Teaching Machine Learning in Argentina: the ClusterAI pipeline

Anonymous Authors[1]

Abstract

Teaching machine learning has been a growing activity in almost any educational establishment. Despite the high availability on study materials, Latin America region has seen a lack of educational programs focused on machine learning. Additionally the majority of educational materials are available only in English. In this work we propose the ClusterAI pipeline based on a curated list of topics in Spanish and a collaboration with the Buenos Aires city government that open public data-sets that let students to apply machine learning models on real data.

1. Introduction

The last decade has experienced a bloom in the machine learning and data science fields propelled mainly by the improvement of new processors, data availability and new statistical learning algorithms. Additionally the availability of thousands of new papers, open course materials and new platforms to run and implement learning models allow the access of machine learning content to a wide audience of practitioners, researchers and enthusiasts. Despite that the highly availability of materials to learn machine learning has been increasing during the last years there is a lack of content in Spanish for the Latin American region. Multiple meetings and workshops actions have been taken in the South America region such as Khipu (khipu, 2019), the Sao Paulo Advanced School on Learning from Data (spsas, 2019) and the Machine Learning Summer School 2018 (MLSS2018, 2018), nevertheless these events were designed mainly for a semi-senior audience and attendees with at least 1 or 2 years of experience in the machine learning field such as PhD students.
Additionally we found that there are not enough open courses designed for undergraduate students with an initial and basic training in algebra and statistics such as engineering, biology, social sciences, economics and design students to let them their first steps in Machine Learning techniques and applications. In this work we present the ClusterAI pipeline held at the Universidad Tecnologica Nacional Buenos Aires, an free and open source Machine Learning program designed for last year STEM and Social sciences students. The course has multiple objectives: training about the statistical learning approach, use computational tools to run the machine learning methods, use real data-sets from Buenos Aires city data portal and presentation of results to a wide audience.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Teaching Machine Learning Workshop at ECML 2021. Do not distribute.

2. Course Requirements

From a computational point of view the course does not requires students previous programming skills. To help students to deal with the first steps a crash-course workshop has been designed to introduce students to the python and Jupyter notebook framework before getting in the proposed ClusterAI course where libraries such as Numpy, Pandas, Matplotlib and Scipy are introduced for the first steps.
From a theoretical point of view the course assumes a student has a basic knowledge in algebra, statistics and calculus. It is assumed that the incoming students understand the concept of matrix, vector, random variables and probability density functions.

2.1. Student profile

The ClusterAI pipeline initially started as a machine learning course in the Industrial Engineering degree at Universidad Tecnologica Nacional of Buenos Aires for last-year undergrad students. Nevertheless the program has been opened to students from other disciplines such as electronics engineering, computer engineering, biology, economics and political science. Besides any kind of heterogeneity, most of the students share lack of formal training in any programming language nor advanced analytics.

3. The ClusterAI pipeline: Contents of the course and learning path

The course is divided in 7 chapters. We base the idea of making each chapter as a workshop and the full course a sequence of workshops where on each one we study a specific

topic. The first six chapters of the course are: exploratory data analysis, supervised learning, unsupervised learning, dimensionality reduction, introduction to natural language processing and neural networks.

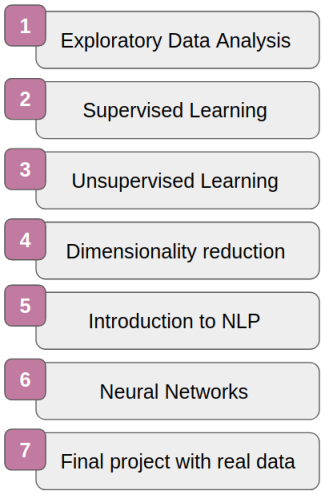

Figure 1. Pipeline of topics covered during the CluterAI course.

Then the chapter 07 corresponds to the final project presented by the students and described in 4 and 6 of this work. The bibliography used are papers corresponding to each topic and the following books: Elements of statistical learning (Friedman et al., 2001), Machine Learning and Pattern Recognition (Bishop, 2006) and the Deep Learning Book (Goodfellow et al., 2016) among others. The following subsections details the first six chapters of the course.

3.1. Exploratory Data Analysis

The exploratory data analysis (EDA) part is the initial stage of the course and the objective is to teach students how to handle, pre-process, explore, visualize and describe tabular data. Despite Machine Learning is commonly applied on different data modalities such as Natural Language, Images, Time Series and Graphs we decide to start with the most common and easy to work data format: tabular data.
The EDA part is composed by three chapters of the course. The first three chapters are more technical than theoretical since the objective is to introduce the students to the Python stack for data analysis. We assume the learning curve of a student to get used with Numpy, Pandas, Matplotlib and Scipy can take three to four weeks. During these chapters students learn to use Jupyter Notebooks and make exploratory data analysis on tabular data, mainly .CVS files. The exploratory analysis aims to explain multiple visualization and analytical approaches such as heat-maps, box-plots, scatter-plots, bar-plots and histograms among others to describe visually the data-set.
To handle tabular data Pandas and Numpy libraries are used to filter, concatenate, merge and process data tables. Additionally pre-processing tasks such as adding dummy variables and cleaning or imputing NaN values are explained. Finally feature pre-processing methods like Auto-scaling and standardization (van den Berg et al., 2006) and Min-Max normalization (Jain & Bhandare, 2011) are presented to transform real-valued and integer features.
All the previous concepts are explained from a theoretical point of view and simultaneously by applying with coding exercises on real data from the Buenos Aires city government open data portal such as the (GCBA, 2019).

3.2. Supervised Learning

Once the exploratory data analysis part is covered the first Machine Learning concept to teach is the Supervised Learning approach. This part has three chapters about statistical learning theory and applications where classification and regression learning tasks are covered.
As an introduction to supervised learning first the concept of how samples are drown from a d-dimensional space $\mathcal{X} \in R^d$ where d is the number of dimensions or features and the multivariate approach is needed to learn from high dimensional data.
Then supervised learning is explained by first introducing a hypothesis space of potential functions to be used in the learning task and the role of the loss function to pick the best one within the available space.
After defining the differences between regression and classification tasks linear and nonlinear models are explained and how these play a role in the required complexity of the model. Linear decision boundaries are studied for binary and multi-class classification problems along with the variance versus bias concept and how this impact in the model selection task. To deal with the model selection step cross validation, grid search and model evaluation are studied.
The classifiers used in the course are support vector machines, k-nearest neigh boors, logistic regression and random forests. For regression tasks linear regression, polynomial regression, support vector regression and k-nearest neigh boors regression are studied.
To study the non-linear case of the support vector machines Kernel Methods are presented.

3.3. Unsupervised Learning

The unsupervised learning part is based mainly on the clustering and community detection methods. The idea on this part is to explain the similarity between sample vectors concept and how different measures of similarity can be used to understand if a pair of samples are similar or not. Clusters are explained and downstream analysis studied once the segmentation of samples is obtained.
Two popular clustering methods, K-means and Hierarchical

clustering, are explained and applied on Buenos Aires subway stations or house prices in Buenos Aires City data-sets. Evaluation metrics such as Silhouette Index (Brun et al., 2007) and Rand Index (Hubert & Arabie, 1985) are studied to let the student decide which number of clusters is the best for a given problem.
Despite dimension reduction can be included within unsupervised learning we considered to studied it on a separate chapter since it can be used for both supervised and unsupervised downstream learning tasks.

3.4. Dimensionality Reduction

In the dimensionality reduction chapter linear and nonlinear unsupervised approaches are studied.
The first concept to introduce is the curse of dimensionality and how the sample to feature ratio is and important aspect to analyze before implementing a machine learning task. Additionally a section explaining the differences between feature selection and feature extraction methods is included. The two method studied in this chapter are Principal Component Analysis, Kernel Principal Component analysis (kPCA) and T-distributed Stochastic Neighbor Embedding (t-SNE). The idea for this chapter is to introduce the students to visualize high dimensional data in two dimensions, to reduce the complexity of a learning task due to the dimension reduction and to improve the sample to feature ratio to avoid the curse of dimensionality.

3.5. Introduction to Natural Language Processing

The idea of the NLP chapter is to make a brief introduction to other types of data formats beyond the tabular data case such as Natural language. In this chapter simple and introductory techniques such as Tokenization, Bag of Words and the TF-IDF are studied and applied on toy and real data-sets. We encourage students to build their own data-sets with real data by taking more than 200 headlines of two newspapers from Argentina: Clarin and Pagina12. Both newspapers are known to be ambassadors of two extreme opposite political parties thus their economic headlines tend to encode signal according to two classes. They are encouraged to build the data-sets with these newspapers and learn a low dimensional representation of the economic headlines. Then a supervised and unsupervised approaches are used to analyze how headers from each newspaper tend to group.
Another application is the pre-processing and classification of positive and negative movie reviews, which is a popular and common teaching example for NLP tasks.

3.6. Neural Networks

The neural network chapter is the last one before the student project part. To make easier the first steps for the students neural networks are explained to be trained only for tabular data since convolutions are more complex and out of the scope for the students in this initial stage.
Before studying neural nets the Perceptron model is presented followed by different activation functions. Then the multilayer perceptron model is presented and how different architectures such as number of hidden layers or number of neurons per layer. Then loss functions such as Mean Squared Error and Cross Entropy are studied in addition to the concept of local minimum in the loss landscape function. Additionally the gradient descent and backpropagation algorithm are explained. Finally it is explained how to improve the neural network training by regularization terms, reducing learning rate on plateau, Dropout and Batch Normalization. With the explained concepts students are encouraged to train neural networks on simple and tabular labeled data-sets such as the Wisconsin Breast Cancer or the Iris data-set and perform classification. The idea is to let students to train neural networks on simple problems to understand how all the hyper-parameters affect the results.
Once basic neural network implementations for supervised learning such as classification problems are introduced then the Autoencoder model is studied as an unsupervised nonlinear dimension reduction method. This part students benchmark low dimensional visualization tasks of the autoencoder against the PCA and Kernel-PCA. The application case involving real data is the gene expression cancer dataset from the International Cancer Genome Consortium to let students learn low dimensional latent space of tumors and perform supervised or unsupervised downstream tasks.
Finally to show the potential of neural networks to deal with time-series and sequences the Recurrent Neural Network model is studied with an application of time-series classification from signals measured in engines where each one has a different manufacturing setup.
This topic concludes the methodology and theoretical aspects of the course. Students are encouraged to use the presented methods on real data-sets from the Buenos Aires city open data portal presented in the next section.

4. Student Projects

One of the main goals for the students in this course is to develop from the ground up a applied Machine Learning based project aiming for solving real problems or getting new insights of an existing situation.
By making groups of three students each group has the objective to pick a data-set from the Open Data Portal from the Buenos Aires city government (GCBA, 2019) to make an exploratory data analysis and implement a supervised or unsupervised learning approach to discover insights from the selected data-set. All the implementations are required to be done in a Python framework as explained in section 5. Additionally, each group has assigned one alumni student serving as a mentor and helping them to get along with the

objective to ensure a consistent result to be shared at the end of the course. During the course, students have different checkpoints with professors and mentors on how to choose a profitable data-set, working strategies, best suited algorithms and other technical issues.
The final delivery is divided into 3 pillars: first, a Jupyter Notebook with the development of the project, second, a technical report explaining the root of the problem they were trying to solve, how they overcome all difficulties and the conclusions and last but not least a research poster that is presented to the public in an open event at the university where not only members of the university attend but also people from different backgrounds and disciplines.
With this methodology, students are exposed to a holistic view, from understanding the problem, building the pipeline, overcoming technical issues and presenting results in a open Data Science fair with +100 participants from different backgrounds.

5. Technologies used

The ClusterAI pipeline is a Python-based class. All classroom exercises and explanations are coded in Jupyter notebooks (Pérez & Granger, 2007), that serve for multiple purposes: explaining and visualizing theoretical concepts with toy data, solving simple exercises with popular datasets such as the Wisconsin Breast Cancer or Iris dataset and coding implementations on real data as case studies class by class.
Sklearn and Tensorflow Keras are the main libraries to implement ML algorithms during the course as well as Pandas, Numpy (Harris et al., 2020) and Matplotlib for data wrangling and visualization in are used.
Additionally all workshop material is stored in a public repository in Github (clusterai, 2018) where notebooks and scripts are built and published by professors in collaboration with mentors. Each year new content is updated and created.
For communication purposes between mentors, professors and students the course uses Slack channels (slack, 2021). The tool is useful to communicate the schedule for each class, upload recorded classes and complementary content, create polls and share related news, posts, blogs, etc. During the COVID-19 pandemic the channel Moreover resulted useful to replace the physical room allowing students to share content and decentralize the flow of information.
As additional content, there is a YouTube channel where the intention is to upload tutorials on common questions among the students, for example, installing Python and Anaconda (youtube, 2021).

6. Alliance with Buenos Aires Government

As explained in section 5 the last stage of the ClusterAI pipeline is to request students an machine learning application project divided in two parts: an exploratory data analysis and a machine learning application. By an alliance with the Buenos Aires City Government the datasets used are obtained from the Buenos Aires Open Data portal `data.buenosaires.gob.ar` (BA Data). The BA DATA portal is open sourced and its data-sets record approximately 60.000 downloads per month. It showcases 421 data-sets from 31 different organizations within the Buenos Aires City Government related to 12 central government initiatives: Public Administration, Culture and Tourism, Human Development, Economy and Finances, Education, Gender, Environment, Transportation, Health, Security, Urbanism and territory and COVID-19. Since December 2019, the Undersecretary of Evidence-Based Public Policies is the central team responsible for data and open data management and manages the platform. BA DATA is Buenos Aires City's open data portal, where public data-sets are generated, saved and published. It's goal is to strengthen the city government's transparency, encourage citizenship participation, promote data reuse and facilitate innovation from data.
Since 2018 the ClusterAI students have done more than 50 projects using datasets from BA data. Some published projects done by the students include Regresion models applied to familiar violence estimation (Bellini, 2019), detection of financial behaviour using machine learning models (Weigandi, 2019), prediction of the type of vegetation (Tettamanti, 2019), commercial opportunities map (Libertun, 2019), quality air analysis via machine learning models (Cavallucci, 2019) and car robbery analysis (Carpaneto, 2019) among others.

7. Conclusions

In this work we present the open ClusterAI pipeline used to teach machine learning to undergraduate students in Argentina. The proposed pipeline is composed by seven chapters where the first six are dedicated to theory and applications of multiple machine learning methods while the last chapter is focused on student projects using the Buenos Aires open data portal. Student projects include a wide range of applications using real data from city sensors. Finally students are encouraged to present via poster sessions, technical reports and github repositories the projects in order to promote open developments and to encourage the local community to use public data. Despite the clusterAI pipeline has been designed by engineering students it has also been validated with students from other disciplines such as social science, economics and biology.

References

Bellini. Family violence estimation, 2019. URL https://github.com/MartinBellini/

Bellini-Grupo14.

Bishop, C. M. Pattern recognition. *Machine learning*, 128 (9), 2006.

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., and Dougherty, E. R. Model-based evaluation of clustering validation measures. *Pattern recognition*, 40 (3):807–824, 2007.

Carpaneto. Roberry car analysis, 2019. URL https://github.com/ilcarbo/data_science_clusterai_carpaneto.

Cavallucci. Air quality analysis, 2019. URL https://github.com/Aldicavallucci/data_science_clusterai_cavallucci.

clusterai. Clusterai machine learning course, 2018. URL https://github.com/clusterai/clusterai_2021.

Friedman, J., Hastie, T., Tibshirani, R., et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

GCBA. Buenos aires data, 2019. URL http://https://data.buenosaires.gob.ar/.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Jain, Y. K. and Bhandare, S. K. Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology*, 2(8):45–50, 2011.

khipu. Khipu, 2019. URL https://khipu.ai/.

Libertun. Commercial opportunity map, 2019. URL https://github.com/FedericoLibertun/TP-Mapa-de-Oportunidades-Comerciales.

MLSS2018. Machine learning summer school 2019, 2018. URL https://mlss2018.net.ar/.

Pérez, F. and Granger, B. E. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.53. URL https://ipython.org.

slack. Slack technologies, 2021. URL https://slack.com/.

spsas. Sao paulo school of advanced science on learning from data, 2019. URL https://sites.usp.br/datascience/spsas-learning-from-data/.

Tettamanti. Plant type prediction, 2019. URL https://github.com/atettamanti/TP-grupo-0.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., and van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7(1): 1–15, 2006.

Weigandi. Financial behaviour, 2019. URL https://github.com/iweigandi/data_science_clusterai_weigandi.

youtube. Youtube clusterai course, 2021. URL https://www.youtube.com/channel/UC3AG3wNJDoMbWpwEGtNO0wQ/featured.