Machine Learning - F8203B040

Master Degree in Biostatistics

Mirko Cesarini Stefano Peluso
mirko.cesarini@unimib.it stefano.peluso@unimib.it

University of Milan Bicocca

Lesson 1

Presentation

- Who am I?
  - Mirko Cesarini
  - mirko.cesarini@unimib.it
  - Office Tel. (+39) 02 6448 5849
- What about me?
  - Researcher and Professor at University of Milan Bicocca
  - Machine Learning Enthusiast
  - ...
- Some Machine Learning projects where I was involved
  - Real-time Labour Market Information on Skill Requirements. Cedefop Europa Scientific Paper
  - Italian Labour Market Digital Monitor. UniMiB Spin-off, WollyBI
  - ...

Program

- Statistical methods for machine learning
    - Supervised and unsupervised learning
    - Recall to regression analysis
    - Classification analysis
    - Cross validation and bootstrap
    - Model selection and regularization
    - Beyond linear models
    - Tree-based methods
    - Support vector machines
- Feature Engineering and ML Tuning
    - Feature Engineering and Selection
    - Data Observability and Model existence issues
    - Hyperparameters optimization
- Artificial Neural Networks and Deep Learning
    - Artificial Neural Networks (ANNs), ANN types (feed forward, recurrent, convolutional, ...), ANN Training (Gradient Descent)
    - Deep learning
    - Industrial applications and open research issues

What is ML - Introductory Video (Moley Kitchen)

- Is it real or fake? It is real
- Youtube Video, credits: www.moley.com

What is Machine Learning (ML)?

- What is your answer? Open discussion
- Machine Learning definition: an computer program is said to learn from . . .
  - **experience E**,
  - with respect to some class of **tasks T**, as evaluated by
  - the **performance measure P**,

  if its performance for tasks T, as measured by P, improves with experience E.
- What is learning?
  - "Learning is any process by which a system improves performance from experience." Herbert Simon
  - "Learning is constructing or modifying representations of what is being experienced." - Ryszard Michalski
- Machine Learning is a subfield of Artificial Intelligence (AI)

Why Machine Learning is getting so Popular?

- Getting computers to program themselves i.e., the machine learns from examples, rather than being explicitly programmed for a particular outcome
  - We humans know more than we can tell: we can't explain exactly how we're able to do a lot of things...
  - Prior to ML, this inability to articulate our own knowledge meant that we couldn't automate many tasks. Now we can! [1]
  - ML systems are often excellent learners. They can achieve superhuman performance in a wide range of activities e.g., detecting fraud, recognising faces, and diagnosing diseases
- ML algorithms can learn relationships and models from data and examples during the training phase. If the learnt models work, then, they can be investigated
- Tackling (big) data and complex scenarios

[1] Brynjolfsson. The Business of Artificial Intelligence. Harvard Business Review. July 2017

Human in the Loop supporting Data Science

- Considerations
  - Humans really are very good at finding patterns and noticing odd things
  - Computers are really good at doing repetitive work and working on a large scale
  - The vice versa doesn't hold
- Machine learning can complement what an analyst can do. So, human smart and pattern-recognition abilities can be applied on a big data scale
- Interactive machine learning (iML)
  - Algorithms that can interact with agents and can optimize their learning behaviour through these interactions, where the agents can also be human
  - This *human-in-the-loop* can be beneficial in solving computationally hard problems, where human expertise can help to reduce an exponential search space through heuristic selection of samples

Not only Managing Large Datasets

- Chris McCubbin Talk at BSides Boston 2016.
    - ... There were problems in speech and text recognition and natural language processing that were stagnant and that had been stagnant for a very long time, around 15 years ...
    - ... Then people applied new spins on (neural network) machine learning techniques, and there was a huge increase in accuracy and potential use of these things in other applications.
    - ... This reawakened a focus on neural networks and machine learning in general.
    - ... Things like Siri and speech recognition on a phone are being used by everybody now because it's feasible. You don't need a supercomputer anymore.
    - In terms of the future, **we have barely scratched the surface of these things**.
- (Maybe) the same will be said in few years about ML and ... (you can choose what write here)

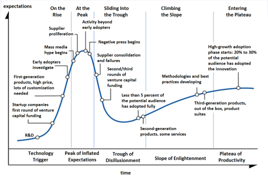
A Few Quotes (when the Hype was rising)

- "A breakthrough in machine learning would be worth ten Microsofts" (Bill Gates, Microsoft co-founder)
- "Machine learning is the next Internet" (Tony Tether, Director, DARPA)
- "Machine learning is the hot new thing" (John Hennessy, President, Stanford University)
- "Web rankings today are mostly a matter of machine learning" (Prabhakar Raghavan, Dir. Research, Yahoo)
- "Machine learning is going to result in a real revolution" (Greg Papadopoulos, CTO, Sun)
- "Machine learning is today's discontinuity" (Jerry Yang, CEO, Yahoo)
- ...[2]

[2] Credits: Pedro Domingos, University of Washington

Machine Learning/AI Winter?

- Previous quotes came from 2016 or earlier. What about now?
- A very pessimistic article wrote in 2018



Hype Cycle

- Self-driving-cars
  - In February 2018, Elon Musk . . . when asked about the coast to coast drive:
  - “We could have done the coast-to-coast drive, but it would have required too much specialized code to effectively game it . . . it would work for one particular route, but not the general solution. . . . which is not really a true solution . . . "
- In my (very personal) opinion
  - Some claims were excessive e.g., we will have self driving cars by 2020, we won’t need any more (medical) radiologists, . . .
  - Current estimates in Silicon Valley is that AI and Machine Learning could reduce costs or improve revenues of about 10%. That’s a great deal anyway!

Starting Example: Classification for supporting Decision Making

- Let's classify the following customers into k categories: e.g., Penurious, Occasional, Good ... (further details not provided)

CustomerID	Annual Purchase			Store Visits		
	Year 1	...	Year N	Year 1	...	Year N
...	...	...	...	...	...	...

- Let's discretise
  - *Annual Purchase* into **8 bins** [a]: ... (details not provided)
  - *Store Visits* into **10 bins**: ... (details not provided)
- Let's limit the scope to **3 years**. We have to map an input space of $(\mathbf{10} \times \mathbf{8})^{\mathbf{3}} = \mathbf{512\ 000}$ possible values into k possible categories. How?
- ML approach
  - An algorithm learns the classification criteria from some labelled examples. Then, it can later classify unlabelled data
  - Shortly: Machine Learning classification helps tackling *complexity* in *large datasets*.

[a] Bins are consecutive, non-overlapping intervals of a variable

Using Models

- We have just seen an example of using models to manage data, specifically to reduce data complexity
- Model: a simplified representation of a concept, phenomenon, system, or an aspect of the real world focusing on those features that are of primary importance to the model maker's purpose.
  - Different point of views, goals, ... different models
  - All models are wrong, but some are useful[3]
- Machine Learning is about (automatically) learning models
- Models are not exclusive to Machine Learning
  - E.g. a room temperature is detected every millisecond for 24h.
  - We have a dataset of $1000 \times 60 \times 60 \times 24 = 86\,400\,000$ observations
  - Let's average over the hour
    - It helps reducing the dataset size
    - Maybe it doesn't affect our analysis

[3] Box G.E.P., Draper N.R. (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley.

What can/can't be achieved by Machine Learning?

- Consider a bingo draw.
  Let's suppose you succeed in training a classifier that can predict the drawn numbers. What should you do?
  - You'd better go to the police.
  - A ML classification algorithm can learn a classification model from data
  - If the ML can learn a model (that works), it means that the number extraction is not random i.e., someone is cheating
- Take home message: if the model doesn't exist, the classifier can learn nothing

Question

- Company A sells Internet contracts by telephone.
  - It periodically collects information about new prospect clients e.g., name, age, location, existing contract price, existing contract technical features (bandwidth)
  - Sales representatives try to contact prospect clients by phone to sell new contracts
  - The company would like to develop a ML algorithm that can predict whether the customer will buy or not (to prioritise calls)
  - Training is based on historical data of selling attempts. Several years data is available, including the outcomes (success or failure)
- Company B sells natural gas and tries to develop a new business: to sell furnaces (caldaie)
  - ...same as above. Customers are contacted by telephone to sell new furnaces
  - ML is used to predict customer behaviour
  - Training is based on historical data of selling attempts
- Only one company is successful, which one in your opinion?

Observability

- Company A was successfull. They had all the information to estimate the customer decision making process:
    - Quality of service. The *Location* provides information about the performances of both the actual provider and the new proposed one
    - New and old fees
    - Demographic (e.g., age)
- Company B had very bad results
    - After interviewing some people, the management realized that old-furnace-owners only are akin to accept the proposal (i.e., furnace age $\geq$ 8 years old) .
    - Unfortunately, no data about furnace age is available
- To summarise: Company B faces a lack-of-data issue
- This issue can be framed in the general problem of *model observability* i.e., the ability to guess a model from the available data

Learning vs Explicitly Coding

-
- I have to evaluate some decision criteria over a huge dataset . . .
- Question: Better to use ML or let a pool of experts investigate the data and explicitly code an algorithm?
- Long answer
  - As long as the variables are few, it is simpler to explicitly code an algorithm
  - The more are the variables, the increasingly (exponentially) high will be the algorithm complexity
  - Even the smartest guy can fail to identify properly all the decision criteria

Video: Rethink Robot - Sawyer

- Credits: http://www.rethinkrobotics.com/sawyer/
- Youtube Video

Suggestion

- How can I make decision about whether to use ML or not?
- How complex can be a model (to be learnt with a reasonable effort)?
- Rule of thumb: If a **mental task** takes **less than one second** of thought to a typical person, we can probably automate it using AI either now or in the near future [4]
- What can ML learn?
  - ML can learn very complex models (e.g., consider the self driving cars), provided that a model exists!
  - The real challenge is preparing a suitable training set. High quality labelling a lot of examples can be very expensive

[4] Andrew Ng, What Artificial Intelligence Can and Can't Do Right Now. Harvard Business Review

Machine Learning Tasks

- Machine Learning (reminder): a computer program (CP) which performs a Task (T) where Performances (P) improve with experience(E).
- Which kind of task are related to Machine Learning?
  - Prediction
    - Classification (or Categorization)
    - Regression
  - Clustering
  - Planning (we won't touch this topic)
  - (...)
- What have in common these tasks? Which differences?
- Let's introduce some concepts before answering
  - Supervised vs Unsupervised learning
  - Continuous vs Discrete Variables

Terminology Introduction

- Consider the following classification problem

X (Input)		**Y** (Output)
N. of Calls	**Avg Duration**	**Client Type**
7	10	Good
2	2	Bad
...	...	...

- Terminology
  - X values/variables can be called: input, covariate(s), or (simply) **X**
  - Y values/variables can called: output, target, label, or (simply) **Y**

- A model can be described as a function $Y = f(X)$ which can predict the Y given the X

Learning

- Problem
  - Given two variables X and Y (e.g., the X and Y in the previous slide)
  - we need to identify a function h() so that
  - $Y = h(X)$

$$\mathrm{Y} \longleftarrow \boxed{h()} \longleftarrow \mathrm{X}$$

- Learning/training: the process of guessing $h()$ from the X and Y data

Supervised vs Unsupervised Learning

- Supervised learning: the Y is available during learning activities. E.g.,

X (Input)		**Y** (Output)
N. of Calls	**Avg Duration**	**Client Type**
7	10	Good
2	2	Bad
...	...	...

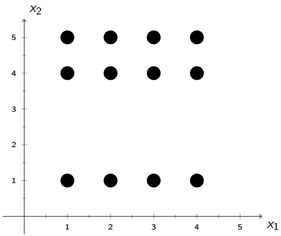
- Unsupervised learning: Y not available. Next table is like the previous but no Y.

X (Input)	
N. of Calls	**Avg Duration**
7	10
2	2
...	...

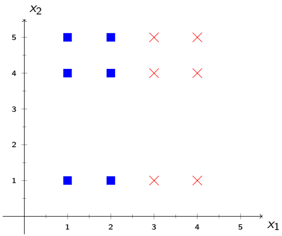
  - We try to identify groups using only X
  - E.g., clustering tries to identifies meaningful subsets based on data patterns/regularities).

Supervised vs Unsupervised Learning 2

- Let's try to identify 2 clusters (clustering is an unsupervised learning techniques):



Case 1: no Y



Case 2: Y available

- Supervised approaches are usually more effective
  - E.g., clustering works with no prior knowledge about the desired subsets.
  - The identified clusters might be meaningless
- If Y is not available during learning, only unsupervised approaches can be used

Continuous vs Discrete Variables

- Now, we will refer to the (general) variable w (it might be either X or Y, we don't care now). We will come back later to X and Y
- Formal definition. ~~Suppose X and Y are metric spaces, $E \subset X$, $p \in E$, and f maps E into Y. Then f is said to be *continuous* at p if for every $\epsilon > 0$ for all points $x \in E \ldots$~~
- Quick and dirty selection criteria. Given a variable w, its domain is
  - **Discrete** if the number of elements in the domain is **finite**, e.g., $w \in \{"Good\ Prospect\ Client", "Bad\ Prospect\ Client"\}$
  - **Continuous** if the values are numeric and the number of elements in the domain is **infinite** e.g., $w \in \mathbb{R}$ (real numbers)

Questions

- I need to classify movies using *1 to 5 stars* rating. No partial star allowed.
- Question. Is the domain *1 to 5 stars* continuous or discrete? Answers not allowed from people having a major in Mathematics!
  - It is *discrete*. The domain cardinality (the n. of elements of the domain is finite).
  - Even if numbers are used
- Let k be the temperature degree of this room. Is the k domain continuous or discrete?
  - It is *continuous*. The actual temperature is a real number (potentially, unlimited decimal digits), the cardinality of $\mathbb{R}$ is infinite
- Let k be a stock quotation at NASDAQ. The domain of quotes is continuous or discrete?
  - It is *continuous*. Although a NASDAQ stock price can have a maximum of 4 decimal digits, it is worth managing quote prices as a continuous domain. Furthermore, there is no upper limit to stock prices

ML Tasks/Problem Types

- Type of learning (recap)
  - Supervised learning: the Y is available during learning activities
  - Unsupervised learning: the Y is not available during learning activities
  - …(some hybrid approaches canalso be used e.g., semi-supervised, …). Now we will focus only on the first two
- Type of Y (Output)
  - Discrete
  - Continuous

The Machine Learning problems can be classified using the following table

	Supervised Learning	**Unsupervised Learning**
Discrete Output	Classification or Categorization	Clustering, …
Continuous Output	Regression	…

Classification

	Supervised Learning	**Unsupervised Learning**
Discrete Output	**Classification or Categorization**	Clustering, ...
Continuous Output	Regression	...

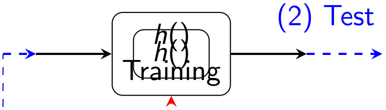
$$Y \longleftarrow \boxed{h()} \longleftarrow X$$

- Classification: automatically identify a function $Y = h(X)$ where X are observable features, and Y is discrete, e.g.,
  - $X = [\#$ of phone calls, average duration$]$,
    $y \in Y = \{"Good\ Prospect\ Client", "Bad\ Prospect\ Client"\}$
  - $X = [\%Alcohol, P.h.\ level]$, $y \in Y = \{"Grape\_Juice", "Wine", "Vinegar"\}$

Training/Testing a Classifier

X		Y
N. of Calls	**Avg Duration**	**Client Type**
7	10	Good
2	2	Bad
10	8	Good
3	2	Bad
8	7	Good
4	1	Bad
...	...	...
6	12	Good
3	4	Bad
9	6	Good
1	4	Bad

- Training: the classifier is fed with X and Y
- Y values are also called labels



Y Predicted
Good
Bad
Good
Bad

(1) Training (only a subset of records is used)

Evaluating Classification Performances

- How can I measure the performances of a classifier?
- Some useful metrics:
  - Accuracy: $accuracy = \frac{N.\ of\ \ correctly\ \ predicted\ \ items}{N.\ of\ \ predicted\ \ items}$
  - Precision (will be described later)
  - Recall (will be described later)
  - ...
- Some examples. Y_{True} is the real value, Y_{Pred} is the value predicted by the (trained) classifier.

Y_{True}	Y_{Pred}
G	G
B	B
G	G
B	B

Acc. = **1.0** (100%)

Y_{True}	Y_{Pred}
G	G
B	G
G	B
G	G

Acc. = **0.5** (50%)

Y_{True}	Y_{Pred}
G	G
B	G
G	B
G	B

Acc. = **0.25** (25%)

Precision, Recall, F1 Score

- Accuracy can be enriched by other measures
- Suppose there are n documents $(d_1, d_2, \ldots, d_n)$ that should be classified according to k classes $(c_1, c_2, \ldots c_k)$. Each document d_i has to be classified into at most one class c_j
  - Let Y_j^{Pred} be the subset of predictions related to c_j
  - Let Y_j^{True} be the subset of predictions that actually belong to class c_j
- $Precision_j = \frac{|Y_j^{Pred} \cap Y_j^{True}|}{|Y_j^{Pred}|} = \frac{n.\ of\ correctly\ classified\ documents}{n.\ of\ classified\ ones}$
- $Recall_j = \frac{|Y_j^{Pred} \cap Y_j^{True}|}{|Y_j^{True}|} = \frac{n.\ of\ correctly\ classified\ documents}{n.\ all\ documents\ in\ the\ category\ j}$
- There is a trade-off between precision and recall
- $F1\ Score_j = 2\frac{precision_j \cdot recall_j}{precision_j + recall_j}$
- Each class has a specific *Precision*, *Recall*, and *F1 Score*. The overall *Precision*, *Recall*, and *F1 Score* are the average of all the single class values.

Training and Test Set

- Classifier evaluation summary:
  - I evaluate $h()$ on a dataset (X, Y_{True}). The real labels are known
  - During prediction, the $h()$ works on a subset of X (no Y i.e., blind prediction)
  - Then Y_{Pred} is collected and compared with Y_{True}
- Which accuracy value will I get if I evaluate a classifier on the training set?
  - Answer: accuracy=1.0 ...
  - ...unless the training phase incurred in serious problems
- To check the classifier generalizability, it is paramount to evaluate a classifier on a dataset which was not used during training

Classification Recap

- A labelled dataset is necessary for classification (i.e. I need the Y in addition to the X).
- A labelled dataset is split in *train* and *test* subsets e.g., (50%, 50%), or (75%, 25%)
- The classifier is trained on the *train* subset (this task is also called *fitting* the classifier), the evaluation is performed on the *test* set
- Some considerations
    - I can perform the process above using several classifiers (i.e. several algorithms), each classifiers has parameters whose value strongly affect the classification performances
    - It takes time to identify the best combination of classifier and parameters for a given task (later, we will go deeper on this topic)

Regression

	Supervised Learning	**Unsupervised Learning**
Discrete Output	Classification or Categorization	Clustering, ...
Continuous Output	**Regression**	...

- Regression: automatically identify a function $y = h(x)$ where y belongs to an **continuous set** e.g., the **Real Numbers**. E.g.,
  - Identify house prices based on location and square meters, $x \in X =$ [location, square meters], $y \in \mathcal{R}$ where y is the price
  - Daily euro/dollar exchange rate, $x \in X = [date]$ and $y \in \mathcal{R}$ where y is the rate
- Regression is similar to Classification ...
  - It is based on a supervised approach (I need both X and Y)
  - Y is continuous

Regression Metric

- E.g., given a target house price to predict (i.e. $y^{True} = 200\,000$), …
- …suppose two regressors (reg_1 and reg_2) predicts two values:
  - $y^{Pred_1} = 200\,002$
  - $y^{Pred_2} = 290\,000$
- What about using accuracy to evaluate the classifier performances?
  - I can't use accuracy to evaluate Regression results (accuracy is on/off)
  - The results are both wrong, but y^{Pred_1} is a better than y^{Pred_2}
- Regression is usually evaluated using an error function e.g., …
  - …(Introducing) the Mean Squared Error (MSE) …
  - let Y^{True} the be the real prices of some houses and
    let Y^{pred} be the predictions of a classifier
  - $MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i^{True} - y_i^{Predic})^2$
  - Where $n = |Y^{True}| = |Y^{Pred}|$

The next two slides were not shown during lesson due to lack of time.
We will discuss them in one of the next lessons.

No free Lunch

Why can't I use Machine Learning for the task XYZ?

- Resource limitation
  - e.g., training algorithm ABC on the scenario XYZ will take 6 weeks
  - Preparing training datasets can be very expensive
  - Not enough data for training
- Model issues (we will come back on this topic during next lessons)

Will Artificial Intelligence take over all the Working Places?

- The *Great Horse Manure Crisis of 1894*. In 1894, the Times newspaper predicted... "In 50 years, every street in London will be buried under nine feet of manure."
  - E.g., about 100,000 horses in New York producing around 2.5M pounds of manure a day
  - By 1912, this seemingly insurmountable problem had been resolved (motor vehicles)
- The luddites movement: a group of English textile workers who destroyed weaving machinery in the 19th century because they fear for job loss
- The phrase "technological unemployment" was popularised by John Maynard Keynes in the 1930s.
  - It is widely accepted that technological change can cause short-term job losses
  - The view that it can lead to lasting increases in unemployment has long been controversial. Participants in the technological unemployment debates can be broadly divided into optimists and pessimists

- Thank you for your attention!
- Are there any questions?