Introduction to Machine Learning

Pascal Fua
IC-CVLab
Human vs Machine Learning

Learn from experience

Learn from experience
What is Machine Learning?

• Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.

• Machine learning algorithms provide knowledge to computers through data and interaction with the real world. It is then used to make accurate predictions given new observations.

• Machine learning is applied statistics.
Recognizing Hand-Written Digits

LeNet (1989-1999)
Recognizing Hand-Written Digits

![Image of handwritten digits 2 and 9 with corresponding diagrams.]
Predictor and Labels

28x28 pixels

$x$ is a 784-D Vector

$y: x \in \mathbb{R}^{784} \rightarrow \{0, 1, 2, \ldots, 9\}$

Predictor

Labels

$= x$
Labeled Training Set

\[ T = \{(x_n, t_n) \text{ for } 1 \leq n \leq N\} \]
Supervised Classification

Minimize:

\[ E(w) = \sum_{n=1}^{N} L(y(x_n; w), t_n) \]

- **x**: Feature vector
- **w**: Model parameters
- **t**: Label
- **y**: Predictor
- **L**: Loss Function
- **E**: Error Function

\[ \rightarrow \text{ML is an optimization problem} \]
Generic Scheme

Data → Algorithms → Insight
**Medical Research**

**Data:** Feature vectors that characterize mothers.

![Feature vector table](Image from Lumen Learning)

**Insight:** What characteristics of a mother contribute most to low birth weight.
Spam Detection

Feature vector: \[ \mathbf{x} = \begin{pmatrix} \#\text{viagra} \\ \#\text{pills} \\ \vdots \end{pmatrix} \]

Labels: Spam, Not Spam

Model parameters: \( \mathbf{w} \)

Predictor: \( y(\mathbf{x}, \mathbf{w}) = \{\text{Spam, Not Spam}\} \)
Recommender Systems

Feature vector:
- What films have you watched?
- Did you like them?

Predictor:
- List of films to propose.
Face Detection

\[ y : x \in \mathbb{R}^{W \times H} \rightarrow v \in \{\text{face, not face}\} \]
Labeled Training Set

Faces: Near frontal with varying ages, ethnicity, gender, lighting, ...

Non-faces: Images containing anything else.
Supervised Learning

Train using an annotated training set:

\{(\text{face}), (\text{face}), (\text{face}), \ldots, (\text{not-face}), (\text{not-face}), (\text{not-face}), \ldots\}

Run on images that do not belong to the test set:

→ Face or not?
Under the IBM Board Corporate Governance Guidelines, the Directors and Corporate Governance Committee and the full Board annually review the financial and other relationships between the independent directors and IBM as part of the assessment of director independence. The Directors and Corporate Governance Committee makes recommendations to the Board about the independence of non-management directors, and the Board determines whether those directors are independent. In addition to this annual assessment of director independence, independence is monitored by the Directors and Corporate Governance Committee and the full Board on an ongoing basis.
Binary Classification

Salmon or sea bass?
Brightness as a Feature

Some algorithm

$\begin{align*}
+1 \text{ (salmon)} \\
-1 \text{ (sea bass)}
\end{align*}$
1D Model

Model:
\[
\begin{cases}
-1 & \text{if brightness } < T \\
1 & \text{otherwise}
\end{cases}
\]

Model parameters:
\[ w = \{T\} \]

Error function:
\[
E(w) = \#\text{salmons with brightness } < T + \#\text{seabasses with brightness } \geq T
\]

Learning: Minimizing \( E(w) \) w.r.t. to \( w \)
Each circle denotes a 2-dimensional sample $\mathbf{x}_i$ of dimension 2, which is assigned a label $t_i \in \{-1, 1\}$.
Decision Boundary

In this class, we will discuss:
• How to define $y$.
• How to choose $w$.

Decision boundary:

$$C = \{ x \in \mathbb{R}^2, y(x, w) = 0 \}$$

$$y(x, w) > 0 : \text{Salmon}$$

$$y(x, w) < 0 : \text{Bass}$$
Training vs Testing

Some algorithm

\[
\begin{pmatrix}
\text{brightness} \\
\text{length}
\end{pmatrix}
\]

Training set = \{ ● , ● \}

Test set = \{ ○ , ○ , ○ , ..., \}

1. Use the training set to learn the model.

→ Makes sense only if both have the same statistical distribution.
Linear 2D Model

Some algorithm

\[
\begin{pmatrix}
\text{brightness} \\
\text{length}
\end{pmatrix}
\]

\[
y(x, w) = w_0 + w_x b + w_y l
\]

\[
x = [b, l]
\]

\[
w = [w_0, w_x, w_y]
\]

\[
\rightarrow +1
\]

\[
\rightarrow -1
\]

\[
\text{ Important special case.}
\]
Course Outline

Introduction
• ML Basics
• K Nearest Neighbors
• K Means

Linear ML
• Linear Regression
• Logistic Regression
• Max Margin Classifiers

Nonlinear ML
• AdaBoost
• Polynomial Support Vector Machines
• Kernel Methods
• Artificial Neural Networks
Kaggle Survey (2019)

- We will focus on methods that do not rely heavily on probabilities.
- We may cover some of the others at the end of the class time permitting.

What data science methods do you use at work?
My Erroneous Perception

- Will it evolve in that direction?
- Time will tell.
General Organization

• Lectures: Tuesdays 8:15-10 online
• Exercises: Tuesdays 8:15-10 online.
• Written exam with one page of notes.

• Slides: https://moodle.epfl.ch/course/view.php?id=16159

• Main references:
  • C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
  • M. Welling, A First Encounter with Machine Learning, 2011
• Available for free on the web.
• We will use the same notations.

Normal slide: It is part of the course and I may ask exam questions about it.

Reminder slide: We have already covered this earlier in the class. Go back to the appropriate lecture if you do not remember.

Optional slide: This is additional material for people interested in more details. I will not ask direct exam questions on this.

Reference to book or paper for even more details.