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

1. Some more logic
   - Quantifiers
   - Previous lab session
   - Proof by resolution
   - Quantifiers and implications

2. Symbolic vs statistical machine learning

3. Symbolic machine learning
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
⚠️ This is a joke about quantifiers

*In this country a woman gives birth every fifteen minutes.*
⚠️ 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
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

1. Some more logic
   - Quantifiers
   - Previous lab session
   - Proof by resolution
   - Quantifiers and implications

2. Symbolic vs statistical machine learning

3. Symbolic machine learning
Proof by resolution

\[
\begin{align*}
[\neg A, B] \\
[A] \\
\hline \\
[B]
\end{align*}
\]

Why do we do this?
Goal: prove that \((\neg A \lor B) \land A\) is a tautology
Proof by resolution

Goal: prove that \(((\neg A \lor B) \land A)\) is a tautology

→ show that \(\neg((\neg A \lor B) \land A)\) is not satisfiable
Proof by resolution

Goal: prove that \((\neg A \lor B) \land A\) is a tautology
→ show that \(\neg((\neg A \lor B) \land A)\) is not satisfiable
→ show that whatever valuation \(I\) pick, \(\nu(\neg((\neg A \lor B) \land A)) = False\)
Proof by resolution

Goal: prove that \( ((\neg A \lor B) \land A) \) is a tautology

→ show that \( \neg((\neg A \lor B) \land A) \) is not satisfiable

→ show that whatever valuation I pick, \( \nu(\neg((\neg A \lor B) \land A)) = \text{False} \)

\[ \neg((\neg A \lor B) \land A) \]
Proof by resolution

Goal: prove that \( ((\neg A \lor B) \land A) \) is a tautology
→ show that \( \neg((\neg A \lor B) \land A) \) is not satisfiable
→ show that whatever valuation I pick, \( \nu((\neg((\neg A \lor B) \land A)) = \text{False} \)

\[
[\neg((\neg A \lor B) \land A)] \\
\text{______________} \\
\text{...} \\
\text{____________________} \\
(1) [\neg A, B] \\
(2) [A]
\]
Proof by resolution

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

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

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

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

Let $\nu$ be a valuation.
Proof by resolution

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

\begin{align*}
(1) & \ [\neg A, B] \\
(2) & \ [A]
\end{align*}

Let $\nu$ be a valuation.

- If $\nu(A) = \text{True}$, $\nu((1)) = \nu(B)$ and $\nu((2)) = \text{True}$, so the valuation of the whole thing is $\nu(B)$. 
Proof by resolution

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

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

Let \( v \) be a valuation.

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

- If \( v(A) = \text{False} \), \( v((1)) = \text{True} \) and \( v((2)) = \text{False} \) so the valuation of the whole thing is False.
Proof by resolution

Goal: show that whatever valuation I pick, $\nu((-A \lor B) \land A)) = False$

(1) $[-A, B]$
(2) $[A]$

Let $\nu$ be a valuation.

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

- If $\nu(A) = False$, $\nu((1)) = True$ and $\nu((2)) = False$ so the valuation of the whole thing is False.

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

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 \(\land\) and \(\lor\)

\[
(((\forall x) A \supset B) \equiv (((\forall x) A) \lor B) \\
\equiv (((\exists x)(\neg A)) \lor B) \\
\equiv (\exists x)(\neg A \lor B) \\
\equiv (\exists x)(A \supset B)
\]
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

1. Some more logic

2. Symbolic vs statistical machine learning
   - Knowledge
   - Explanations
   - Anomalies
   - Mechanics

3. Symbolic machine learning
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

1. Some more logic

2. Symbolic vs statistical machine learning
   - Knowledge
   - Explanations
   - Anomalies
   - Mechanics

3. Symbolic machine learning
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
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

1. Some more logic

2. Symbolic vs statistical machine learning
   - Knowledge
   - Explanations
   - Anomalies
   - Mechanics

3. Symbolic machine learning
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
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?
Al-generated images

Which image is Al-generated?

→ there are anomalies

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

1. Some more logic

2. Symbolic vs statistical machine learning
   - Knowledge
   - Explanations
   - Anomalies
   - Mechanics

3. Symbolic machine learning
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          Marie-Jean Meurs 6

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

1. Some more logic

2. Symbolic vs statistical machine learning

3. Symbolic machine learning
   - Reinforcement learning
   - Analogies
   - Inductive logic programming
   - Machine learning as compression
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
Symbolic machine learning
Reinforcement learning

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

\[
\begin{align*}
\text{ghi} & \rightarrow \text{ghj} \\
\text{uuvvw} & \rightarrow 
\end{align*}
\]
Analogies

芙hi → ghj
uuvvww → uuvvxx
uuvvjj
uuvvwx
ghj
uuvvwx
uuvvj
uuvvww
uuvvwj
error
Analogies

- On-the-fly learning of rules
- Many tasks are a form of analogy
  - solve → solves, get → ?
  - rosa → rosam, vita → ?
  - orang → orang-orang, burung → ?

- Analogies are highly discrete, but may be approximated by continuous representations, e.g. word embeddings

---

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

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
Deduction vs induction

Deduction: rules $\rightarrow$ conclusions (Prolog)

Induction: conclusions $\rightarrow$ 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

\[
\text{cute}(X) :\neg \text{dog}(X), \text{small}(X), \text{fluffy}(X). \quad (1) \\
\text{cute}(X) :\neg \text{cat}(X), \text{fluffy}(X). \quad (2) \\
\]

*Least-general generalization* of (1) and (2): \text{cute}(X) :\neg \text{fluffy}(X).

\[
\text{pet}(X) :\neg \text{dog}(X). \quad (3) \\
\text{pet}(X) :\neg \text{cat}(X). \quad (4) \\
\text{small}(X) :\neg \text{cat}(X). \quad (5) \\
\text{tame}(X) :\neg \text{pet}(X). \quad (6) \\
\]
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 $\neq i$
- 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

1. Some more logic

2. Symbolic vs statistical machine learning

3. Symbolic machine learning
   - Reinforcement learning
   - Analogies
   - Inductive logic programming
   - Machine learning as compression
Tycho Brahe

1546 - 1601

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

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

1643 - 1727

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.

\[ \frac{d\vec{p}}{dt} = \sum_i \vec{F}_i \]

\[ \vec{F}_{A\rightarrow B} = -\vec{F}_{B\rightarrow A} \] and \[ \vec{F}_{A\rightarrow B} \cdot \vec{AB} = 0 \]

Compression

Reality
Compression

Reality

Observations
Compression

Reality

Observations

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

Empirical laws

Tobias Stöger
Compression

Reality

Observations

Empirical laws

Principles

\[ \frac{d \rho}{dt} = \sum_i \vec{F}_i \]

\[ \frac{T^2}{a^3} = \text{constant} \]
Compression

Reality

Observations

Empirical laws

Principles

\[ \frac{d\bar{p}}{dt} = \sum_i \bar{F}_i \]

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

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

\[ DL(data) = DL(model) + DL(data|model) \]
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

\[ DL(\text{data}) = DL(\text{model}) + DL(\text{data}|\text{model}) \]