U3 MACHINE LEARNING ALGORITHMS

U3.E3 SEMI-SUPERVISED MODELS

Machine Learning Engineer

January 2021, Version 1

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

LEARNING OBJECTIVES

The student is able to

MLE.U3.E3.PC1	Define and understand semi-supervised learning.
MLE.U3.E3.PC2	Know some of the most commonly used algorithms in semi-supervised learning.
MLE.U3.E3.PC3	Know the domain application of semi-supervised models.

Is it possible to improve the quality of learning by combining labeled and unlabeled data?

- There is usually a lot more unlabeled data available than labeled.
- Semi-Supervised Learning (SSL) is a mixture between supervised and unsupervised approaches.
- In addition to unlabeled data, the algorithm is provided with some supervision information – but not necessarily for all examples.

DRIVES
Development and Research on Innovative Vocational Education Skills

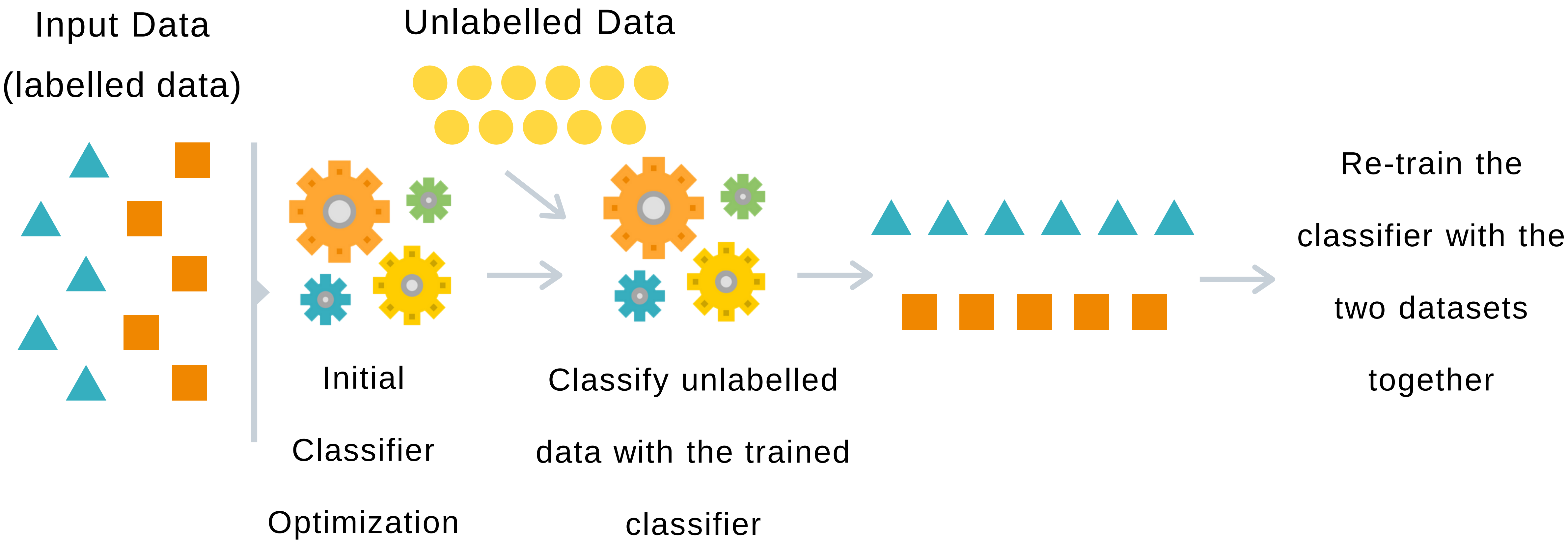
As the name suggests, SSL is halfway between supervised and unsupervised learning.

The dataset consists of a combination of labelled and unlabelled data

First, similar data is grouped using an unsupervised learning algorithm and only then the supervised learning algorithms are applied

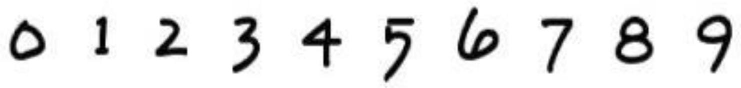
It can refer to either transductive or inductive learning

SEMI-SUPERVISED LEARNING



SEMI-SUPERVISED LEARNING

Labeled vs. Unlabeled Data



“0” “1” “2” “3” “4” “5” “6” “7” “8” “9”

Human Expert / Special Equipment

“dog” “cat “dog”

Unlabeled data, X_i

Labeled data, Y_i

Cheap and Abundant

Expensive and Scarce

DRIVES
Development and Research on Innovative Vocational Education Skills

Feature Space *X*

Label Space *Y*

GOAL: Construct a predictor $f: X \to Y$ to minimize $R(f) \equiv \mathbb{E}_{XY}[loss(Y, f(X))]$

An optimal predictor (Bayes Rule) depends on unknown P_{XY}. So, instead *learn* a good prediction rule from training data $\{(X_i, Y_i)\}_{i=1}^{n} \overset{\text{iid}}{\sim} P_{XY}(unknown)$

Training Data $\{(X_i, Y_i)\}_{i=1}^{n}$ *labeled* → Learning Algorithm → Prediction Rule $\hat{f}_n$

DRIVES
Development and Research on Innovative Vocational Education Skills

Training Data

$\{(X_i, Y_i)\}_{i=1}^n$

$\{X_i\}_{i=1}^m$

Learning Algorithm

Prediction Rule

$\hat{f}_{n,m}$

Supervised Learning

Labeled Data $\{(X_i, Y_i)\}_{i=1}^n$

"dog"

X_i Y_i

Semi-Supervised Learning

Labeled Data $\{(X_i, Y_i)\}_{i=1}^n$ and Unlabeled Data $\{X_i\}_{i=1}^m$

$m \gg n$

Goal: Learn a better prediction rule than based on labeled data alone.

COGNITIVE SCIENCE

Computational modal of how humans learn from labeled and unlabeled data:

- Concept learning in children: x=animal, y=concept (e.g. cat)
- Dad/Mum points to an animal and says “this is a cat”
- Children also observe animals by themselves

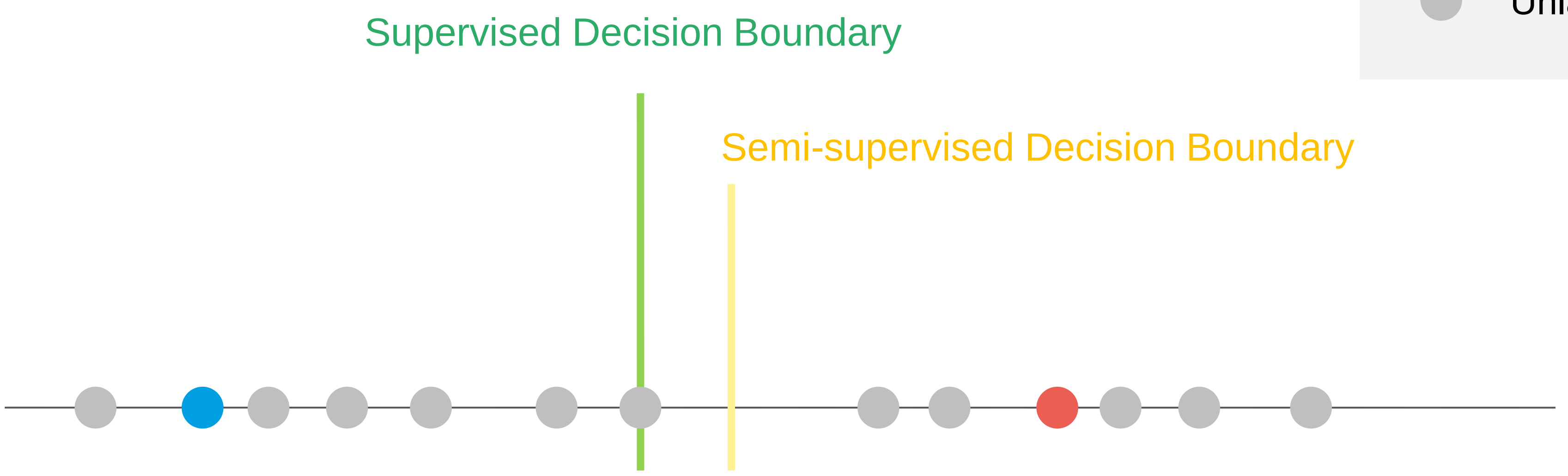
CAN UNLABELED DATA HELP?

“Similar” data points have “similar” labels

Positive labeled data

Negative labeled data

Unlabeled data



- Assume each class is a coherent group (e.g. Gaussian)
- Then, unlabeled data can help identify the boundary more accurately

DRIVES
Development and Research on Innovative Vocational Education Skills

WHY SHOULD WE USE SSL?

- Labeling data is usually expensive, specially when a significant number of cases are being addressed;

  I have a good idea, but I can't afford to label lots of data!

- Even when labeled data are available in significant amounts, it is useful to use more data to introduce as much variability as possible in the studies carried out;

  I have lots of labeled data, but I have even more unlabeled data available!

- Domain adaptation.

  I have labeled data from a domain, but I want a model for a different domain!

SSL ALGORITHMS

- Self-Training
- Generative models
- Mixture models
- Graph-based models
- Co-Training
- Semi-supervised SVM or Transductive Support Vector Machine (TSVM)
- …

Transduction refers to reasoning from observed specific (training) cases to specific (test) cases. In contrast, **induction** means reasoning from observed training cases to general rules, which are then applied to the test cases.

Hence, **Inductive Learning** can be seen as the same as traditional supervised learning, in which a machine learning model is built and trained based on the labeled training data. Then, the trained machine learning model is used to predict the labels of the testing data, which the model has never seen before.

Transductive Learning techniques, on the contrary, observe all the data, both the training and testing datasets, beforehand. Here the model learns from the already observed training dataset and then predicts the labels of the testing dataset. Although the labels of the testing datasets are unknown, the model can make use of the patterns and additional information present in these data during the learning process.

TRANSDUCTIVE LEARNING VS INDUCTIVE LEARNING

DRIVES
Development and Research on Innovative Vocational Education Skills

Transductive learning	Inductive learning
Can only predict the points encountered in the test set based on the observed training set.	Can predict any point in the space of points beyond the unlabeled points.
Does not build a predictive model. If a new unlabeled data point is added to the test data, we will have to re-run the algorithm from the beginning.	Builds a predictive model. If a new unlabeled data point is added to the test data, we can use the initially built model.
More computational cost	Less computational cost

HOW SEMI-SUPERVISED LEARNING WORKS

- Most approaches make strong model assumptions (guesses).

- Some commonly used assumptions:
  - Clusters of data are from the same class
  - Data can be represented as a mixture of parameterized distributions
  - Decision boundaries should go through non-dense areas of the data
  - Model should be as simple as possible

- Semi-Supervised Learning is a mixture between supervised and unsupervised approaches.

- In SSL, the dataset consists of a combination of labeled and unlabeled data. The similar data is grouped using unsupervised learning algorithms and then, supervised learning algorithms are applied.

- SSL can be either transductive or inductive learning. The main difference between these types of SSL is the fact that transductive learning uses the test data to train the model whereas inductive learning only uses the training data and then applies the model to predict unseen instances.

REFERENCE TO AUTHORS

Diana Ferreira

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

[0000-0003-2326-2153](0000-0003-2326-2153)

Regina Sousa

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

[0000-0002-2988-196X](0000-0002-2988-196X)

José Machado

- Associate Professor with Habilitation at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

[0000-0003-4121-6169](0000-0003-4121-6169)

REFERENCE TO AUTHORS

António Abelha
- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0001-6457-0756

Victor Alves
- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0003-1819-7051

REFERENCE TO AUTHORS

This Training Material has been certified according to the rules of **ECQA – European Certification and Qualification Association.**

The Training Material was developed within the international job role committee **“Machine Learning Engineer”:**

UMINHO – University of Minho (https://www.uminho.pt/PT)

The development of the training material was partly funded by the EU under Blueprint Project DRIVES.

REFERENCES

Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, *20*(3), 542-542.

Zhu, X. J. (2005). Semi-supervised learning literature survey.

Zhou, X., & Belkin, M. (2014). Semi-supervised learning. In *Academic Press Library in Signal Processing* (Vol. 1, pp. 1239-1269). Elsevier.

Zhu, X., Lafferty, J., & Rosenfeld, R. (2005). *Semi-supervised learning with graphs* (Doctoral dissertation, Carnegie Mellon University, language technologies institute, school of computer science).

DRIVES

Development and Research on Innovative Vocational Education Skills

Thank you for your attention

DRIVES project is project under [The Blueprint for Sectoral Cooperation on Skills in Automotive Sector](), as part of New Skills Agenda.

The aim of the Blueprint is **to support an overall sectoral strategy and to develop concrete actions to address short and medium term skills needs.**

Follow DRIVES project at:

More information at:

www.project-drives.eu

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.