Adapted from AIMA slides

Full Bayesian inference (Learning)

Peter Antal

antal@mit.bme.hu

Outline

- Learning paradigms
  - Learning as inference
  - Bayesian learning, full Bayesian inference, Bayesian model averaging
  - Model identification, maximum likelihood learning
- Probably Approximately Correct learning

Principles for induction

- Epicurus' (342? B.C. - 270 B.C.) principle of multiple explanations which states that one should *keep all hypotheses that are consistent with the data*.
- The principle of Occam's razor (1285 - 1349, sometimes spelt Ockham). Occam's razor states that when inferring causes *entities should not be multiplied beyond necessity*. This is widely understood to mean: Among all hypotheses consistent with the observations, choose the simplest. In terms of a prior distribution over hypotheses, this is the same as giving simpler hypotheses higher a priori probability, and more complex ones lower probability.

Bayesian inference with multiple models

Assume multiple models $M_i = (S_i, \theta_i)$ with prior $p(M_i)$ $i = 1, \ldots, M$.

The inference $p(Q = q|E = e)$ can be performed as follows:

$$p(q|e) = \Sigma_{i=1,\ldots,M} p(q, M_i|e) = \Sigma_{i=1,\ldots,M} p(q|M_i, e)p(M_i|e)$$

Note that $p(M_i|e)$ is a posterior over models with evidence e:

$$p(M_i|e) = \frac{p(e|M_i)p(M_i)}{p(e)} \propto p(e|M_i)p(M_i)$$

i.e., the evidence e reweight our beliefs in multiple models.

The inference is performed by **Bayesian Model Averaging** (BMA). Epicurus' (342(?) B.C. - 270 B.C.) **principle of multiple explanations** which states that one should keep all hypotheses that are consistent with the data.

Bayesian model averaging

Beside models, assume N multiple complete observations D_N.

The standard inference $p(Q = q|E = e, D_N)$ is defined as:

$$p(q|e, D_N) = \Sigma_{i=1,\ldots,M} p(q, M_i|e, D_N) = \Sigma_{i=1,\ldots,M} p(q|M_i, e, D_N) p(M_i|e, D_N)$$

Because $p(q|M_i, e, D_N) = p(q|M_i, e)$ and $p(M_i|e, D_N) \approx p(M_i|D_N)$:

$$p(q|e, D_N) \approx \Sigma_{i=1,\ldots,M} p(q|M_i, e) p(M_i|D_N)$$

where again $p(M_i|D_N)$ is a posterior after observations D_N:

$$p(M_i|D_N) = \frac{p(D_N|M_i)p(M_i)}{p(e)} \propto \underbrace{p(D_N|M_i)}_{likelihood} \underbrace{p(M_i)}_{prior}.$$

i.e., our rational foundation, probability theory, automatically includes and normatively defines learning from observations as standard Bayesian inference!

Full Bayesian learning

View learning as Bayesian updating of a probability distribution over the hypothesis space

H is the hypothesis variable, values $h_1, h_2, \ldots$, prior $\mathbf{P}(H)$ jth observation d_j gives the outcome of random variable D_j training data $\mathbf{d} = d_1, \ldots, d_N$

Given the data so far, each hypothesis has a posterior probability:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

where $P(\mathbf{d}|h_i)$ is called the likelihood

Predictions use a likelihood-weighted average over the hypotheses:

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i\, \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \Sigma_i\, \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!

Bayesian model averaging

View learning as Bayesian updating of a probability distribution over the hypothesis space

H is the hypothesis variable, values $h_1, h_2, \ldots$, prior $\mathbf{P}(H)$

jth observation d_j gives the outcome of random variable D_j
training data $\mathbf{d} = d_1, \ldots, d_N$

Given the data so far, each hypothesis has a posterior probability:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

where $P(\mathbf{d}|h_i)$ is called the likelihood

Predictions use a likelihood-weighted average over the hypotheses:

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i\, \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \Sigma_i\, \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!

Russel&Norvig: Artificial intelligence, ch.20

Bayesian Model Averaging example

Suppose there are five kinds of bags of candies:

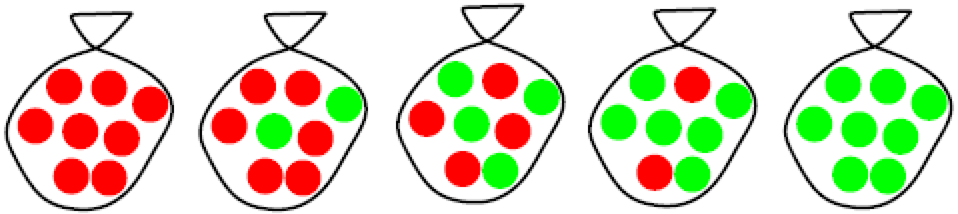
10% are h_1: 100% cherry candies

20% are h_2: 75% cherry candies + 25% lime candies

40% are h_3: 50% cherry candies + 50% lime candies

20% are h_4: 25% cherry candies + 75% lime candies

10% are h_5: 100% lime candies

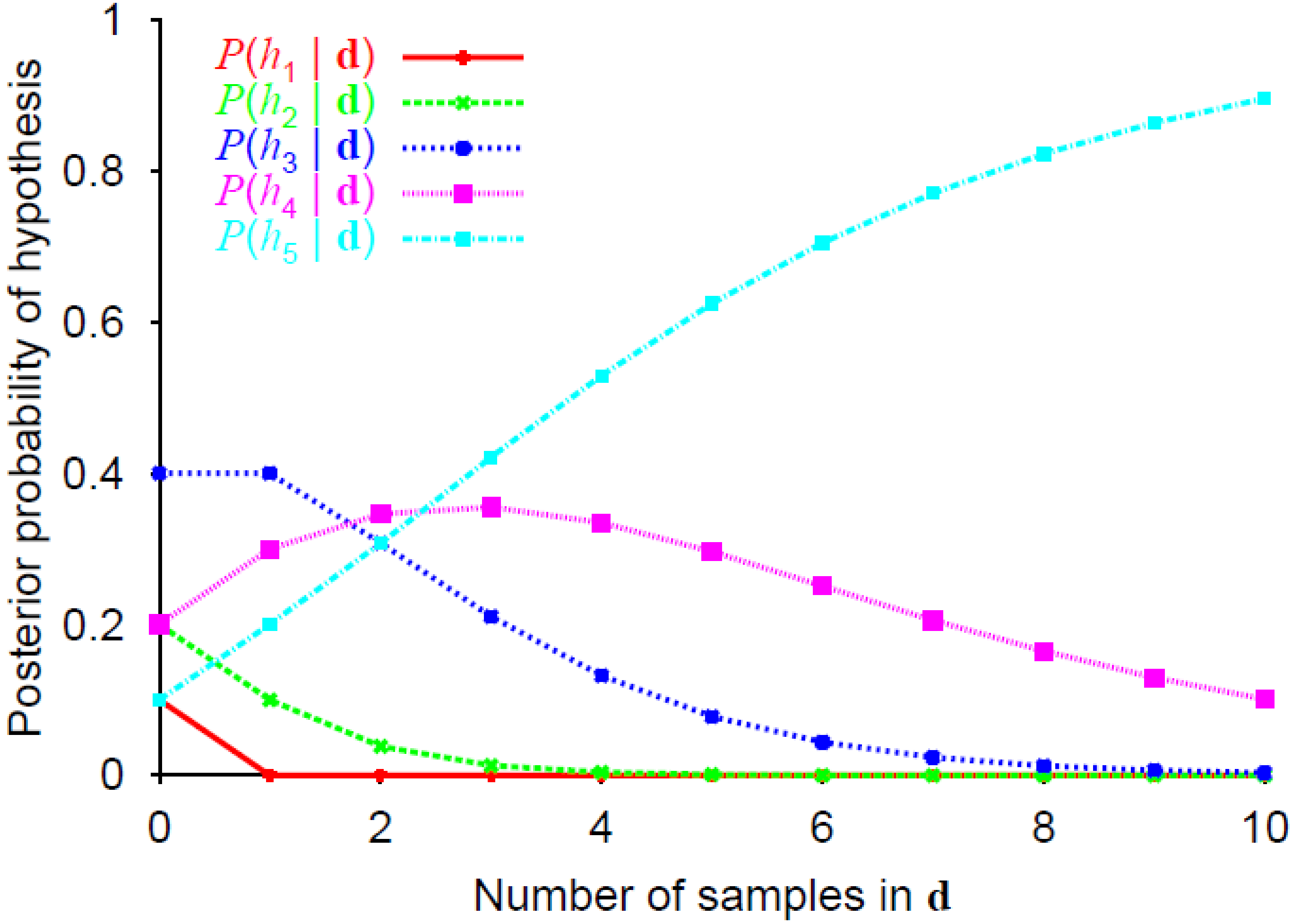


Then we observe candies drawn from some bag: ● ● ● ● ● ● ● ● ● ●

What kind of bag is it? What flavour will the next candy be?

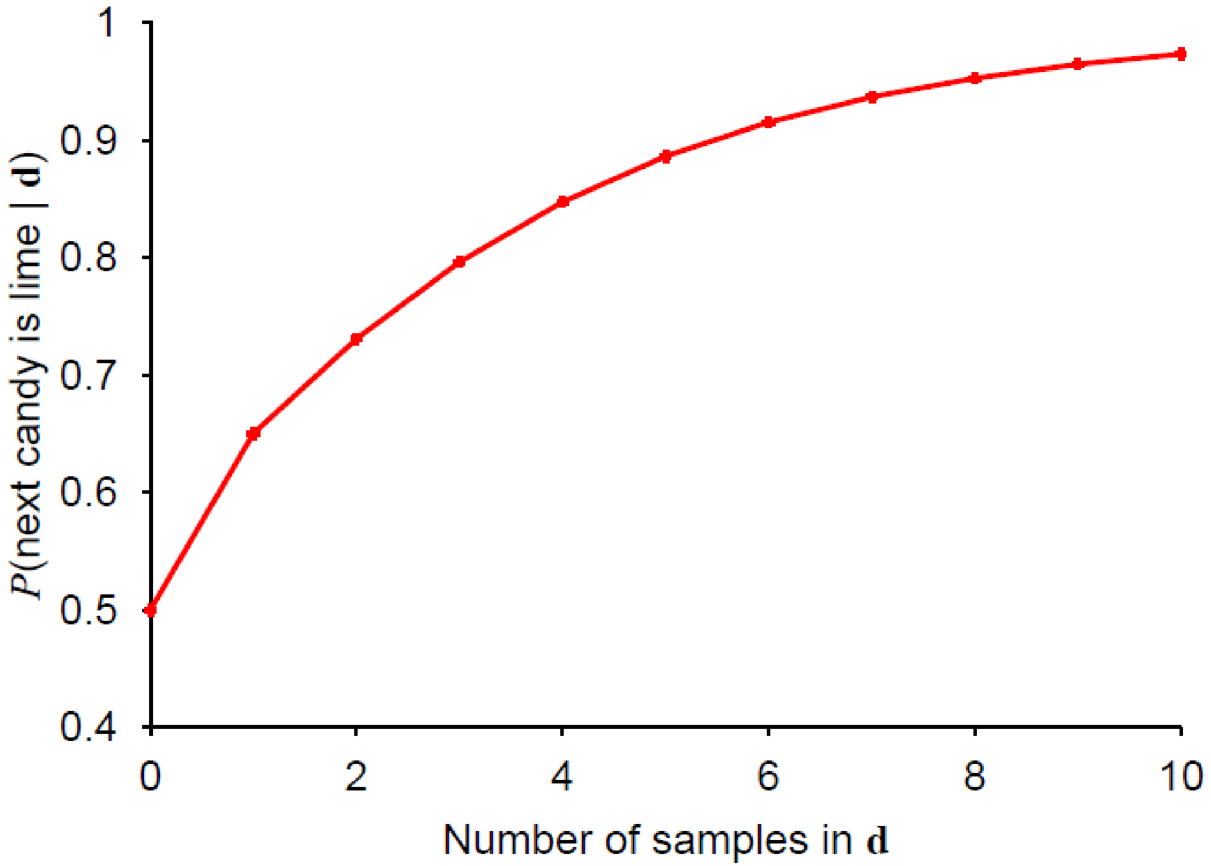
Russel&Norvig: Artificial intelligence

Learning rate for models



Russel&Norvig: Artificial intelligence

Learning rate for model predictions



Russel&Norvig: Artificial intelligence

MAP approximation

Summing over the hypothesis space is often intractable
(e.g., 18,446,744,073,709,551,616 Boolean functions of 6 attributes)

Maximum a posteriori (MAP) learning: choose h_{MAP} maximizing $P(h_i|\mathbf{d})$

I.e., maximize $P(\mathbf{d}|h_i)P(h_i)$ or $\log P(\mathbf{d}|h_i) + \log P(h_i)$

Log terms can be viewed as (negative of)
bits to encode data given hypothesis + bits to encode hypothesis
This is the basic idea of minimum description length (MDL) learning

For deterministic hypotheses, $P(\mathbf{d}|h_i)$ is 1 if consistent, 0 otherwise
⇒ MAP = simplest consistent hypothesis (cf. science)

ML approximation

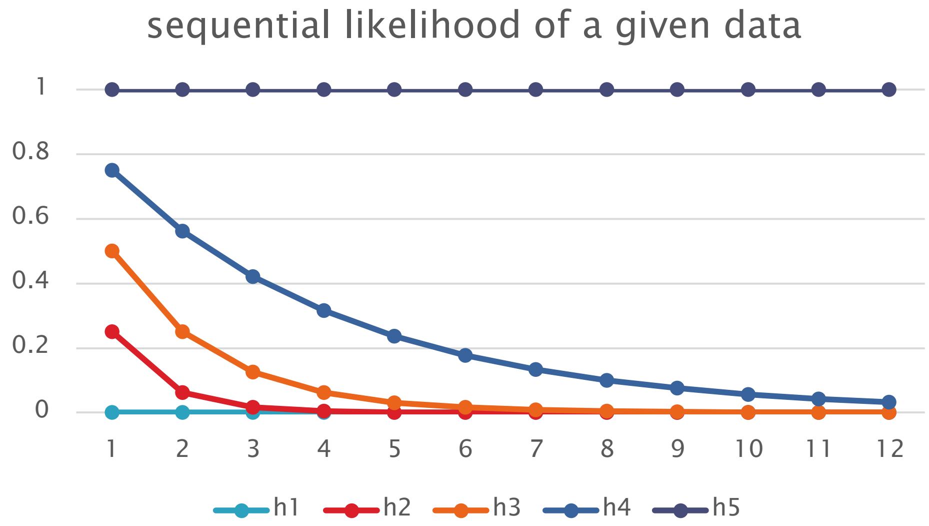
For large data sets, prior becomes irrelevant

Maximum likelihood (ML) learning: choose h_{ML} maximizing $P(\mathbf{d}|h_i)$

I.e., simply get the best fit to the data; identical to MAP for uniform prior
(which is reasonable if all hypotheses are of the same complexity)

ML is the "standard" (non-Bayesian) statistical learning method

Maximum likelood model selection



Inductive learning

- Simplest form: learn a function from examples

f is the target function

An example is a pair (*x*, *f(x)*)

Problem: find a hypothesis *h*
such that $h \approx f$
given a training set of examples

(This is a highly simplified model of real learning:

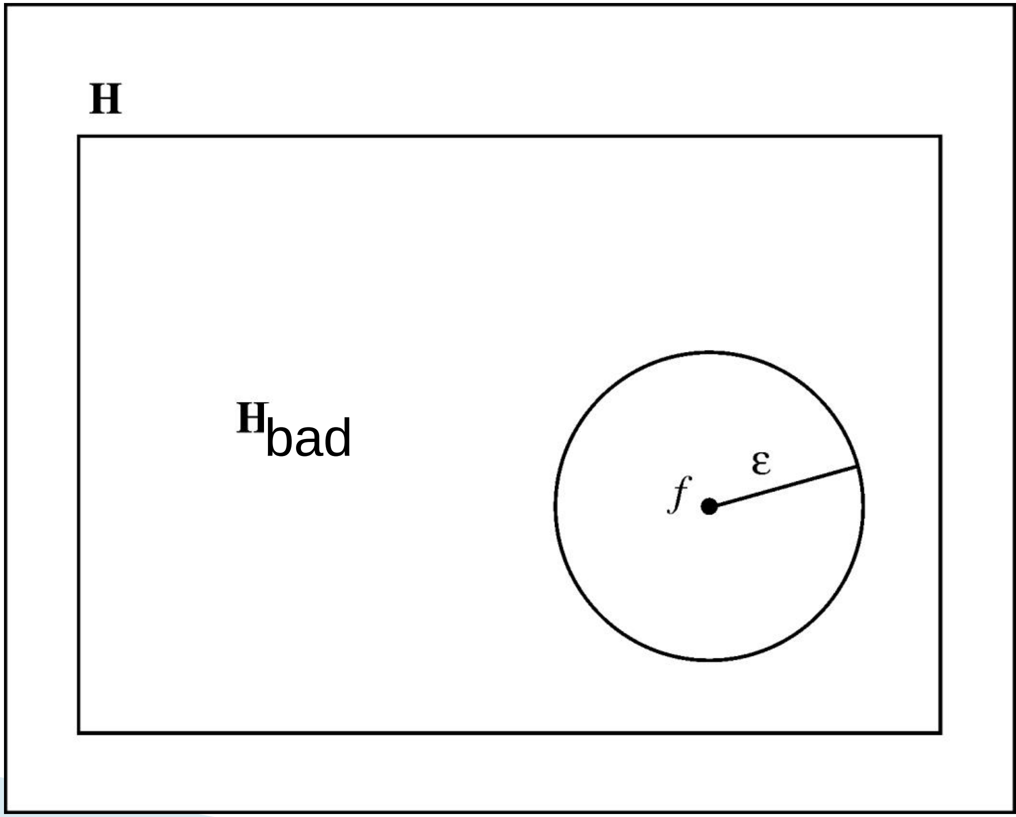
- Ignores prior knowledge
- Assumes examples are given)

The Probably Approximately Correct PAC-learning

A single estimate of the expected error for a given hypothesis is convergent, but can we estimate the errors for all hypotheses uniformly well??

Example from concept learning

X: i.i.d. samples.
n: sample size
H: hypotheses



Assume that the true hypothesis f is element of the hypothesis space **H.**

Define the error of a hypothesis h as its misclassification rate:

$$error(h) = p(h(x) \neq f(x))$$

Hypothesis h is **approximately correct** if

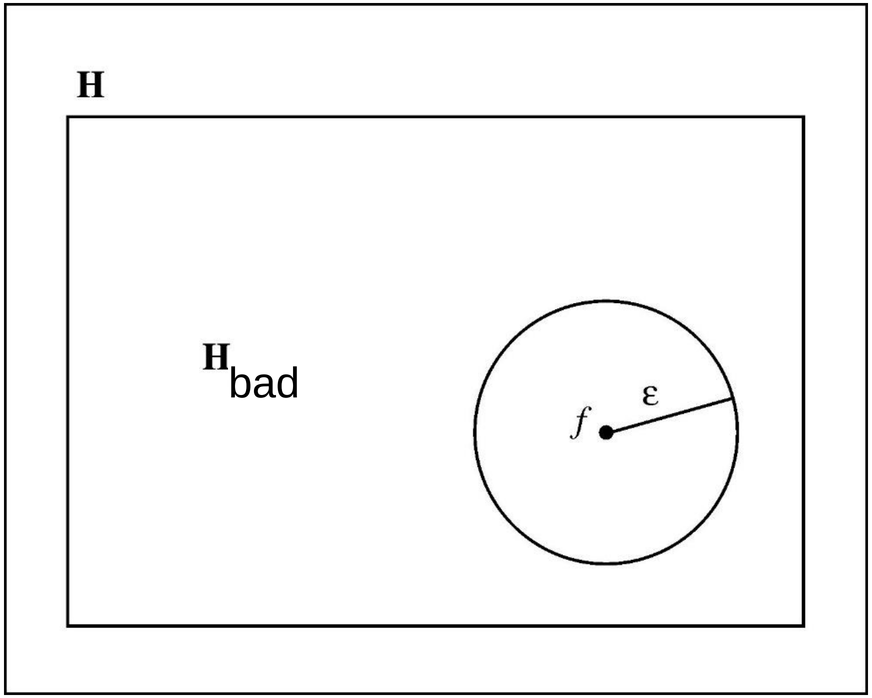
$$error(h) < \varepsilon$$

(ϵ is the “accuracy”)

For h∈H_{bad}

$$error(h) > \varepsilon$$

H can be separated to $H_{<\epsilon}$ and H_{bad} as $H_{\epsilon<}$



By definition for any h ∈ H_{bad}, the probability of error is larger than ε

➔thus the probability of no error is less than $\leq (1-\varepsilon)$

Thus for m samples for a $h_b \in H_{bad}$:

$$p\left(D_n\!:\!h_b(x) = f(x)\right) \leq (1-\varepsilon)^n$$

For any $h_b \in H_{bad}$, this can be bounded as

$$p\left(D_n\!:\!\exists h_b \in H, h_b\ (x) = f(x)\right) \leq$$
$$\leq |H_{bad}|(1-\varepsilon)^n$$
$$\leq |H|\ (1-\varepsilon)^n$$

To have at least δ “probability” of approximate correctness:

$$|H|\,(1-\varepsilon)^{n} \le \delta$$

By expressing the sample size as function of *ϵ accuracy* and *δ confidence* we get a bound for ***sample complexity***

$$1/\varepsilon(\ln|H| + \ln\left(\frac{1}{\delta}\right)) \le n$$

Hypothesis spaces

How many distinct concepts/decision trees with *n* Boolean attributes?

= number of Boolean functions

= number of distinct truth tables with 2^n rows = 2^{2^n}

- E.g., with 6 Boolean attributes, there are 18,446,744,073,709,551,616 trees

Summary

- Normative predictive probabilistic inference
  - performs Bayesian model averaging
  - implements learning through model posteriors
  - avoids model identification
- Model identification is hard
  - Probably Approximately Correct learning