Machine Learning

CSE 142

Xin (Eric) Wang

Wednesday, October 13, 2021

Today

- Concept learning, Ch. 4

Concept learning

- Concept learning means learning (typically binary) concepts from examples
  - The learned concept is the positive class
  - Everything else is the negative class

- We'll now use logical models – logical expressions describe concepts and divide the instance space appropriately
  - See "Background 4.1" on p. 105 in the textbook for an overview of the logical concepts and notation
  - Propositional logic
    - Logical manipulation of propositions (symbols that have values)
  - (First-order) predicate logic
    - Add variables, predicates (binary functions), functions, and variable quantification (for all x…, there exists an x such that…)

Propositional (Boolean) logic

- Symbols represent propositions (statements of fact, sentences)
  - *P* means "San Francisco is the capital of California"
  - *Q* means "It is raining in Seattle"
  - *Length = 3* means "The value of the feature *Length* is 3"
  - *Teeth=many* means "The value of the feature *Teeth* is *many*

- Expressions are generated by combining proposition symbols with Boolean (logical) connectives
  - *True, false,* propositional symbols
  - *feature/value* relations – e.g., *feature* = *value*, *feature* < *value, …*
  - $\neg$ (not), $\wedge$ (and), $\vee$ (or), $\Rightarrow$ (implies), $\Leftrightarrow$ (equivalent)

Propositional logic

- Semantics
  - Defined by clearly interpreted symbols and straightforward application of truth tables
  - Rules for evaluating truth: Boolean algebra
  - Simple method: truth tables

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
False	*False*	*True*	*False*	*False*	*True*	*True*
False	*True*	*True*	*False*	*True*	*True*	*False*
True	*False*	*False*	*False*	*True*	*False*	*False*
True	*True*	*False*	*True*	*True*	*True*	*True*

2^N rows for N propositions

Simple Boolean logic

Commutative, associative, and distributive laws

$$P \wedge Q \Leftrightarrow Q \wedge P$$

$$(P \wedge Q) \wedge R \Leftrightarrow P \wedge (Q \wedge R)$$

$$P \wedge (Q \vee R) \Leftrightarrow (P \wedge Q) \vee (P \wedge R)$$

De Morgan's Laws

$$\neg\neg P \Leftrightarrow P$$

$$\neg(P \wedge Q) \Leftrightarrow \neg P \vee \neg Q$$

$$\neg(P \vee Q) \Leftrightarrow \neg P \wedge \neg Q$$

Propositional logic

- Propositional logic has simple syntax and semantics, and limited expressiveness
- However, it only has one representational device, the proposition, and cannot generalize
  - Input: facts; Output: facts
  - Result: Many, many rules are necessary to represent any non-trivial world
  - It is impractical for even very small worlds
- The solution?
  - **First-order logic**, which can represent propositions, objects, and relations between objects
  - Worlds can be modeled with many fewer rules

First-Order Logic (FOL)

- Also known as First-Order Predicate Calculus
  - Propositional logic is also known as Propositional Calculus

- An extension to propositional logic in which quantifiers can bind variables in sentences
  - Universal quantifier ($\forall$) – “For all…”
  - Existential quantifier ($\exists$) – “There exists…”
  - Variables: *x, y, z, a, joe, table*…

- Examples
  - $\forall x$ **Beautiful** (x)
  - $\exists x$ **Beautiful** (x)

Propositional logic vs. FOL

Extra

- Propositional logic:
  - **P** stands for "All men are mortal"
  - **Q** stands for "Socrates is a man"
  - What can you infer from P and Q?
    - Nothing!

- First-order logic:
  - $\forall x\ \mathbf{Man}(x) \Rightarrow \mathbf{Mortal}(x)$
  - $\mathbf{Man}(\mathit{Socrates})$
  - What can you infer from these?
    - $\mathbf{Mortal}(\mathit{Socrates})$

Concept learning

- In concept learning, we want to learn a Boolean function over a set of attributes+values
  - I.e., derive a Boolean function from training examples
    - Positive and negative examples of the concept
      - Positive: Temperature = high $\wedge$ Coughing = yes $\wedge$ Spots = yes
      - Negative: Temperature = medium $\wedge$ Coughing = no $\wedge$ Spots = yes
  - This is our hypothesis
- The target concept $\boldsymbol{c}$ is the true concept
  - We want the hypothesis to be a good estimate of the true concept
  - Thus we wish to find $\boldsymbol{h}$ (or $\hat{\boldsymbol{c}}$) such that $\boldsymbol{h} \approx \boldsymbol{c}$ (or $\hat{\boldsymbol{c}} \approx \boldsymbol{c}$)
- The hypothesis $\boldsymbol{h}$ is a Boolean function over the features
  - E.g., some combinations of {Temperature, Coughing, Spots} are <u>in</u> the concept, and others are <u>not in</u> the concept

The hypothesis space

- Using a set of features, what concepts can possibly be learned?
- The space of all possible concepts is called the hypothesis space
  - What is the hypothesis space for a given problem?
- First, how many possible instances are there for a given set of features?
  - All combinations of feature values
    - In set theory, the Cartesian product of all the features
      - $F_1 \times F_2 \times \ldots \times F_N$
  - UCSC courses: Quarter (4), Dept (40), courselevel (2), topic (500)
    - $4 \times 40 \times 2 \times 500 = 160{,}000$ possible instances
    - E.g., (fall, CSE, ugrad, ML), (spring, Music, grad, StringTheory), …

The hypothesis space

- The hypothesis space is the number of binary functions on these instances, which is… $2^{160,000}$!!
  - I.e., the number of all subsets you can make from 160,000 elements
  - Or if you laid out all possible instances, the number of different contours you could draw separating some instances from the rest
  - Each of these hypotheses… sets… contours… defines a concept
- The challenge in concept learning is deciding which hypothesis is best, given the training data
  - As with all problems in machine learning, generalization is of key importance – we don't only want to memorize the training data (the overfitting problem)
  - We want to learn a concept that will generalize well to new, unseen instances
- But even for a simple problem the hypothesis space is huge!

The conjunctive hypothesis space

- To make the problem tractable, we'll limit our hypothesis space to conjunctive concepts – i.e., hypotheses that can be expressed as a conjunction of literals (features)
  - Hypothesis: Quarter=? ∧ Dept=? ∧ courselevel=? ∧ topic=?
- We add "absence" or "don't care" to each feature, so now the total number of combinations is $5 \times 41 \times 3 \times 501 = 308{,}115$
  - That's a lot, but much better than $2^{160{,}000}$! (between 2^{18} and 2^{19})
- The most general conjunctive hypothesis is (X, X, X, X), which includes all possible instances
  - (fall, X, X, X) is the concept of all fall quarter courses
  - (fall, CSE, grad, X) is the concept of all CS graduate courses in the fall
- In this conjunctive hypothesis space, we can't represent concepts like "all courses in AI or Graphics"

An example hypothesis space

Target concept:
c = *Owns a house*

Two binary features:

		College student: Yes	College student: No
Age < 30	Yes	**A**	**B**
	No	**C**	**D**

Instance space:
{Age × Student}

$2 \times 2 = 4$ instances:
(Yes, Yes) – A (Yes, No) – B
(No, Yes) – C (No, No) – D

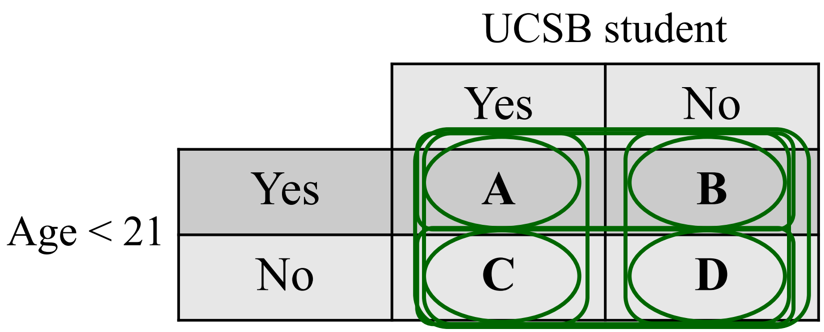
(1) How many possible hypotheses are there?
$2^4 = 16$ possible hypotheses (concepts)

(2) The training example (Yes, No) provides evidence for which hypotheses?
– All the ones that contain B

(3) Now what if we observe a second training example (No, No)?

{A}
★ {B}
{C}
{D}
★ {A, B}
{C, D}
{A, C}
★ {B, D}
{A, D}
★ {B, C}
★ {B, C, D}
{A, C, D}
★ {A, B, D}
★ {A, B, C}
★ {A, B, C, D}
{ } or ϕ

Our example using conjunctive hypothesis space



Instance space:
{Age × Student}

CHS: Hypotheses that can be represented as
Age={Yes, No, X} ∧ Student={Yes, No, X}

That's 9 hypotheses:

(Yes, Yes)	(Yes, No)	(No, Yes)
(No, No)	(Yes, X)	(No, X)
(X, Yes)	(X, No)	(X, X)

{A}	{A, D}
{B}	{B, C}
{C}	{B, C, D}
{D}	{A, C, D}
{A, B}	{A, B, D}
{C, D}	{A, B, C}
{A, C}	{A, B, C, D}
{B, D}	{ } or ϕ

Conjunctive = combining rows and columns via AND (not by OR)

An example of CHS learning

Suppose you come across a number of sea animals that you suspect belong to the same species. You observe their length in meters, whether they have gills, whether they have a prominent beak, and whether they have few or many teeth. The first animal can described by the following conjunction of features:

$$\text{Length} = 3 \;\wedge\; \text{Gills} = \text{no} \;\wedge\; \text{Beak} = \text{yes} \;\wedge\; \text{Teeth} = \text{many}$$

The next one has the same characteristics but is a meter longer, so you drop the length condition and generalize the conjunction to

$$\text{Gills} = \text{no} \;\wedge\; \text{Beak} = \text{yes} \;\wedge\; \text{Teeth} = \text{many}$$

The third animal is again 3 meters long, has a beak, no gills and few teeth, so your description becomes

$$\text{Gills} = \text{no} \;\wedge\; \text{Beak} = \text{yes}$$

All remaining animals satisfy this conjunction, and so your hypothesis is formed.

Someone tells you what these animals are called: Dolphins

An example of CHS learning

We took a specific-to-general approach in coming up with a hypothesis here.

Instances: Hypotheses:

(1) Length = 3 ∧ Gills = no ∧ Beak = yes ∧ Teeth = many

Length = 3 ∧ Gills = no ∧ Beak = yes ∧ Teeth = many

(2) Length = 4 ∧ Gills = no ∧ Beak = yes ∧ Teeth = many

Length = X ∧ Gills = no ∧ Beak = yes ∧ Teeth = many

(3) Length = 3 ∧ Gills = no ∧ Beak = yes ∧ Teeth = few

Length = X ∧ Gills = no ∧ Beak = yes ∧ Teeth = X

An example of CHS learning

Features and possible values:

Length = { 3, 4, 5 }

Gills = { yes, no }

Beak = { yes, no }

Teeth = { few, many }

In this problem, there are $3 \times 2 \times 2 \times 2 = 24$ possible instances and 2^{24} possible hypotheses over the instances (about 16.8 million)

But with the conjunctive hypothesis space, we have only $4 \times 3 \times 3 \times 3 = 108$ possible conjunctive hypotheses

- In our earlier example, we went from 16 hypotheses to 9 using CHS
- Here we go from 16.8 million to 108

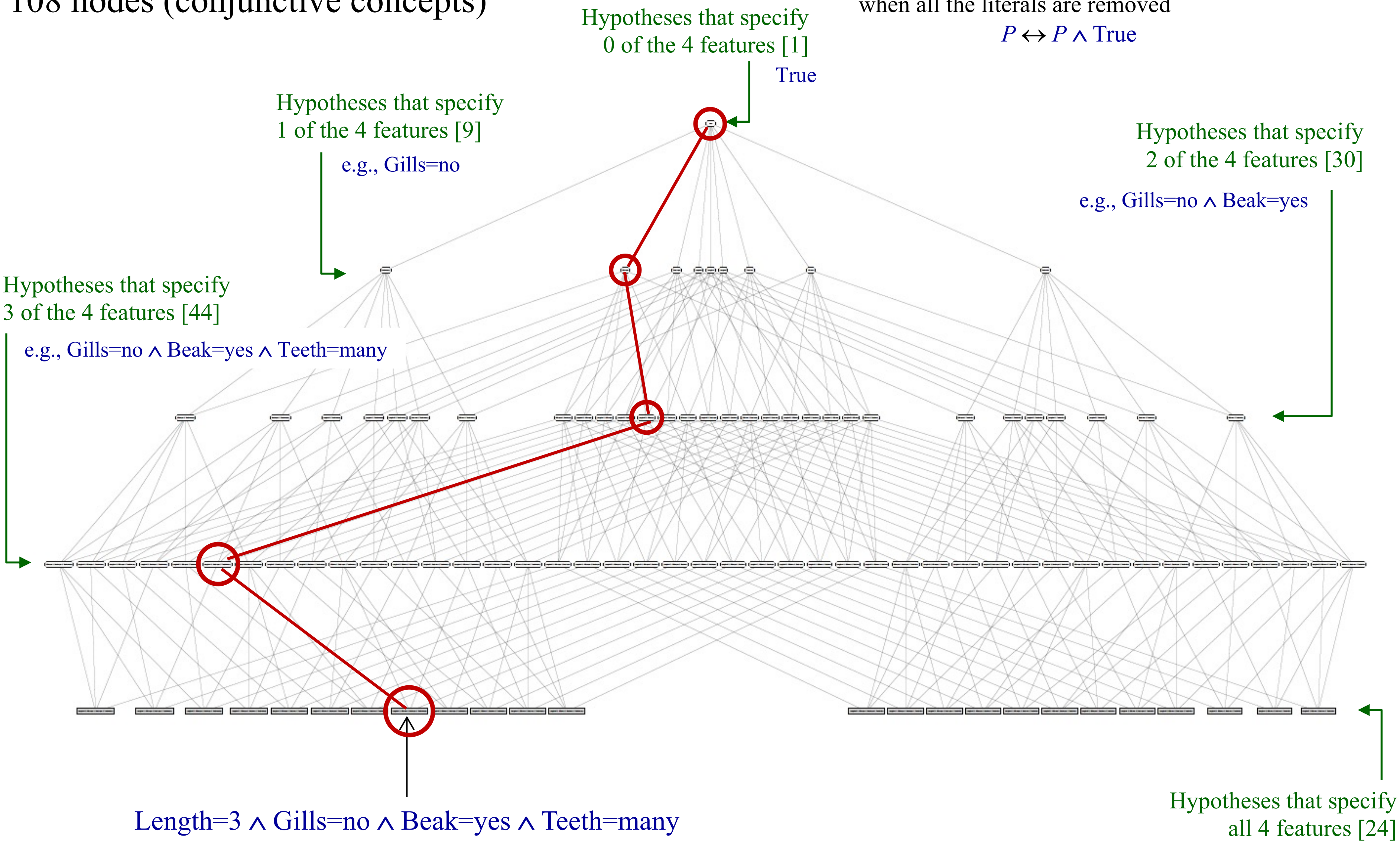
Let’s visualize the conjunctive hypothesis space for this problem:

The Conjunctive Hypothesis Space

108 nodes (conjunctive concepts)

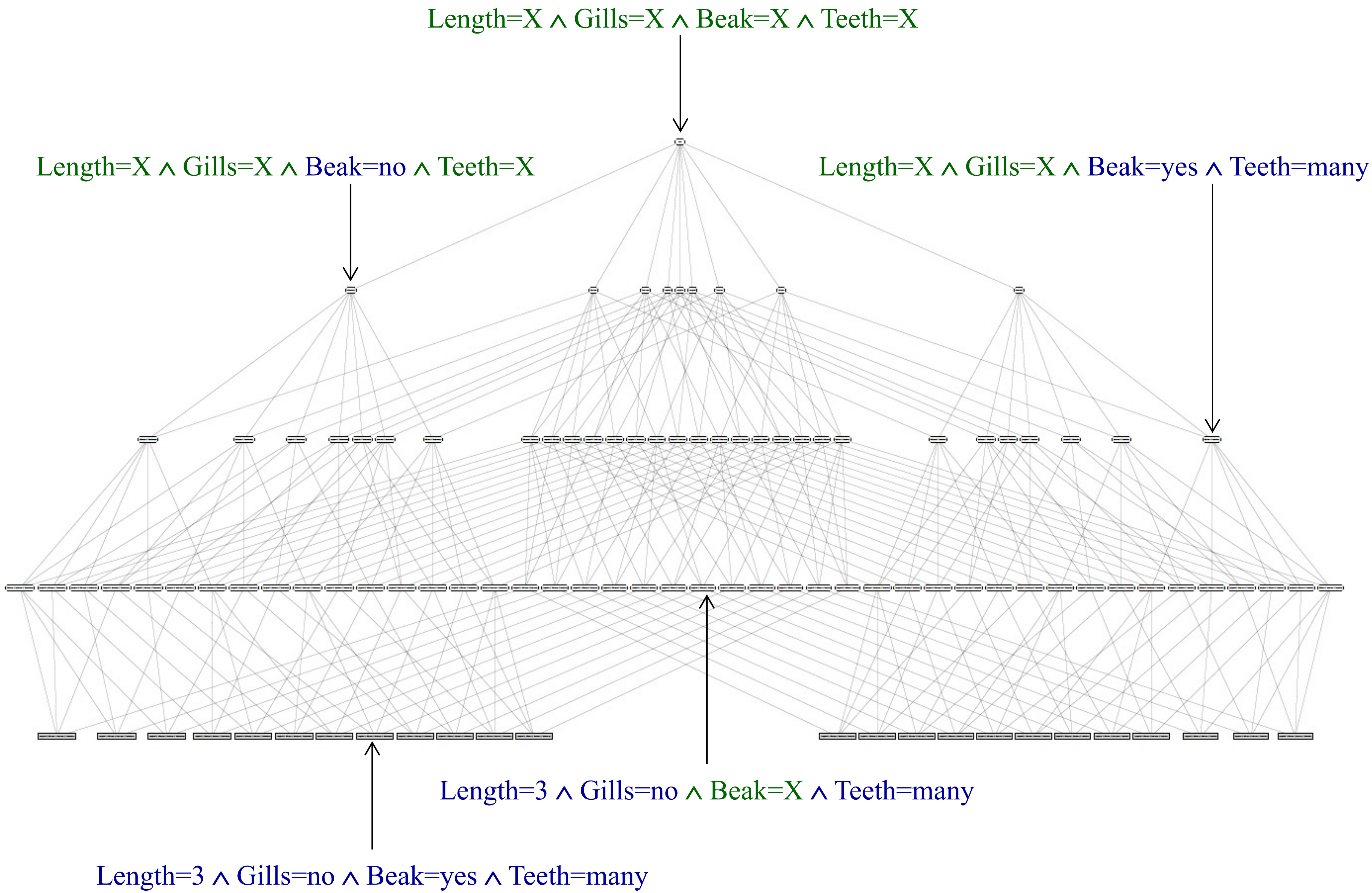
Note that ∧ True can be appended to <u>any</u> proposition without changing its truth value, so this is what's left when all the literals are removed

$$P \leftrightarrow P \wedge \text{True}$$



A node connects upward to every more general hypothesis that includes it.
A node connects downward to every more specific hypothesis that includes it.

The Conjunctive Hypothesis Space



Number of “Don’t care” (X) increases by 1 each level

The Conjunctive Hypothesis Space

Without writing the "Don't care"s