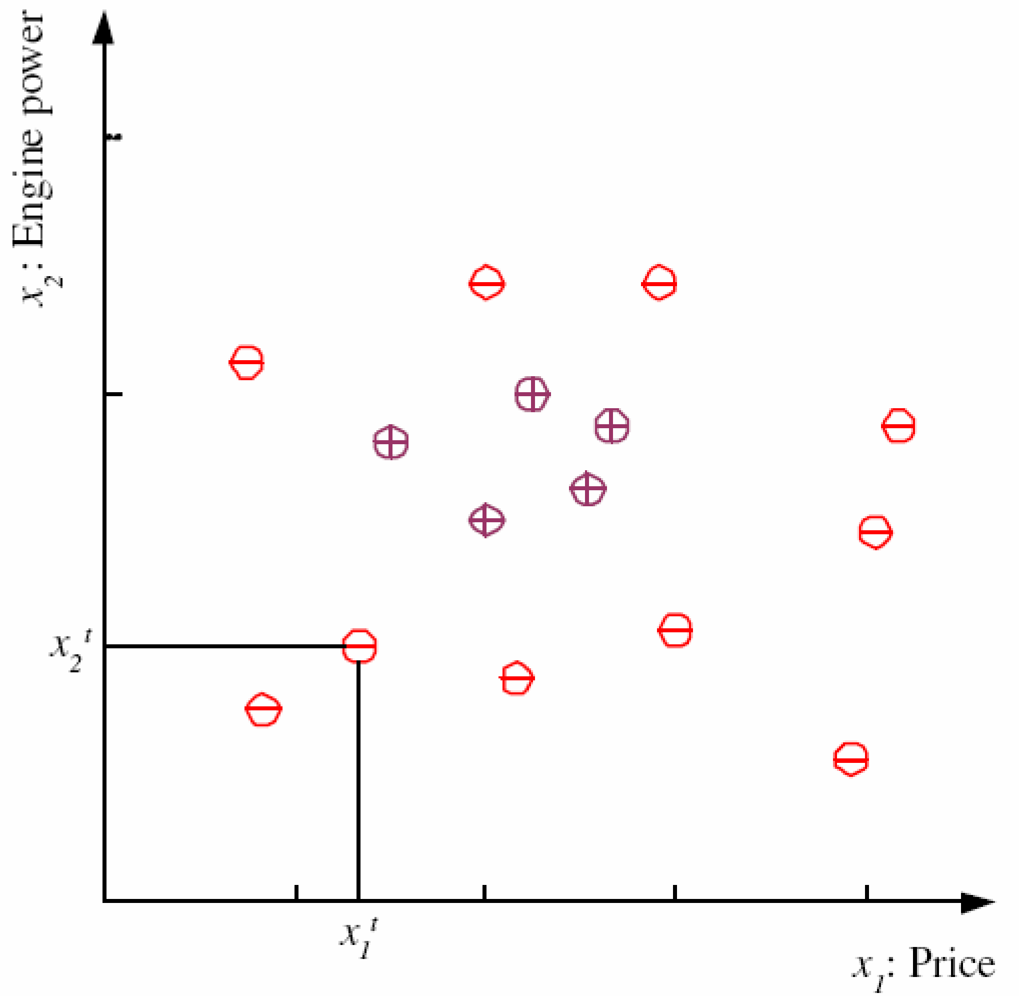
CS340:
Bayesian concept learning

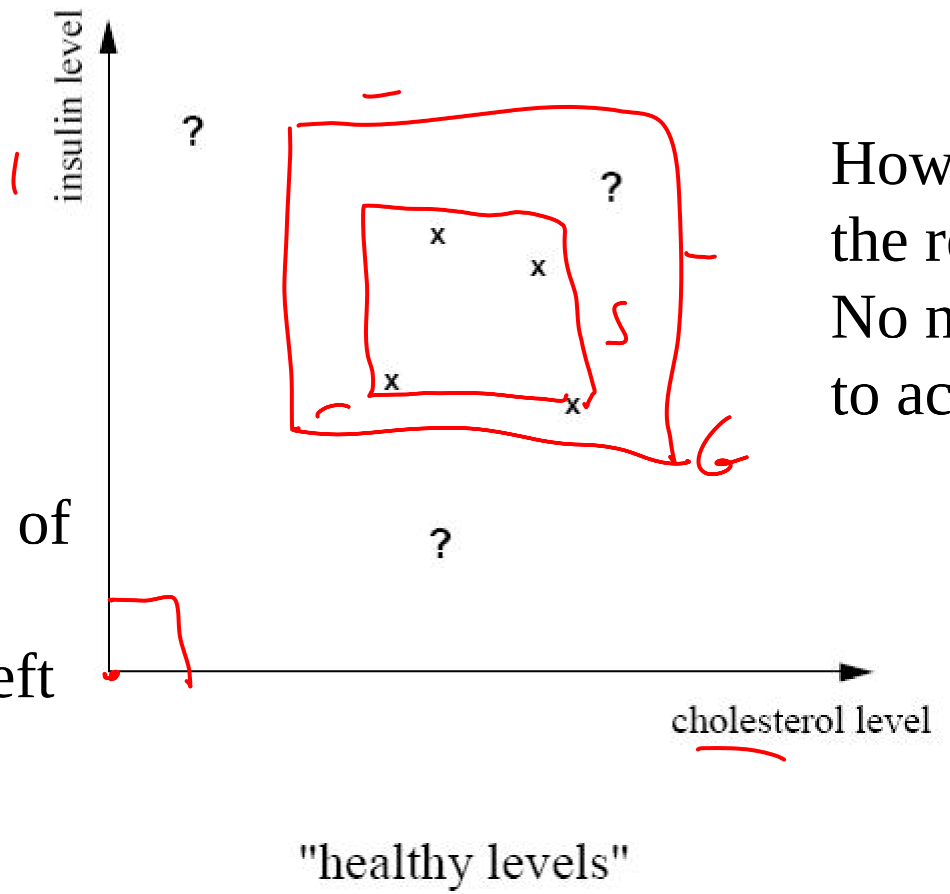
Kevin Murphy

Based on Josh Tenenbaum's PhD thesis (MIT BCS 1999)

Concept learning from positive and negative examples



Concept learning from positive only examples

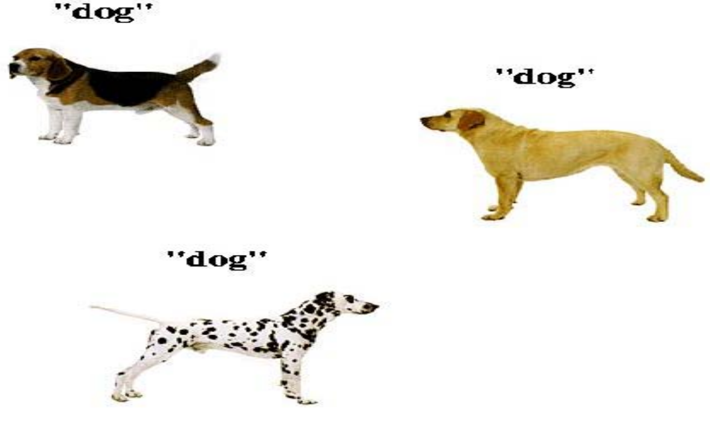


How far out should the rectangle go? No negative examples to act as an upper bound.

"Safe levels" of toxins would be in lower left

Human learning vs machine learning/ statistics

- Most ML methods for learning "concepts" such as "dog" require a large number of positive and negative examples
- But people can learn from small numbers of positive only examples (look at the doggy!)
- This is called "one shot learning"



Everyday inductive leaps

How can we learn so much about . . .

- Meanings of words
- Properties of natural kinds
- Future outcomes of a dynamic process
- Hidden causal properties of an object
- Causes of a person’s action (beliefs, goals)
- Causal laws governing a domain

. . . from such limited data?

The Challenge

- How do we generalize successfully from very limited data?
  - Just one or a few examples
  - Often only positive examples
- Philosophy:
  - Induction called a “problem”, a “riddle”, a “paradox”, a “scandal”, or a “myth”.
- Machine learning and statistics:
  - Focus on generalization from many examples, both positive and negative.

The solution: Bayes' rule

Posterior probability

Likelihood

Prior probability

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')} = p(d)$$

Sum over space of hypotheses

The origin of Bayes' rule

- A simple consequence of using probability to represent degrees of belief
- For any two random variables:

$$P(A \wedge B) = P(A)\, P(B \mid A)$$

$$P(A \wedge B) = P(B)\, P(A \mid B)$$

$$P(B)\, P(A \mid B) = P(A)\, P(B \mid A)$$

$$P(A \mid B) = \frac{P(A)\, P(B \mid A)}{P(B)}$$

$p(h|d) \propto p(h)p(d|h)$

Bayesian inference

- Bayes' rule: $P(H \mid D) = \dfrac{P(H)P(D \mid H)}{P(D)}$

- What makes a good scientific argument? $P(H|D)$ is high if:
  - Hypothesis is plausible: $P(H)$ is high
  - Hypothesis strongly predicts the observed data:

    $P(D|H)$ is high
  - Data are surprising: $P(D)$ is low

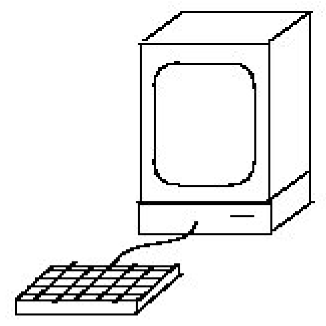
Bayesian inference: key ingredients

- Hypothesis space H
- Prior p(h)
- Likelihood p(D|h)
- Algorithm for computing posterior

post = normalize(lik .* prior)

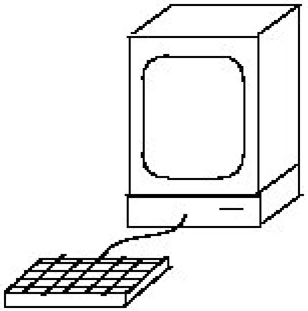
$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

The number game



- Program input: number between 1 and 100
- Program output: “yes” or “no”

The number game



- Learning task:
  - Observe one or more positive (“yes”) examples.
  - Judge whether other numbers are “yes” or “no”.

The number game

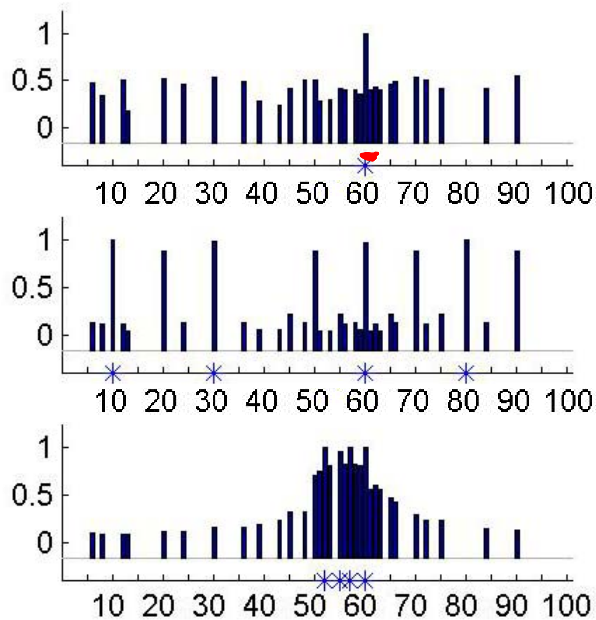
Examples of "yes" numbers	Hypotheses
60	multiples of 10 even numbers ? ? ?
60 80 10 30	multiples of 10 even numbers
60 63 56 59	numbers "near" 60

Human performance

60

60 80 10 30

60 52 57 55



Diffuse similarity

Rule:
"multiples of 10"

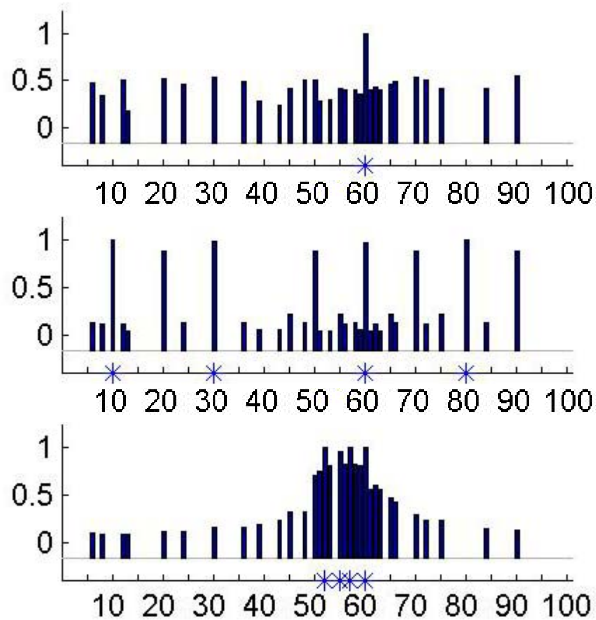
Focused similarity:
numbers near 50-60

Human performance

60

60 80 10 30

60 52 57 55



Diffuse similarity

Rule:
"multiples of 10"

Focused similarity:
numbers near 50-60

Some phenomena to explain:

- People can generalize from just positive examples.
- Generalization can appear either graded (uncertain) or all-or-none (confident).

Bayesian model

- H: Hypothesis space of possible concepts:
- $X = \{x_1, \ldots, x_n\}$: n examples of a concept C.
- Evaluate hypotheses given data using Bayes' rule:

$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{\sum_{h' \in H} p(X \mid h')\, p(h')}$$

  - $p(h)$ ["prior"]: domain knowledge, pre-existing biases
  - $p(X|h)$ ["likelihood"]: statistical information in examples.
  - $p(h|X)$ ["posterior"]: degree of belief that h is the true extension of C.

Hypothesis space

- Mathematical properties (~50):
  - odd, even, square, cube, prime, …
  - multiples of small integers
  - powers of small integers
  - same first (or last) digit
- Magnitude intervals (~5000):
  - all intervals of integers with endpoints between 1 and 100

Likelihood p(X|h)

- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as *n* increases.

$$p(X \mid h) = \left[\frac{1}{\mathrm{size}(h)}\right]^n \text{ if } x_1, \ldots, x_n \in h$$

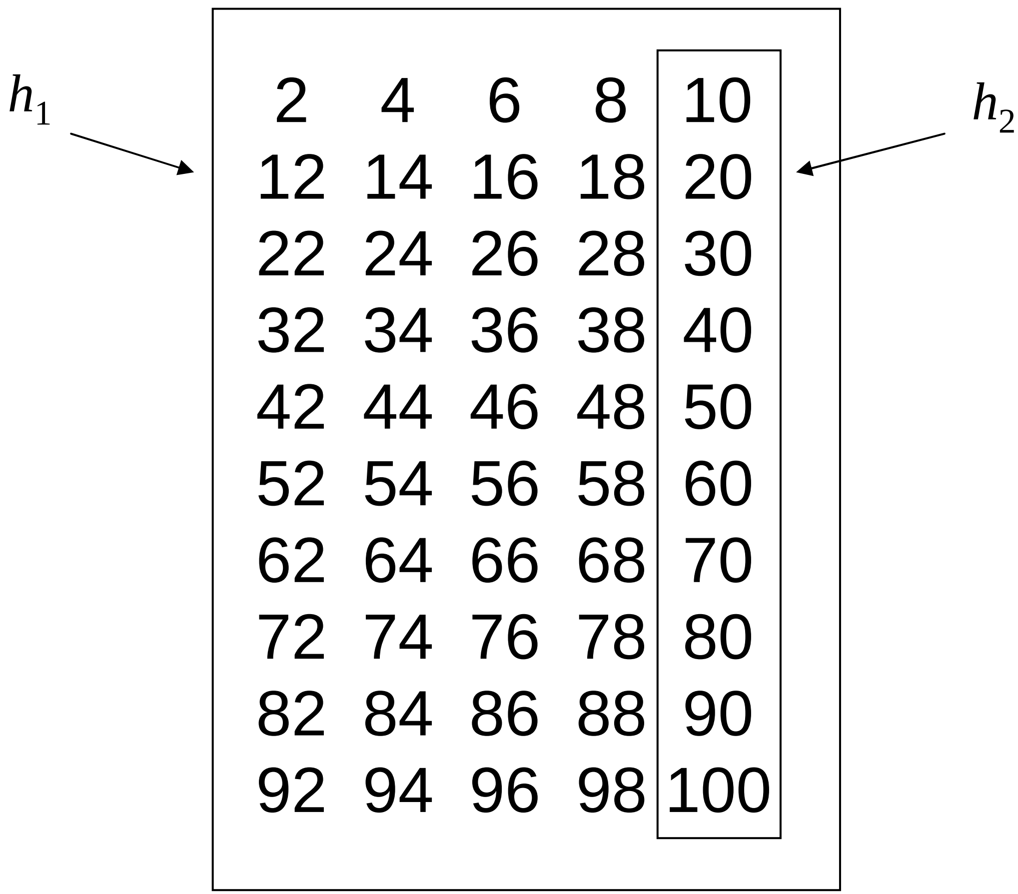
$$= 0 \text{ if any } x_i \notin h$$

- Follows from assumption of randomly sampled examples (**strong sampling**).
- Captures the intuition of a representative sample.

Example of likelihood

- $X=\{20,40,60\}$
- $H1$ = multiples of 10 = $\{10,20,\ldots,100\}$
- $H2$ = even numbers = $\{2,4,\ldots,100\}$
- $H3$ = odd numbers = $\{1,3,\ldots,99\}$
- $P(X|H1) = 1/10 * 1/10 * 1/10$
- $p(X|H2) = 1/50 * 1/50 * 1/50$
- $P(X|H3) = 0$

Illustrating the size principle



Illustrating the size principle

h_1

h_2

2	4	6	8	10
12	14	16	18	20
22	24	26	28	30
32	34	36	38	40
42	44	46	48	50
52	54	56	58	60
62	64	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

Data slightly more of a coincidence under h_1

Illustrating the size principle

h_1

h_2

2	4	6	8	10
12	14	16	18	20
22	24	26	28	30
32	34	36	38	40
42	44	46	48	50
52	54	56	58	60
62	64	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

Data *much* more of a coincidence under h_1

Prior p(h)

- X={60,80,10,30}
- Why prefer "multiples of 10" over "even numbers"?
  - Size principle (likelihood)
- Why prefer "multiples of 10" over "multiples of 10 except 50 and 20"?
  - Prior
- Cannot learn efficiently if we have a uniform prior over all 2^{100} logically possible hypotheses

Need for prior (inductive bias)

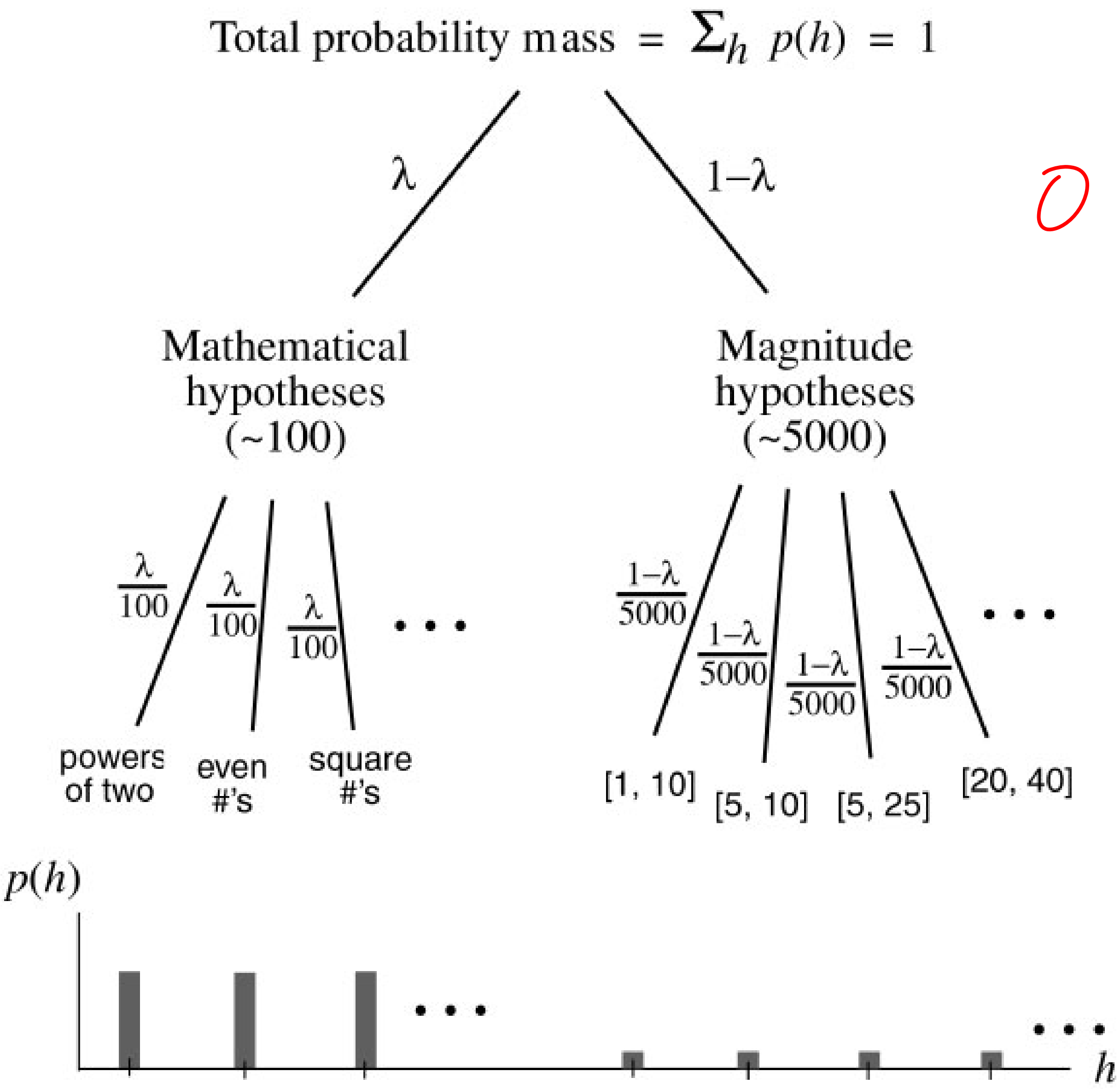
- Alpaydin p33
- Consider all $2^{2^2} = 16$ possible binary functions on 2 binary inputs

Boolean functions.

x_1	x_2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

- If we observe (x_1=0, x_2=1, y=0), this removes h_5, h_6, h_7, h_8, h_{13}, h_{14}, h_{15}, h_{16}
- Still leaves exponentially many hypotheses!

Hierarchical prior



$0 < \lambda < 1$

Computing the posterior

- In this talk, we will not worry about computational issues (we will perform brute force enumeration or derive analytical expressions).

$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{\sum_{h' \in H} p(X \mid h')\, p(h')}$$

Bayesian Occam’s Razor

- Which hypothesis is better supported by the examples {54, 6, 22}?
  - “even numbers”
  - “numbers between 6 and 54”
- Intuition: simpler hypotheses come from smaller (more constrained) hypothesis spaces.
  - “Entities should not be multiplied without necessity”
  - Prefer models with fewer free parameters.
- Both prior and likelihood contribute to this, since p(h|X) \propto p(h) p(X|h)

Minimum Description Length (MDL)

- Intuition: choose the hypothesis in terms of which the data is simplest/cheapest to encode.
- Basic information theory:
  - For a random variable X with distribution $P(X = x_i)$, the optimal code (shortest expected code length) uses

$$-\log P(X = x_i)$$

    bits to represent the proposition that $X = x_i$.
  - Examples:
    - Coding a uniform distribution over 1, …, 2^n
    - Alternatively: optimal strategy for playing “Twenty Questions”.

Relation between Bayes and MDL

- Bayesian inference:

$$P(h \mid X) \propto P(X \mid h)\, P(h)$$

- MDL inference:

$$-\log P(h \mid X) = -\log P(X \mid h) \quad + \quad -\log P(h) \quad + \quad \text{Const}$$

Cost to encode the data given the hypothesis

Cost to encode the hypothesis

MDL principle