Neuro-Symbolic Artificial Intelligence
Chapter 5
Symbolic Machine Learning

Nils Holzenberger

March 19, 2024

Halftime

Some statistics:

- You are (more than) halfway through this class
- There are 3 lab sessions left and 1 exam (no documents, no switched-on devices)
- I have posted 3 past exams with solutions

Outline

1. Some more logic
   - Quantifiers
   - Previous lab session
   - Proof by resolution
   - Quantifiers and implications
2. Symbolic *vs* statistical machine learning
   - Knowledge
   - Explanations
   - Anomalies
   - Mechanics
3. Symbolic machine learning
   - Reinforcement learning
   - Analogies
   - Inductive logic programming
   - Machine learning as compression

Outline

Outline

Quantifiers in natural language

⚠ This is a joke about quantifiers

In this country a woman gives birth every fifteen minutes.

Quantifiers in natural language

⚠ This is a joke about quantifiers

In this country a woman gives birth every fifteen minutes. Our job is to find that woman and stop her.
— Groucho Marx

Outline

1. Some more logic
   - Quantifiers
   - Previous lab session
   - Proof by resolution
   - Quantifiers and implications
2. Symbolic *vs* statistical machine learning
3. Symbolic machine learning

Previous lab session

Error in question "Resolution with a trap"

The implication was in the wrong direction in the question

Thank you for telling me this

This question will not be graded

Outline

Proof by resolution

$$\begin{array}{l} [\neg A, B] \\ [A] \\ \hline [B] \end{array}$$

Why do we do this?

Proof by resolution

Goal: prove that $((\neg A \vee B) \wedge A)$ is a tautology

Proof by resolution

Goal: prove that $((\neg A \vee B) \wedge A)$ is a tautology

$\rightarrow$ show that $\neg((\neg A \vee B) \wedge A)$ is not satisfiable

Proof by resolution

Goal: prove that $((\neg A \vee B) \wedge A)$ is a tautology

$\rightarrow$ show that $\neg((\neg A \vee B) \wedge A)$ is not satisfiable

$\rightarrow$ show that whatever valuation I pick, $v(\neg((\neg A \vee B) \wedge A)) = \text{False}$

Proof by resolution

Goal: prove that $((\neg A \vee B) \wedge A)$ is a tautology

$\rightarrow$ show that $\neg((\neg A \vee B) \wedge A)$ is not satisfiable

$\rightarrow$ show that whatever valuation I pick, $v(\neg((\neg A \vee B) \wedge A)) = \text{False}$

$$[\neg((\neg A \vee B) \wedge A)]$$

Proof by resolution

Goal: prove that $((\neg A \vee B) \wedge A)$ is a tautology

→ show that $\neg((\neg A \vee B) \wedge A)$ is not satisfiable

→ show that whatever valuation I pick, $v(\neg((\neg A \vee B) \wedge A)) = \mathrm{False}$

$$[\neg((\neg A \vee B) \wedge A)]$$

...

---

(1) $[\neg A, B]$
(2) $[A]$

Proof by resolution

Goal: show that whatever valuation I pick, $v(\neg((\neg A \lor B) \land A)) = \text{False}$

$$\begin{array}{ll} (1) & [\neg A, B] \\ (2) & [A] \end{array}$$

Proof by resolution

Goal: show that whatever valuation I pick, $v(\neg((\neg A \vee B) \wedge A)) = \text{False}$

$$
\begin{array}{l}
(1)\ [\neg A, B] \\
(2)\ [A]
\end{array}
$$

Let v be a valuation.

Proof by resolution

Goal: show that whatever valuation I pick, $v(\neg((\neg A \lor B) \land A)) = \mathrm{False}$

$$\begin{array}{l}(1)\ [\neg A, B]\\(2)\ [A]\end{array}$$

Let v be a valuation.

- If $v(A) = \mathrm{True}$, $v((1)) = v(B)$ and $v((2)) = \mathrm{True}$, so the valuation of the whole thing is $v(B)$.

Proof by resolution

Goal: show that whatever valuation I pick, $v(\neg((\neg A \vee B) \wedge A)) = \mathrm{False}$

$$\begin{array}{ll} (1) & [\neg A, B] \\ (2) & [A] \end{array}$$

Let v be a valuation.

- If $v(A) = \mathrm{True}$, $v((1)) = v(B)$ and $v((2)) = \mathrm{True}$, so the valuation of the whole thing is $v(B)$.
- If $v(A) = \mathrm{False}$, $v((1)) = \mathrm{True}$ and $v((2)) = \mathrm{False}$ so the valuation of the whole thing is False.

Proof by resolution

Goal: show that whatever valuation I pick, $v(\neg((\neg A \vee B) \wedge A)) = \mathrm{False}$

$$\begin{array}{l}(1)\ [\neg A, B] \\ (2)\ [A]\end{array}$$

Let v be a valuation.

- If $v(A) = \mathrm{True}$, $v((1)) = v(B)$ and $v((2)) = \mathrm{True}$, so the valuation of the whole thing is $v(B)$.
- If $v(A) = \mathrm{False}$, $v((1)) = \mathrm{True}$ and $v((2)) = \mathrm{False}$ so the valuation of the whole thing is False.

$\rightarrow$ I only need to consider $v(B)$

Exercise: why can I merge $[A, X, B]$ and $[C, \neg X, D]$ to $[A, B, C, D]$?

Outline

1. Some more logic
   - Quantifiers
   - Previous lab session
   - Proof by resolution
   - Quantifiers and implications
2. Symbolic *vs* statistical machine learning
3. Symbolic machine learning

Why $(((\forall x)A) \supset B) \equiv (\exists x)(A \supset B)$ and not $(((\forall x)A) \supset B) \equiv (\forall x)(A \supset B)$?

Proof using equivalence with $\wedge$ and $\vee$

$$
\begin{aligned}
(((\forall x)A) \supset B) \quad & \equiv ((\neg((\forall x)A)) \vee B) \\
& \equiv (((\exists x)(\neg A)) \vee B) \\
& \equiv (\exists x)(\neg A \vee B) \\
& \equiv (\exists x)(A \supset B)
\end{aligned}
$$

Why $(((\forall x)A) \supset B) \equiv (\exists x)(A \supset B)$ and not $(((\forall x)A) \supset B) \equiv (\forall x)(A \supset B)$?

Example where $(((\forall x)A) \supset B) \not\equiv (\forall x)(A \supset B)$:

$B = \bot$

Domain $D = \{0, 1\}$

Interpretation of A: $A^I = x == 0$

- Left side
  - $((\forall x)A)$ is False
  - $((\forall x)A) \supset B)$ is True
- Right side
  - For assignment $x = 0$, $A^I \supset B^I$ is False
  - $(\forall x)(A \supset B)$ is False

Why $(((\forall x)A) \supset B) \equiv (\exists x)(A \supset B)$ and not $(((\forall x)A) \supset B) \equiv (\forall x)(A \supset B)$?

Examples where $(((\forall x)A) \supset B) \equiv (\forall x)(A \supset B)$:

- If the domain D contains a single element, then $\forall x$ and $\exists x$ are the same.
- If x occurs neither in A nor in B, then $\forall x$ and $\exists x$ behave the same in that formula.

Outline

Symbolic *vs* statistical machine learning

- Symbolic machine learning: define syntax over symbols to prove theorems
- Statistical machine learning: define random variables and parameterize the probabilities

Outline

Background knowledge

- In symbolic ML: background knowledge can be added easily
  - Add a rule
  - Add an entire knowledge base
  - Tweak one parameter
- In statistical ML: background knowledge is acquired as part of the target task

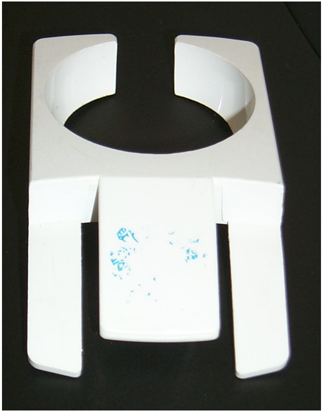
Auditability

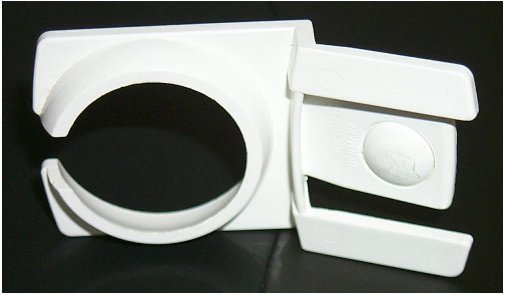
- What does ChatGPT know?
- Symbolic models can be *audited*
- Statistical models, not so much

Editability

- The knowledge in symbolic models can be edited (insert, delete, replace)
- In statistical ML it's possible (see Lake et al) but takes many repetitions
  - Acquiring a new word for a language model is estimated to take ~10k occurrences of the word
  - There are ways to construct one-shot learning, e.g. Lake *et al*, *One shot learning of simple visual concepts*, CogSci 2011

One-shot learning of unknown object





Generalization

- The point of machine learning is to build a model using training data, and then to use it on new data
- A model that works well on new data has good *generalization*
- Historically, statistical ML has generalized better than symbolic ML
- Statistical systems also learn structure: *While deep networks are capable of memorizing noise data, our results suggest that they tend to prioritize learning simple patterns first.*[1]

[1] Arpit *et al*, *A Closer Look at Memorization in Deep Networks*, ICML 2017

Outline

Criteria for explanations

- Relevance
  - Adapted to the level of expertise of the user
  - Specific: just highlighting the part of the input that led to the decision is not specific enough
- Faithfulness: Is the reason provided the actual reason that was used to get to the output?

Symbolic ML

- Typically, a model is its own explanation
- The rules define how the input is mapped to the output (→ faithfulness)
- Rules can be translated to match the desired level of expertise and specificity (→ relevance)
- Generally this translation is a challenge

Statistical ML

Yes, when you add two odd numbers together, the result is always an odd number. This is because any odd number can be expressed as 2n+1, where n is an integer. When you add two numbers in this form, the result is (2n+1)+(2m+1) = 2(n+m) + 2, which is also in the form 2p+1, where p is an integer. This means that the result is an odd number.
— ChatGPT, early 2023

- Numerical computations need to be translated to relevant and faithful explanations
- Post-hoc models of explanability have no guarantee of being faithful

Outline

1. Some more logic
2. Symbolic *vs* statistical machine learning
   - Knowledge
   - Explanations
   - Anomalies
   - Mechanics
3. Symbolic machine learning

AI-generated images

Which image is AI-generated?

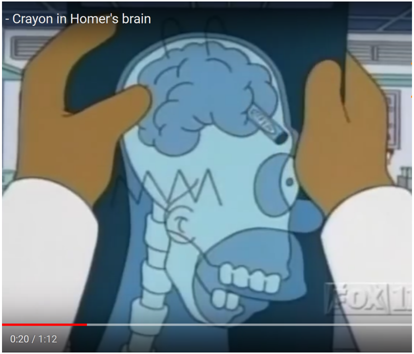
AI-generated images

Which image is AI-generated?

→ there are anomalies

https://hyperallergic.com/808778/ai-image-generators-finally-figured-out-hands/

Homer Simpson's brain



An AI image recognition software would not understand the anomaly because

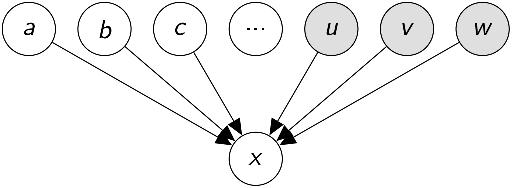
- a brain with a crayon in it looks almost like a brain and
- it has never seen crayons in brains

Contradiction

- Symbolic ML is sensitive to it
- Statistical ML is not

Outline

Randomness



- Many factors cause x, but we only know some of them, so it *appears* that the behavior is random
- Saying that x is random is like saying "I don't know the mechanisms that govern the behavior of x"
- The best thing would be to find out the mechanism; the next best thing is to model the probability
- Imagine modeling the trajectory of the Earth around the sun by interpolating the curve with a polynomial

Independently controllable features

Independently Controllable Factors

Valentin Thomas [*1 2] Jules Pondard [*1 2 3] Emmanuel Bengio [*4]
Marc Sarfati [1 5] Philippe Beaudoin [2] Marie-Jean Meurs [6] Joelle Pineau [4]
Doina Precup [4] Yoshua Bengio [1 7]

August 29, 2017

Models

- Symbolic and statistical systems are models of reality, not reality itself
- *All models are wrong, some of them are useful* — George E. P. Box

Outline

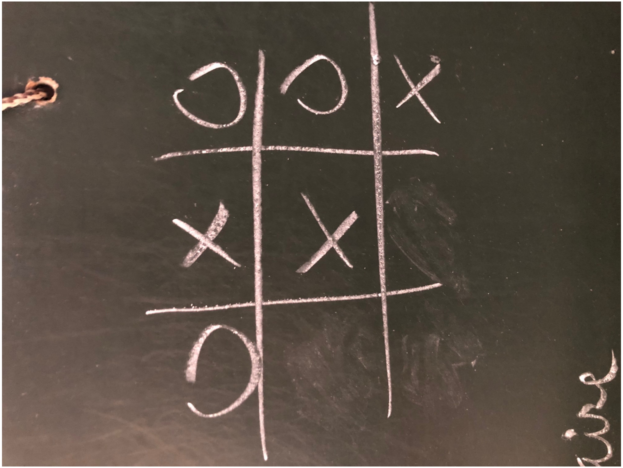
Symbolic *vs* statistical machine learning

- This lecture is mostly about symbolic machine learning
- The next lectures will be about statistical machine learning

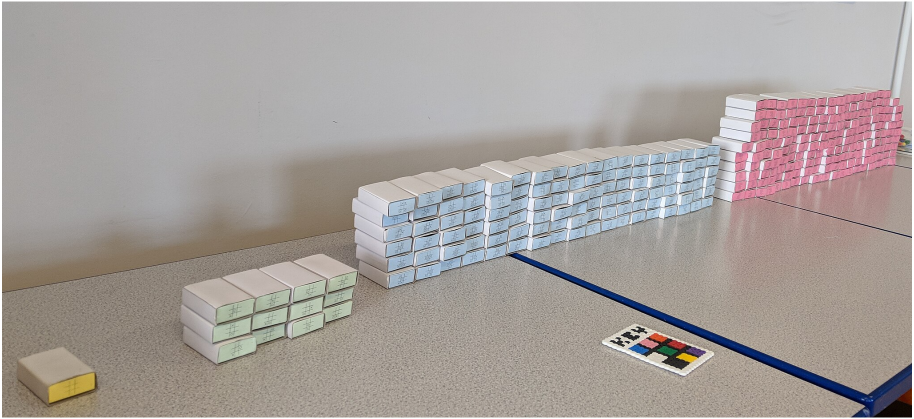
Outline

1. Some more logic
2. Symbolic *vs* statistical machine learning
3. Symbolic machine learning
   - Reinforcement learning
   - Analogies
   - Inductive logic programming
   - Machine learning as compression

Noughts and Crosses/Tic-Tac-Toe



Matchbox Educable Noughts and Crosses Engine

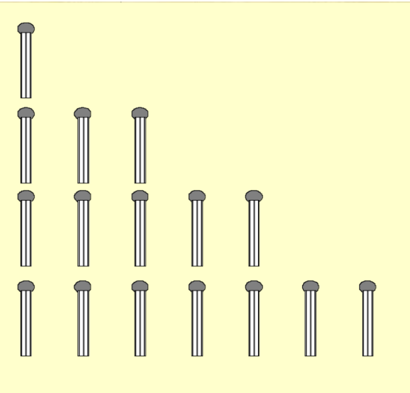


https://en.wikipedia.org/wiki/Matchbox_Educable_Noughts_and_Crosses_Engine

Matchbox Educable Noughts and Crosses Engine

- Donald Michie, 1961
- 304 matchboxes, one for each state of the game (up to rotation and symmetry)
- Beads of 9 different colors (one for each possible move)
- To decide which move to make:
  - Go to the matchbox corresponding to the game state
  - Draw a bead from it, and take that move
- If the game was won, **return the beads** to their original box, and **add 3 more beads** of that color
- If the game was lost, **don't return** the beads to their original box
- If the game was a draw, **return the beads** and **1 more** to their original box

Nim



- Players take turns removing matches
- Each player can remove as many matches as they like (at least 1), as long as they all come from the same row
- The last player to remove a match loses

Outline

1. Some more logic
2. Symbolic *vs* statistical machine learning
3. Symbolic machine learning
   - Reinforcement learning
   - Analogies
   - Inductive logic programming
   - Machine learning as compression

Analogies

$$\begin{array}{rcl} \text{ghi} & \rightarrow & \text{ghj} \\ \text{uuvvww} & \rightarrow & \end{array}$$

Analogies

ghi	→	ghj
uuvvww	→	uuvvxx
		uuvvjj
		uuvvwx
		ghj
		uuwvwx
		uuvvj
		uuvvww
		uuvvwj
		error

Analogies

- On-the-fly learning of rules
- Many tasks are a form of analogy
  - solve → solves, get → ? conjugation in English[2]
  - rosa → rosam, vita → ? declension in Latin
  - orang → orang-orang, burung → ? plural in Indonesian
- Analogies are highly discrete, but may be approximated by continuous representations, e.g. word embeddings[3]

[2] Murena *et al*, *Solving Analogies on Words based on Minimal Complexity Transformation*, IJCAI 2020

[3] Mikolov *et al*, *Distributed Representations of Words and Phrases and their Compositionality*, NIPS 2013; Chen *et al*, *Evaluating vector-space models of analogy*, CogSci 2017

Outline

Deduction *vs* induction

Deduction: rules → conclusions (Prolog)

Induction: conclusions → rules (Progol, Stephen Muggleton, 1995)

Learning rules

```
cute(X) :- dog(X), small(X), fluffy(X).  (1)
cute(X) :- cat(X), fluffy(X).            (2)
```

Learning rules

```
cute(X) :- dog(X), small(X), fluffy(X).  (1)
cute(X) :- cat(X), fluffy(X).            (2)
```

Least-general generalization of (1) and (2): `cute(X) :- fluffy(X).`

Learning rules

```
cute(X) :- dog(X), small(X), fluffy(X).  (1)
cute(X) :- cat(X), fluffy(X).            (2)
```

Least-general generalization of (1) and (2): `cute(X) :- fluffy(X).`

```
pet(X) :- dog(X).      (3)
pet(X) :- cat(X).      (4)
small(X) :- cat(X).    (5)
tame(X) :- pet(X).     (6)
```

Learning rules

```
cute(X) :- dog(X), small(X), fluffy(X).   (1)
cute(X) :- cat(X), fluffy(X).             (2)
```

Least-general generalization of (1) and (2): `cute(X) :- fluffy(X).`

```
pet(X) :- dog(X).       (3)
pet(X) :- cat(X).       (4)
small(X) :- cat(X).     (5)
tame(X) :- pet(X).      (6)
```

Least-general generalization of (1)-(6):
`cute(X) :- pet(X), small(X), fluffy(X).`

Inverse resolution

Association Rule Mining

Data-driven version of inverse resolution

- The data D is a set of *transactions*
  e.g. Transaction = list of items someone bought in a shop
- Every transaction has a set of binary attributes
  e.g. Attribute i = whether person bought item $\#i$
- An *itemset* is a subset of a transaction
- *Support* of itemset X is number of occurrences in D
  $\text{support}(X) = |\{t | t \in D, X \subseteq t\}|$
- *Confidence* in rule $X \rightarrow Y$ is $\frac{\text{support}(X \cap Y)}{\text{support}(X)}$

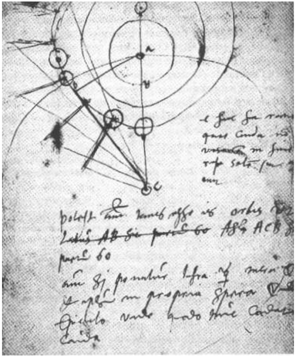
⚠ This is based on co-occurrence in data, while inverse resolution is based on existing rules.

Agrawal *et al*, *Mining association rules between sets of items in large databases*, SIGMOD 1993;
Belyy and Van Durme, *Script Induction as Association Rule Mining*, NUSE@ACL 2020

Outline

Tycho Brahe

Tycho Brahe
1546 - 1601



https://en.wikipedia.org/wiki/Tycho_Brahe

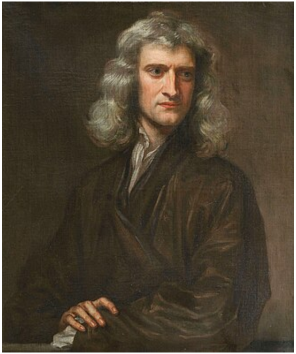
Johannes Kepler

Johannes Kepler
1571 - 1630

1. The orbit of every planet is an ellipse with the sun at one of the two foci.
2. A line joining a planet and the Sun sweeps out equal areas during equal intervals of time.
3. The ratio of the square of an object's orbital period with the cube of the semi-major axis of its orbit is the same for all objects orbiting the same primary.
$$\frac{T^2}{a^3} = \text{constant}$$

https://en.wikipedia.org/wiki/Johannes_Kepler

Isaac Newton



Isaac Newton
1643 - 1727

1. A body remains at rest, or in motion at a constant speed in a straight line, except insofar as it is acted upon by a force.
2. $\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = \sum_i \vec{F}_i$
3. $\vec{F}_{A \to B} = -\vec{F}_{B \to A}$ and $\vec{F}_{A \to B} \cdot \vec{AB} = 0$

https://en.wikipedia.org/wiki/Isaac_Newton

Compression

Reality

Compression

Reality

Observations

Compression

Reality

Observations

Empirical laws

$$\frac{T^2}{a^3} = \text{constant}$$

Compression

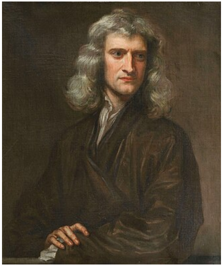
Reality

Observations

Empirical laws

$$\frac{T^2}{a^3} = \text{constant}$$

Principles



$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = \sum_i \vec{F}_i$$

Compression

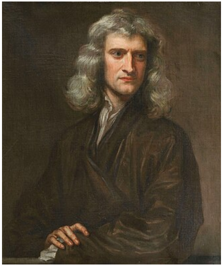
Reality

Observations

Empirical laws

$$\frac{T^2}{a^3} = \text{constant}$$

Principles



$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = \sum_i \vec{F}_i$$

ChatGPT as compression

ANNALS OF TECHNOLOGY

CHATGPT IS A BLURRY JPEG OF THE WEB

OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?

By **Ted Chiang**

February 9, 2023

Minimum description length

Which one is the best model?

- An equation with 8 parameters that explains 92% of observations
- A parametric function with 12M parameters trained on 1M samples that explains 96% of observations

Minimum description length

Which one is the best model?

- An equation with 8 parameters that explains 92% of observations
- A parametric function with 12M parameters trained on 1M samples that explains 96% of observations

The answer depends on:

- Your goal
  - Predict
  - Understand
- The cost of
  - Making inaccurate predictions
  - Computation
    - Training (a.k.a. Parameter estimation)
    - Inference
    - Collecting data samples
- ...

These criteria can be unified using *minimum description length*

$\mathrm{DL}(\text{data}) = \mathrm{DL}(\text{model}) + \mathrm{DL}(\text{data}|\text{model})$