What is the definition of a **good Machine Learning algorithm**?

After 60 years, is this a closed problem? And if not …

Outline

1. What does work
2. Limitations
3. Learning comes with **which guarantees**?
   - Induction: how to win this game?
   - The statistical learning theory
   - A closed case? Not so sure
4. Other paradigms? An **historical perspective**
5. Is there a **paradigmatic change in sight**?
6. Conclusions
Object recognition in images

The ImageNet competition

- More than 15M high resolution labeled images
- Approximately 22K categories
- Taken from the Web and labeled using Amazon Mechanical Turk

Illustration: ImageNet

The ImageNet competition

- More than 15M high resolution labeled images
- Approximately 22K categories
- Taken from the Web and labeled using Amazon Mechanical Turk

Results: 8 ILSVRC-2010 test images

- Results

Object recognition

TEST IMAGE

RETRIEVED IMAGES

[Krizhevsky, Sutskever and Hinton (2012)]
**Image annotating**

*Figure 2.11: “A group of young people playing a game of frisbee”—that caption was written by a computer with no understanding of people, games or frisbees.*

**Speech recognition**

- Works reasonably well

Comparison (2012) of the word error rates achieved by traditional GMMs and DNNs, reported by three different research groups on three different benchmarks.

**The SuperVision network**

Image classification with deep convolutional neural networks


- 7 hidden “weight” layers
- 650K neurons
- 60M parameters
- 630M connections

**Machine translation**

- Still far from perfect, but …

From Hofstädter (2018)
Game playing with Reinforcement Learning

- E.g. AlphaGo

Outline

1. What does work
2. Limitations
3. Learning comes with which guarantees?
   - Induction: how to win this game?
   - The statistical learning theory
   - A closed case? Not so sure
4. Other paradigms? An historical perspective
5. Is there a paradigmatic change in sight?
6. Conclusions

Requires enormous training sets

- Image recognition
  - Object localization for 1000 categories.
    - Millions of images
Requires enormous training sets

- **Image recognition**
  - Object localization for 1000 categories.
  - Millions of images

- **AlphaGo**
  - Training on KGS dataset led to overfitting
  - Self-play data (30 million distinct positions, each sampled from a separate game)
  - Over the course of millions of AlphaGo vs AlphaGo games, the system progressively learned the game of Go from scratch, accumulating thousands of years of human knowledge during a period of just a few days. (In the first three days AlphaGo Zero played 4.9 million games against itself in quick succession.)

Exclusively focused on error rate

- **The Netflix prize**
  - The winner system was not used afterwards!!

- **Machine translation**
  - Good on easy and mundane texts
  - Bad on interesting texts
Weak account of the structure

- **Texts** as bags of words
- **Images** as simple correlations

**Example:** detection of the action “giving a phone call”

![Diagram of image detection](image)

[Oquab et al., CVPR (2014)] (~70% correct (SOTA))

**Learning systems do not work together flawlessly**

- Two sub-systems
  - One locating the ads links
  - The other the adds
- That influence each other
  - Each takes into account the clicks
  - Which depends in part from the actions of the other sub-system
  - In addition of other uncontrolled factors (price, user’s queries, ...)

**The Simpson’s paradox**

- Physicians would like to know whether drug A is more or less efficient than drug B
- Two groups of 350 patients each are chosen. One is given drug A, and the other drug B

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open surgery</td>
<td>78% (273/350)</td>
</tr>
<tr>
<td>Percutaneous nephrolithotomy</td>
<td>83% (289/350)</td>
</tr>
</tbody>
</table>

B is best?
The Simpson’s paradox

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Patients with small stones</th>
<th>Patients with large stones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment A:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open surgery</td>
<td>78% (273/350)</td>
<td>93% (81/87)</td>
<td>73% (192/263)</td>
</tr>
<tr>
<td>Treatment B:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percutaneous nephrolithotomy</td>
<td>83% (289/350)</td>
<td>87% (234/270)</td>
<td>69% (55/80)</td>
</tr>
</tbody>
</table>

• Influencing factor
  The choice of the patients for each group was function of the severity of the pathology

Thus, is the sky so blue?

Learning systems ...

1. Require enormous amounts of training data
2. Are exclusively focused on error rates
3. Do not fully take advantage of structures
4. Do not cooperate well
   - Software engineering with adaptive components is yet to be solved

Learning systems do not work together flawlessly

• Two sub-systems
  - One locating the ads links
  - The other the adds

• That influence each other
  - Each takes into account the clicks
  - Which depends in part from the actions of the other sub-system
  - In addition of other uncontrolled factors (price, user’s queries, ...)

Outline

1. What does work
2. Limitations
3. Learning comes with which guarantees?
   - Induction: how to win this game?
   - The statistical learning theory
   - A closed case? Not so sure
4. Other paradigms? An historical perspective
5. Is there a paradigmatic change in sight?
6. Conclusions
Which guarantees?
The statistical theory of learning

Supervised learning

Given a training set

\[ \mathcal{S}_m = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_m, y_m) \} \]

- Find an hypothesis \( h \in \mathcal{H} \) such that \( h(x_i) \approx y_i \)
- Hoping that it generalizes well:

\[ \forall x \in \mathcal{X} : h(x) \approx y \]

Supervised induction

- We want to be able to predict the class of unseen examples

A decision function

One example that tells a lot ...

- Examples described using:
  - Number (1 or 2);
  - Size (small or large);
  - Shape (circle or square);
  - Color (red or green)
- They belong either to class ‘+’ or to class ‘-’
### One example that tells a lot ...

- Examples described using:
  - `Number (1 or 2); size (small or large); shape (circle or square); color (red or green)`
- They belong either to class `+` or to class `-`

<table>
<thead>
<tr>
<th>Description</th>
<th>Your prediction</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 large red square</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 large green square</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>2 small red squares</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2 large red circles</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1 large green circle</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>1 small red circle</td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

### One example that tells a lot ...

- Examples described using:
  - `Number (1 or 2); size (small or large); shape (circle or square); color (red or green)`

### One example that tells a lot ...

- Examples described using:
  - `Number (1 or 2); size (small or large); shape (circle or square); color (red or green)`

### One example that tells a lot ...

- Examples described using:
  - `Number (1 or 2); size (small or large); shape (circle or square); color (red or green)`

### How many possible functions altogether from X to Y?

- $2^4 = 2^{16} = 65,536$

### How many functions do remain after 6 training examples?

- $2^{10} = 1024$

### How many remaining functions?

- 15

### How many possible functions with 2 descriptors from X to Y?

- $2^{12} = 2^4 = 16$

### How many functions do remain after 3 training examples?

- $2^1 = 2$
Induction: an impossible game?

- A bias is need

- Types of bias
  - Representation bias (declarative)
  - Research bias (procedural)

Interpreting – completion of percepts

Induction and its illusions
Clustering

The perceptron

- Rosenblatt (1958-1962)

\[ y_i = \text{sign} \left( \sum_{j=0}^{d} w_{ji} x_j \right) \]

\[ y_i = \begin{cases} 1 & \text{if } \sum_{j=0}^{d} w_{ji} x_j > 0 \\ -1 & \text{otherwise} \end{cases} \]

The perceptron

- Rosenblatt (1958-1962)
The perceptron: a linear discriminant

The perceptron learning rule

- Adjustments of the weight $W_i$
  
  Principle (Perceptron’s rule): learn only in case of prediction error

Algorithm 1: The perceptron learning algorithm

Data: A training sample $S_m = \{(x_i, y_i)\}_{1 \leq i \leq m}$

Result: A weight vector $w$

while not convergence do
  if the randomly drawn $x_i$ is st. $\text{sign}(w \cdot x_i) = y_i$ then
    do nothing
  else
    $w(t + 1) = w(t) + \eta x_i y_i$
  
Randomly select next training example $x_i$

The perceptron

NO reasoning !!!

Some remarkable properties !!

- Convergence in a finite number of steps
  
  - Independently of the number of examples
  
  - Independently of the distribution of the examples
  
  - Independently of the dimension of the input space

If there exists a linear separator of the training examples
The statistical theory of learning

Guarantees on generalization ??

• Theorems about the performance with respect to the training set

• We want guarantees about future examples

Statistical study for $|\mathcal{H}|$ hypotheses

It leads to:

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[ R(h) \leq \hat{R}(h) + \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m} \right] > 1 - \delta$$

The Empirical Risk Minimization principle is sound only if there exists a limit (a bias) on the expressivity of $\mathcal{H}$

The size $m$ of the training set must be large enough w.r.t. to capacity of $\mathcal{H}$

Bounds on the difference between the true risk and the empirical risk

• $\mathcal{H}$ finite, realizable case

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[ R(h) \leq \hat{R}(h) + \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m} \right] > 1 - \delta$$

• $\mathcal{H}$ finite, non realizable case

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[ R(h) \leq \hat{R}(h) + \sqrt{ \frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2m} } \right] > 1 - \delta$$
**Statistical theory of learning as a theory of justification**

Use of the ERM principle (fitting the data) is justified as long as the expressiveness (or capacity) of $\mathcal{H}$ is controlled (and limited)

$$\forall h \in \mathcal{H}, \forall \delta \leq 1: \quad P_m\left[ R(h) \leq R(h^*) + R_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \right] > 1 - \delta$$

---

**HOW TO ... devise learning algorithms**

1. Define an appropriate **regularized** (inductive) criterion
   1. Translate the cost of errors of prediction in the domain into a loss function
   2. Define a regularization term that expresses assumptions about the underlying regularities of the world
   3. If possible, make the resulting optimization problem a convex one

   $$h_{opt} = \underset{h \in \mathcal{H}}{\operatorname{ArgMin}} \left[ \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) + \lambda \underset{\text{bias on the world}}{\operatorname{reg}(\mathcal{H})} \right]$$

2. Use or develop an efficient optimization solver

---

**Learning sparse linear approximator**

- **The hypothesis** is of the form $h(x) = w \cdot x$
- **A priori assumption**: few non zero coefficients

**Ridge regression**

$$w_{ridge}^* = \underset{w}{\operatorname{ArgMin}} \left\{ \sum_{i=1}^{m} (y_i - w \cdot x_i)^2 + \lambda \|w\|^2 \right\}$$

**Lasso regression**

$$w_{lasso}^* = \underset{w}{\operatorname{ArgMin}} \left\{ \sum_{i=1}^{m} (y_i - w \cdot x_i)^2 + \lambda \|w\|_1 \right\}$$
A very alluring framework

1. Based on a justification theory
   - Bounds on the generalization error can be claimed (very important for having paper accepted)
   - Valid for the worst case: against any possible distribution of the data

2. Seemingly very benign assumptions on the world
   - Data (and future questions) supposedly i.i.d.
   - $f \in H$ or $f \notin H$

3. Provides a recipe to produce learning algorithms
   - Very generic applicability: minimization of a regularized empirical risk
   - Learning = optimization

A lot of “Lamppost theorems”

Theorems that guarantee that:

- If the world obeys my a priori assumptions
- Then the learning algorithm will end up with a good hypothesis (closed to the “real” one)

- Otherwise learning can lead to very bad hypotheses
  (e.g. If the world is not sparse)

« Deep learning »
as THE universal solution (2006– ...)

“The paper focuses on a subject that might be of limited importance at ICML, given the current trend towards neural networks.”
Artificial Neural Networks

- With numerous hidden layers (possibly > 100x)
- And a very large number of parameters (~ $10^2$ – $10^3$ parameters)
- Learn hierarchical and compositional representations

**GoogleNet**

- A mécano of neural networks

Troubling findings

A paper

  "Understanding deep learning requires rethinking generalization"

Extensive experiments on the classification of images

- The AlexNet (> 1,000,000 parameters) + 2 other architectures

- The CIFAR-10 data set:
  - 60,000 images categorized in 10 classes (50,000 for training and 10,000 for testing)
  - Images: 32x32 pixels in 3 color channels
Troubling findings

Experiments

1. **Original dataset without modification**
   - Results?
     - Training accuracy = 100% ; Test accuracy = 89%
     - Speed of convergence ~ 5,000 steps

2. **Random labels**
   - Training accuracy = 100% !?? ; Test accuracy = 9.8%
   - Speed of convergence = similar behavior (~ 10,000 steps)

Expected behavior if the capacity of the hypothesis space is limited

i.e. the system cannot fit any (arbitrary) training data

\[
\forall h \in \mathcal{H}, \forall \delta \leq 1: \quad \Pr\left[R(h) \leq \bar{R}(h) + 2 \text{rad}_{\delta}(\mathcal{H}) \right] \geq 1 - \delta
\]
Troubling findings

- Deep NNs can accommodate **ANY** training set

\[ \forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[ R(h) \leq \hat{R}(h) + 2 \hat{\text{Rad}}(\mathcal{H}) + 3 \frac{\ln(2/\delta)}{\ln m} \right] > 1 - \delta \]

But then, **why are deep NNs so good on image classification tasks?**

Explanations and deep neural networks

Identification object categories in an image
- Here, two classes: « **dog** » and « **tiger cat** »

Adversarial learning

Illustration

Example: Adversarial input can fool a machine-learning algorithm into misinterpreting images.

Evaluation protocol: comparison between explanations
- Which robot do you trust most?

Both robots predicted: Person

<table>
<thead>
<tr>
<th>What do you see?</th>
<th>Robot A based it's decision on</th>
<th>Robot B based it's decision on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your options:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Which robot is more reasonable?
- Robot A seems clearly more reasonable than robot B
- Robot A seems slightly more reasonable than robot B
- Both robots seem equally reasonable
- Robot B seems clearly more reasonable than robot A
- Robot B seems slightly more reasonable than robot A

54 subjects on Amazon Turk: robot B evaluated 1.27 (between -2 et +2)

---

[Selvaraju et al. (2017)] Grad-CAM: Visual explanations from deep networks via gradient-based localization •

[Selvaraju et al. (2017)] Grad-CAM: Visual explanations from deep networks via gradient-based localization •
Explanations and deep neural networks

Optical illusions: how to explain them?

- Boxer: 0.40 Tiger Cat: 0.18

- Airliner: 0.9999

- Boxer: 1.3e-20

- Tiger Cat: 8.3e-11

!!??


Assessment

1. A theory
   - for stationary environments and i.i.d. data and queries!!
   - focused on the expectation of the cost of errors
     - Prior knowledge must be encoded in the cost
   - that can produce learning algorithms when combined with optimization techniques

2. Deep NNs
   - depart from this framework
     - Demand at least a reworking of the theory
     - Prior knowledge encoded in the architecture
   - Still
     - require enormous amount of data
     - Focused on error rates
     - Based on correlations

Outline

1. What does work
2. Limitations
3. Learning comes with which guarantees?
   - Induction: how to win this game?
   - The statistical learning theory
   - A closed case? Not so sure
4. Other paradigms? An historical perspective
5. Is there a paradigmatic change in sight?
6. Conclusions
Are there other paradigms?
An historical perspective on ML

E.g. The PRODIGY system

PRODIGY: An Integrated Architecture for Planning and Learning
Jaime Carbonell, Daren Estrin, Willy Gil, Robert Joseph
Craig Knoblock, Steve Minton, and Monica Veloso

PRODIGY's basic reasoning engine is a general-purpose problem solver and planner [10] that searches for sequences of operators (i.e., plans) to accomplish a set of goals from a specified initial state description. Search in PRODIGY is guided by a set of control rules that apply at each decision point.

PRODIGY's reliance on explicit control rules which can be learned for specific domains distinguishes it from most domain independent problem solvers. Instead of using a least-commitment search strategy, for example, PRODIGY expects that any important decisions will be guided by the presence of appropriate control knowledge. If no control rules are relevant to a decision, then PRODIGY makes a quick, arbitrary choice. If in fact the wrong choice is made, and costly backtracking proves necessary, an attempt will be made to learn the control knowledge that must be missing.

Illustration: LEX (Tom Mitchell)

... as

a means to improve the efficiency of a problem solver
Calculer la primitive de :
\[ \int 3x \cos(x) \, dx \]

Illustration: LEX (Tom Mitchell)

Génération de problèmes

Résolution de problèmes

Critique

Généralisation

Un des exemples positifs proposés :
\[ \int 3x \cos(x) \, dx \rightarrow \text{Appliquer OP2 avec :} \]
\[ a = 3x \]
\[ dv = \cos(x) \, dx \]

Illustration: LEX (Tom Mitchell)

Génération de problèmes

Résolution de problèmes

Critique

Généralisation

Un des exemples positifs proposés :
\[ \int 3x \cos(x) \, dx \rightarrow \text{Appliquer OP2 avec :} \]
\[ a = 3x \]
\[ dv = \cos(x) \, dx \]

Learning from one example

Explanation-Based Learning

1. From a single example
2. Try to prove the “fork”
3. Generalize
**Explanation-Based Learning**

Ex: learn the concept  \( \text{stackable} \text{(Object1, Object2)} \)

- **Domain theory:**
  
  (T1): \( \text{weight}(\text{X}, W) \leftarrow \text{volume}(\text{X}, V), \text{density}(\text{X}, D), W = V \times D \).
  
  (T2): \( \text{weight}(\text{X}, 50) \leftarrow \text{is\_a}(\text{X}, \text{table}) \).
  
  (T3): \( \text{lighter\_than}(\text{X}, Y) \leftarrow \text{weight}(\text{X}, W1), \text{weight}(\text{X}, W2), W1 < W2 \).

- **Operationality constraint:**
  
  - Concept should be expressible using \( \text{volume}, \text{density}, \text{color}, \ldots \)
  
  - Positive example (solution):

    - \( \text{on}(\text{obj1}, \text{obj2}) \).
    - \( \text{is\_a}(\text{obj1}, \text{box}) \).
    - \( \text{is\_a}(\text{obj2}, \text{table}) \).
    - \( \text{color}(\text{obj1}, \text{red}) \).
    - \( \text{color}(\text{obj2}, \text{blue}) \).
    - \( \text{made\_of}(\text{obj1}, \text{cardboard}) \).
    - \( \text{made\_of}(\text{obj2}, \text{wood}) \).
    - \( \text{owner}(\text{obj1}, \text{frederic}) \).
    - \( \text{color}(\text{obj2}, \text{blue}) \).
    - \( \text{made\_of}(\text{obj1}, \text{cardboard}) \).
    - \( \text{made\_of}(\text{obj2}, \text{wood}) \).

**Explanation-Based Learning**

- Induction from a single example
  
  - … plus a strong domain theory

- Based on
  
  - Logic-based knowledge representation
  
  - Reasoning Operators (deduction, goal regression in a proof tree, …)

Now used in SAT “solvers”
Explanation-Based Learning

- What was the aim of learning?
- What was a good method of learning?

1. Method improving the problem solving performances
   - [Steve Minton (1990) « Quantitative results concerning the utility of Explanation-Based Learning »]

2. Method that simulates the performances (and limits) of a natural cognitive agent (human or animal)

Learning and reasoning

Papers like


But

No measure of generalization performance independent of the problem-solver

Difficulties to scale up and to face noisy data

... when data started to pour down

Outline

1. What does work
2. Limitations
3. Learning comes with which guarantees?
   - Induction: how to win this game?
   - The statistical learning theory
   - A closed case? Not so sure
4. Other paradigms? An historical perspective
5. Is there a paradigmatic change in sight?
6. Conclusions

New learning scenarios

=> Is there a paradigmatic change in sight?
Furthermore, we compute the log odds for the presence of the object of interest. Using the previous three featurizations we compute, for each feature, the context image, objects of interest, object image, we featurize the original image. We featurize each image. Experiments will repeat this process for all the twenty objects of interest.

Scores are computed independently from the causal/anticausal scores. For simplicity, the follow-its object score, context score, causal score, and anticausal score. Importantly, the object/context class labels because NCC has been trained on continuous data with full support on classifier. We use features before their nonlinearity and log odds instead of the class probabilities or (activations before the classifier nonlinearity) obtained from the image network classifier formed by two hidden layers of 512 as a feature extractor. This network was trained on the entire ImageNet COCO dataset, we study the same classes. This selection amounts to 99,309 images. We preprocess multiple objects from different categories. The objects may appear at different scales and angles to obtain true zero (black) pixels.

Figure 4: Blackout processes for object of interest "dog". Original images.
Why should we not prescribe tetracycline to a child under the age of 8?

Expert justifications

- Drug depot on developing bones
  - Definitive blackening of the teeth
  - Socially unwanted coloration
  - Do not administer tetracycline to children under the age of

Notion of undesirable side effects

Causality relationships

Transfer learning

Definition [Pan, TL-IJCAI’13 tutorial]

- Ability of a system to recognize and apply knowledge and skills learned in previous domains/tasks to novel domains/tasks

Example

- We have labeled images (person / no person) from a web corpus
- Novel task: is there a person in unlabeled images from a video corpus?
Transfert learning: questions

• What can be the basis of transfer learning?
  How to translate formally: “the target domain is like the source domain”?

• What determine a good transfer?
  – A “good source”?
  – A high “similarity” between source and target?

• What formal guarantees can we have on the transferred hypothesis?

Transfer and analogy

Why should ‘a b a b c d’ be any better than ‘a b d’?

Transfer and sequence effects

Long-life learning

• Learning organized in a sequence of tasks
  – Very far from the i.i.d. scenario

  Learning will be affected by the history of the system

• We need a theory of the dynamics of learning
  1. Which sequence effects can we expect?
  2. How to best organize the curriculum of a learning system?
Conclusions

The current situation

- Inductive learning needs biases
  - No objective bias-free results

- The theory
  - Is focused entirely on the error rate
  - Assumes stationary environment and random inputs (i.i.d.)
  - Requires large enough data sets w.r.t. to the capacity of $H$

- We do not understand well deep neural networks

- Correlations ≠ structures, semantics, causation

We start to pay attention to new demands

1. The need for explanations
   - Structures
   - Causal reasoning
   - No more only error rate

We start to pay attention to new demands

1. The need for explanations
   - Structures
   - Causal reasoning
   - No more only error rate

2. The need for transfer learning
   - What should be transferred?
   - Conditions for positive / negative transfer?
We start to pay attention to new demands

1. The need for explanations
   - Structures
   - Causal reasoning
   - No more only error rate

2. The need for transfer learning
   - What should be transferred?
   - Conditions for positive / negative transfer?

3. Scenarios away from the i.i.d. assumption
   - Online learning / changing environments
   - Curriculum learning
   - Long-life learning

Conclusions: “new” scenarios

• Limited data sources
  – We often learn from (very) few examples

• The past history of learning affects learning: Education
  – Sequence effects

• We learn in order to build “theories”
  – All the time: small and large theories

For instance, what would you like to ask?