade-off. We will analyse the pros and cons of each technique and we will be covering the fundamentals of feature engineering and selection using different techniques, filters, wrapper, embedded.

SESSION 17 (LIVE IN-PERSON)

Feature selection exercise

Continuing with the first dataset, different preprocessing techniques will be applied. This exercise will be used to make sure that all libraries are in place and working and it will be the first contact with feature engineering techniques to see and compare the way they work.

SESSION 18 (LIVE IN-PERSON)

Decision trees, ensembles, bagging and boosting

In this session, we will introduce tree models, which will set the grounds for the rest of the advanced modelling techniques: ensembles, bagging and boosting.

SESSION 19 (LIVE IN-PERSON)

Random Forest and Gradient boosting

In this session, we will move from trees to random forests and gradient boosting. We will cover the fundamentals and applications of each of them and hyper-parameter tuning to improve the accuracy of the model. Hyper-parameter tuning can improve the final accuracy of the model compared to the vanilla one.

SESSION 20 (LIVE IN-PERSON)

Advanced modeling techniques practise

Students will be given a dataset to compare the different models, accuracy and the impact of hyper-parameter tuning.

SESSION 21 (LIVE IN-PERSON)

Support Vector Machines

In this session, we address support vector machines.

SESSION 22 (LIVE IN-PERSON)

Compare classifiers practice

A dataset will be given to the students so they can apply different classifiers that compete against each other. The goal is to find the best model.

SESSION 23 (LIVE IN-PERSON)

Review/Questions/Inquiries.

SESSION 24 (LIVE IN-PERSON)

Clustering

In this session different clustering techniques are covered.

SESSION 25 (LIVE IN-PERSON)

Group segmentation using clustering practise

In this session we will build an applied unsupervised learning solution using the clustering algorithms to perform group segmentation.

SESSION 26 (LIVE IN-PERSON)

Dimensionality reduction

In this session we will cover different dimensional reduction algorithms such as PCA, SVD and t-SNE.

SESSION 27 (LIVE IN-PERSON)

Dimensionality reduction practise

We will carry out different practise exercises using dimensional reduction techniques.

SESSION 28 (LIVE IN-PERSON)

Anomaly detection

In this session we will cover anomaly detection techniques and different applications.

SESSION 29 (LIVE IN-PERSON)

Project's presentations

SESSION 30 (LIVE IN-PERSON)

Final Exam.

EVALUATION CRITERIA

Your final grade in the course will be based on both individual and group work of different characteristics that will be weighted in the following way:

criteria	percentage	Learning Objectives	Comments
Class Participation	15 %		active engagement in class (asking/answering questions), participation in forums, attendance

Assignments	30 %		Assignments
Group Presentation	20 %		Quality and knowledge
Intermediate Tests	15 %		Mid term and individual test
Final Exam	20 %		Final exam

RE-SIT / RE-TAKE POLICY

A. Class participation and discussion [15%]

Class participation will be evaluated based on in-class participation, forums and attendance. The active participation will be evaluated using the following criteria:

- Quality (not quantity) of your participation in class discussion: The most important dimension of participation concerns what it is that you are saying. A high quality comment reveals depth of insight, rigorous use of case evidence, consistency of argument, and realism. Frequency refers to the attainment of a threshold quantity of contributions that is sufficient for making a reliable assessment of comment quality. The logic is simple: if contributions are too few, one cannot reliably assess the quality of your remarks. However, once threshold quantity has been achieved, simply increasing the number of times you talk does not automatically improve your evaluation. Beyond the threshold, it is the quality of your comments that must improve. In particular, one must be especially careful that in claiming more than a fair share of “airtime”, quality is not sacrificed for quantity. Finally, your attempts at participation should not be such that the instructor has to “go looking for you”. You should be attempting to get into the debate on a regular basis.

You might want to avoid being classified as one of the following types of students:

- Repeaters, i.e., students that, consciously or unconsciously, make comments that are really just repeats/rephrasing of what has already been said (by other students, or you). This wastes time and adds nothing to learning.
- Ramblers, i.e., students that take a lot of time to say simple things or they may tell long personal/professional stories, or they roam into topics that are not relevant, or simply make low-quality comments just to participate. They waste valuable time and prevent other students from being able to participate.
- Students that have been distracted (by social networks, etc.) or who have stopped paying attention and then, later on, when they realized they have missed a term or concept, they ask you about it.

B. Assignments [30%]

As part of the practical and hands-on evaluation of this course, students will be asked to work on datasets in assignments to be delivered. In these assignments, students will practice the knoweldge acquired during the previous classes and will cover previous concepts. These assignments are mainly individual work although a some group assignment can be requiered which will be explicitely indicated. Assignments have a hard deadline for submission. Any assignment submitted after such a deadline will be consider as a fail.

C. Group presentation [20%]

Random groups are created and students will select an individual topic related to applications of AI/ML in the "real world". The project is open, students are free to create a model, show it in class or research on a hot topic in AI. The grade will be based both on the presentation, student control of the topic and deliverables (model, application, etc).

D. Intermediate tests [15%]

One or more tests will be carried out in class where students will be asked to complete a set of multiple choice quizzes, or short practical exercises. These quizzes will help you assess your overall understanding of the subject being studied and identify any caveat in your learning.

E. Final exam [20%]

The final exam is split into theory questions (multiple choice, complete, etc) and a programming exercise in class.

BIBLIOGRAPHY

Recommended

- Andreas C. Müller, Sarah Guido. *Introduction to Machine Learning with Python.* O'Reilly. ISBN 9781449369415 (Digital)
- Aurélien Géron. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow..* Second. O'Reilly Media, Inc.. ISBN 9781492032649 (Digital)
- Benjamin Johnston , Aaron Jones , Christopher Kruger. (2019). *Applied Unsupervised Learning with Python.* First. Pack. ISBN 9781789952292 (Digital)
- Christopher M. Bishop. (2007). *Pattern Recognition and Machine Learning.* First. Springer-Verlag New York Inc.. ISBN 9780387310732 (Digital)
- Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong.. (2020). *Mathematics for Machine Learning..* First. Cambridge University Press. ISBN 9781108455145 (Digital)
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2017). *An Introduction to Statistical Learning with Applications in R.* Springer Science+Business Media New York. ISBN 9781461471370 (Digital)
- Sebastian Raschka, Vahid Mirjalili. (2019). *Python Machine Learning..* Third Edition. Packt Publishing. ISBN 9781789955750 (Digital)
- Giuseppe Bonaccorso. (2019). *Hands-On Unsupervised Learning with Python.* First. Packt. ISBN 9781789348279 (Digital)
- Trevor Hastie, Robert Tibshirani, Jerome Friedman. (2017). *The Elements of Statistical Learning.* Second. Springer. ISBN 9780387848570 (Digital)

BEHAVIOR RULES

Please, check the University's Code of Conduct here. The Program Director may provide further indications.

ATTENDANCE POLICY

Please, check the University's Attendance Policy here. The Program Director may provide further indications.

ETHICAL POLICY

Please, check the University's Ethics Code here. The Program Director may provide further indications.