Scribe #4: Bayesian Learning

Jiang Ye, Kam Sing Leung, Byron Gao

February 19, 2004

Contents

1 Introduction **2**
- 1.1 Two Theories of Learning . . . 2
- 1.2 Bayesian Statistics . . . 2
- 1.3 Two Roles for Bayesian Methods . . . 3

2 Basics about Probability Theory **3**
- 2.1 Terms . . . 3
- 2.2 A More Abstract View of Probability . . . 4
- 2.3 Conditional Probability . . . 4
- 2.4 Basic Formulas for Probabilities . . . 4

3 Understanding Bayes Theorem **6**
- 3.1 Prior and Posterior Probabilities . . . 6
- 3.2 Bayes Theorem . . . 6
- 3.3 Proving Bayes Theorem . . . 7

4 Choosing Hypothesis **7**
- 4.1 MAP and ML Hypotheses . . . 7
- 4.2 An Illustrative Example . . . 8
- 4.3 Updating Probabilities . . . 9

5 Relation to Concept Learning **11**
- 5.1 Defining the Algorithms . . . 11
- 5.2 ConsistFilter vs. MAP-learner . . . 12
- 5.3 Find-S vs. MAP-learner . . . 13
- 5.4 Characterizing Learning Algorithms by Equivalent MAP-learners . . . 13

6 Minimum Description Length **14**
- 6.1 Some Background: Entropy Encoding . . . 15
- 6.2 The MDL Principle . . . 15
- 6.3 Understanding MDL . . . 16
- 6.4 Some Observations about MDL . . . 17

7 Probabilistic Views of Learning **17**

8 References **18**

1. Introduction

1.1 Two theories of learning

Computational Learning Theory

- Study the design and analysis of algorithms for making predictions about the future based on past experiences
- Emphasis is on rigorous mathematical analysis
- Used mainly by computer scientists

Statistics

- Science and practice of developing human knowledge through the use of empirical data
- Aim is to produce the "best" information from available data
- Used in much more general culture: business, sociology, manufacturing…

1.2 Bayesian statistics

Thomas Bayes, an English mathematician, was the first to use probability assessments inductively, i.e. calculating the probability of a new event on the basis of earlier probability estimates which have been derived from empirical data. Bayes set down his ideas on probability in "Essay Towards Solving a Problem in the Doctrine of Chances". This work became the basis of a statistical technique, now called Baysian statistics.

Brief history of statistics

- Bayesian philosophy developed in late 18^{th} century
- Classical philosophy formalized in early 20^{th} century and quickly became dominant
- Revival of Bayesian statistics in late 20^{th} century due largely to computational advances (Markov Chain Monte Carlo software, etc). The applications of Bayesian statistics in industry are countless.

Bayesian statistics vs. classical statistics

- Bayesian statistics and classical statistics are different ways of doing statistical analyses
- The key difference is that Bayesian methods require specification of prior knowledge which is updated through further observation to obtain posterior knowledge, while classical statistics does not assume we have the prior

- If the prior is not known, it has to be estimated using background knowledge. Different individuals may estimate differently. Thus, classical statisticians argue that Bayesian methods suffer from a lack of objectivity

- Bayesians argue back that the classical methods of statistical inference have built-in subjectivity (through the choice of a sampling plan and the assumption of "randomness" of distributions) and that an advantage of the Bayesian approach is that the subjectivity is made explicit

- They are still fighting…

1.3 Two Roles for Bayesian Methods

Provides practical learning algorithms:
- Naïve Bayes learning
- Bayesian belief network learning
- Combine prior knowledge (prior probabilities) with observed data

Provides useful conceptual framework
- Provides "gold standard" for evaluating other learning algorithms
- Provides additional insight into Occam's razor

Aside: about Occam's razor
Occam's razor is a logical principle stating that one should not make more assumptions than minimum needed: "Of two competing theories or explanations, all other things being equal, the simpler one is to be preferred." It underlies all scientific modeling and theory building.

2. Basics About Probability Theory

2.1 Terms

Random Variable
In the context of machine learning, we can think of random variable as some attribute that can take some values.

E.g. weather $\in$ {Sunny, Rain, Cloudy, Snow}

Mathematically, a random variable is defined as a measurable function from a probability space to some measurable space. This measurable space is the space of possible values of the variable, and it is usually taken to be the real numbers with the σ algebra. (more explanations in the following "Aside" part)

Domain
Set of possible values that a random variable can take. It could be finite or infinite. E.g. all conjunctions; all Boolean functions; all functions from R^4 to {0,1}, all R^n to R.

Probability Distribution
Mapping from domain to values in [0…1].
P(weather) = (0.7, 0.2, 0.08, 0.02) means
P(weather = Sunny) = 0.7
P(weather = Rain) = 0.2
P(weather = Cloudy) = 0.08
P(weather = Snow) = 0.02

Event
Each assignment of a domain value to a random variable is an "event".
e.g. weather = Rain

2.2 A more abstract view of probability

Probability theory can be viewed as the study of probability spaces and random variables. A probability space is a triple (Ω, F, P), where

- Ω is a non-empty set, sometimes called the "sample space". Each of its members is thought to be a potential outcome of a random experiment.
- F is a sigma-algebra of subsets of Ω. Its members are called "events". To say that F is a sigma-algebra necessarily implies that the complement of any event is an event, and the countable union of any sequence of events is an event, thus any countable intersection is also an event.
- P is a probability measure on F

A random variable is a measurable function on Ω.

2.3 Conditional Probability

P (A | B) = Probability of event A, given that event B has happened
E.g. P (Cavity | Toothache) = 0.8, meaning that 80% of toothache cases are due to cavity

In general,

$$P (A | B) = P (A \wedge B) / P (B)$$

2.4 Basic Formulas for Probabilities

Product Rule: probability of conjunction of events A and B:

$$P (A \wedge B) = P (A | B) * P (B) = P (B | A) * P (A)$$

Sum Rule: probability of disjunction of events A and B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Theorem of total probability:

If events $A_1, \ldots A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i)$$

The intention here is that we break it down to conditional probabilities which are easier to estimate.

3 Understanding Bayes Theorem

3.1 Prior and Posterior Probabilities

- The unconditional (prior) probability of an event is the probability of the event before evidence is presented.
  - For example, $P(cavity) = 0.01$ means that the probability that someone (from this population) has a cavity is 1 in 100.
- Evidence is the percept that affects the degree of belief in an event.
  - Toothache is an evidence for someone's having a cavity.
- The conditional (posterior) probability of an event is the probability of the event after evidence is presented.
  - $P(cavity|toothache) = 0.8$. (Note that posterior probability can be *completely* different from prior probability.)
- In general, $P(A|B)$ is the probability of event A given that event B has happened. It can be defined as follows:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

  Alternatively, we have the following product rule:

$$P(A \wedge B) = P(A|B)P(B)$$

3.2 Bayes Theorem

Bayes theorem provides a way to calculate the probability of a hypothesis h from some space H, given the observed training data D:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$ is the prior probability of hypothesis h.
  - It is the initial probability of h before we have observed the training data.

  - It reflects any background knowledge we have about the chance that h is a correct hypothesis.
  - If we have no such prior knowledge, we might simply assign the same prior probability to each candidate hypothesis.

- $P(D)$, the prior probability of training data D, is the probability of D given no knowledge about which hypothesis holds.
- $P(D|h)$, the likelihood of training data D given hypothesis h, is the probability of observing D given some world in which h holds.
- $P(h|D)$ is the posterior probability of h.
  - It is the probability that h holds given the observed training data D.
  - It reflects the influence of D on our confidence (or degree of belief) that h holds after we have seen the data D. In comparison, the prior probability $P(h)$ is independent of D.

3.3 Proving Bayes Theorem

The proof is (embarrassingly) very simple. By the product rule, we have:

$$\begin{aligned} P(h \wedge D) &= P(h|D)P(D) \\ P(D \wedge h) &= P(D|h)P(h) \end{aligned}$$

But $P(h \wedge D) = P(D \wedge h)$. Thus,

$$\begin{aligned} P(h|D)P(D) &= P(D|h)P(h) \\ P(h|D) &= \frac{P(D|h)P(h)}{P(D)} \end{aligned}$$

An intuitive understanding of Bayes theorem: as one would expect, $P(h|D)$ increases with $P(h)$ and $P(D|h)$; however it is reasonable to see that $P(h|D)$ decreases with $P(D)$, because the more probable it is that D will be observed independently of h, the less evidence D provides in support of h.

4 Choosing Hypothesis

4.1 MAP and ML Hypotheses

We want to know which hypothesis (among candidate hypotheses) is the most probable, given the training data. In other words, our aim is to find the maximum

a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \operatorname*{argmax}_{h \in H} P(h|D) \\ &= \operatorname*{argmax}_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \operatorname*{argmax}_{h \in H} P(D|h)P(h) \end{aligned}$$

The justification for dropping $P(D)$ is that it is a constant independent of h.

If every hypothesis in space H is equiprobable (i.e., $P(h_i) = P(h_j)$), then the maximum *a posteriori* hypothesis is simply the hypothesis h that maximizes the likelihood $P(D|h)$ of the data given the hypothesis. We denote it by h_{ML}:

$$h_{ML} = \operatorname*{argmax}_{h \in H} P(D|h)$$

4.2 An Illustrative Example

Does the patient have cancer, or does he not? A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, 0.008 of the entire population have this cancer. To summarize:

$$\begin{array}{ll|ll} P(cancer) & = 0.008 & P(\neg cancer) & = 0.992 \\ P(\oplus|cancer) & = 0.98 & P(\ominus|cancer) & = 0.02 \\ P(\oplus|\neg cancer) & = 0.03 & P(\ominus|\neg cancer) & = 0.97 \end{array}$$

- To find the MAP hypothesis, we compute the following:

$$\begin{aligned} P(\oplus|cancer)P(cancer) &= 0.98 \times 0.008 = 0.0078 \\ P(\oplus|\neg cancer)P(\neg cancer) &= 0.03 \times 0.992 = 0.0298 \end{aligned}$$

  Thus h_{MAP} is $\neg cancer$, since $P(\oplus|\neg cancer)P(\neg cancer)$ is greater than $P(\oplus|cancer)P(cancer)$.

- For the exact posterior probabilities, we compute the following:

$$\begin{aligned} P(cancer|\oplus) &= \frac{P(\oplus|cancer)P(cancer)}{P(\oplus)} \\ P(\neg cancer|\oplus) &= \frac{P(\oplus|\neg cancer)P(\neg cancer)}{P(\oplus)} \end{aligned}$$

We already know what $P(\oplus|cancer)P(cancer)$, $P(\oplus|\neg cancer)P(\neg cancer)$, and $P(\oplus)$ are. What remains to be found out is $P(\oplus)$, the probability of positive test result:

$$\begin{aligned} P(\oplus) &= P(\oplus, cancer) + P(\oplus, \neg cancer) \\ &= P(\oplus|cancer)P(cancer) + P(\oplus|\neg cancer)P(\neg cancer) \\ &= 0.0078 + 0.0298 \end{aligned}$$

Thus,

$$\begin{aligned} P(cancer|\oplus) &= \frac{0.0078}{0.0078 + 0.0298} \\ &= 0.21 \end{aligned}$$

$$\begin{aligned} P(\neg cancer|\oplus) &= \frac{0.0298}{0.0078 + 0.0298} \\ &= 0.79 \end{aligned}$$

- Note that the posterior probabilities can also be determined by normalizing the quantities $P(\oplus|cancer)P(cancer)$ and $P(\oplus|\neg cancer)P(\neg cancer)$ (i.e., 0.0078 and 0.0298) so that they sum to 1.
- Although the posterior probability of cancer (given the positive test result) is significantly higher than its prior probability (0.21 compared to 0.008), it is still lower than the probability of the patient's not having cancer (0.79). This is due to the low prior probability assigned to cancer.
- What further action should the patient consider in light of the probability calculation?
  - The choice of action also depends on utilities (the agent's preferences between possible outcomes of the various plans). For example, if preserving life is much preferred to saving money or time, the patient will be well advised to take further action such as having a second test to confirm if he has cancer. The stake is so high that it would be irrational for the patient to do nothing and simply hope that chance is on his side.

4.3 Updating probabilities

- Every rational person should adopt prior probabilities that conform to the rules of probability theory.

  - Representation theorem (de Finetti, Ramsey, von Neumann, Savage): a rational person (in his or her choice of actions) can be modelled as if s/he has a *probability assignment* over possible states of the world, and a *utility function* on the outcomes of actions (which can be viewed as his or her goals or preferences).

- Upon learning evidence D, a rational person should update his or her prior probabilities accordingly:

$$\text{Bayes Rules:} \quad P_{t+1}(h) = P_t(h|D)$$

- Note that Bayes rules is not the same as Bayes theorem, which can be stated as:

$$\text{Bayes Theorem:} \quad P_t(h|D) = \frac{P_t(D|h)P_t(h)}{P_t(D)}$$

  - Bayes rule is a rule for updating probabilities over time, while Bayes theorem deals with concurrent probabilities (diachronic vs. synchronic).
  - Bayes rule is not a theorem. A person, in adopting the rule, may update his or her probabilities simply by following the opinion of experts without using Bayes theorem to calculate probabilities himself or herself.
  - Bayes theorem may deal with hypothetical situations: what the probability would have been if such were the case. On the other hand, Bayes rule always deals with the actual situation.

- From prior probability to learner:

  (1) Specify prior probability distribution P.

  (2) Given data D, update $P(h)$ by $P(h|D)$.

  (3) Output $\text{argmax}_{h \in H} P(h)$.

  It is arbitrarily hard to compute $P(h|D)$. Thus, the above is an algorithm only if step (2) is computable.

- From learner to prior probability: if the method is rational, it must be based on prior probability.

5. Relation to Concept Learning

In the following, we discuss the relationship between Bayes theorem and concept learning. We compare several algorithms discussed in earlier chapters, particularly ConsistFilter and Find-S, with "MAP-learner," a brute-force Bayes learning algorithm that outputs MAP hypotheses. As we shall see, under certain conditions, ConsistFilter and Find-S output MAP hypotheses.

5.1 Defining the Algorithms

A standard concept learning task is to learn some target concept c: $X \rightarrow \{0, 1\}$. Some related terms are:

- ✓ Instance space X containing instances $x_1, x_2, \ldots, x_m$
- ✓ Hypothesis / Concept space H containing hypothesis $h_1, h_2, \ldots, h_m$
- ✓ Set of training examples $D = \{<x_i, c(x_i)>\}$ where $x_i \in X$ and $c(x_i) \in H$

The following simplifying assumptions are made without altering the main conclusions of this section:

- ✓ The set of instances $<x_1, x_2, \ldots, x_m>$ is fixed.
  Therefore $D = \{<x_i, c(x_i)>\}$ can be simplified as $D = \{<c(x_i)>\}$
  ie, $<c(x_1), c(x_2), \ldots, c(x_m)>$ corresponding to $<x_1, x_2, \ldots, x_m>$

➢ **MAP-learner**

MAP-learner is a brute-force learning algorithm that outputs MAP hypotheses. Recall a MAP hypothesis is a *maximum a posteriori* hypothesis, i.e., a most probable hypothesis.
MAP-learner algorithm consists of two steps:

1. For each hypothesis h in H, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2. Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

➢ **ConsistFilter**

ConsistFilter(D, H) returns a hypothesis in the version space $VS_{H,D}$.
Recall the version space contains all the hypotheses that are consistent with the set of training examples D. Therefore, ConsistFilter filters out those hypotheses that are inconsistent with D from the hypothesis space H.

➢ **Find-S**

Find-S outputs a maximally specific hypothesis from the version space $VS_{H,D}$.

5.2 ConsistFilter vs. MAP-learner

Can we view ConsistFilter as MAP-Learner? Or equivalently, when does ConsistFilter produce MAP hypotheses? To answer this question, we start with MAP-learner.

In order to specify a learning problem for MAP-learner, we must specify what values are to be used for $P(h)$ and $P(D|h)$.

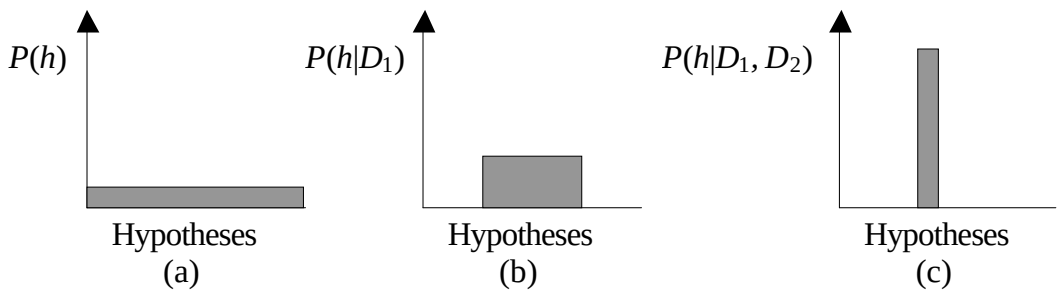
- ✓ $P(h) = 1/|H|$ for all h in H
- ✓ $P(D|h) = \begin{cases} 1 & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases}$

Note $P(h) = 1/|H|$ is a reasonable choice given no prior knowledge that one hypothesis is more likely than another.
Also, $P(D|h) = |VS_{H,D}| / |H|$ because we assume D is noise free and the number of hypotheses consistent with D is by definition $|VS_{H,D}|$.
Therefore, from Bayes theorem we will have the following:

$$P(h|D) = \begin{cases} 1/|VS_{H,D}| & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases}$$

The above analysis implies that under our choice of $P(h)$ and $P(D|h)$, every consistent hypothesis has posterior probability $1/|VS_{H,D}|$, and every inconsistent hypothesis has posterior probability 0. Every consistent hypothesis is therefore a MAP hypothesis. In other words, under such choices, ConsistFilter is a MAP-learner.

The following graph shows that as training data accumulates, the posterior probability for inconsistent hypothesis becomes 0 while the total probability summing to 1 is shared equally by the remaining consistent hypotheses.



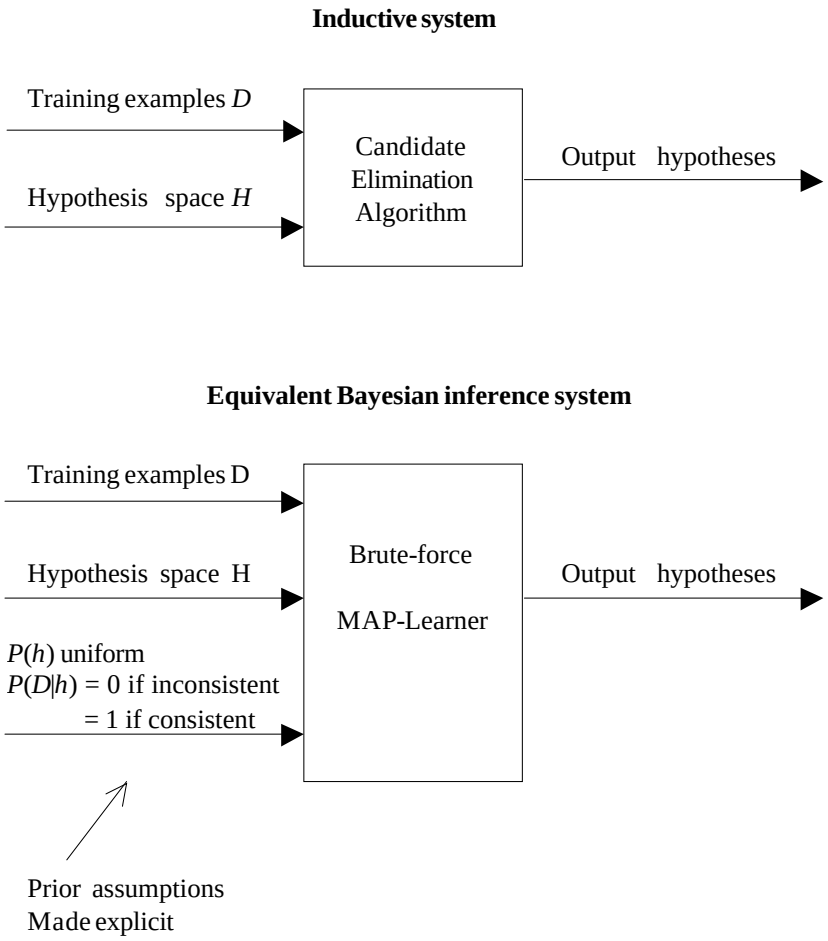
5.3 Find-S vs. MAP-learner

It is clear Find-S outputs a MAP hypothesis under the above choices of $P(h)$ and $P(D|h)$ because Find-S $\leq$ ConsistFilter in the sense that the output of Find-S is also in the version space $VS_{H,D}$. But are there other choices of $P(h)$ and $P(D|h)$ under which Find-S outputs MAP hypotheses?

The answer is yes. Because Find-S outputs maximally specific hypothesis from the version space, its output hypothesis will be a MAP hypothesis under any prior probability distribution that favors more specific hypothesis. To put it more formally:

If $P(h_1) \geq P(h_2)$ whenever h_1 is more specific than h_2, then Find-S is a MAP-learner.

5.4 Characterizing Learning Algorithms by Equivalent MAP-learners



From the graph we can see that a probabilistic reasoning system based on Bayes theorem (MAP-learner) will exhibit input-output behavior equivalent to Candidate-Elimination (or Find-S), provided it is given these assumed probability distributions $P(h)$ and $P(D|h)$.

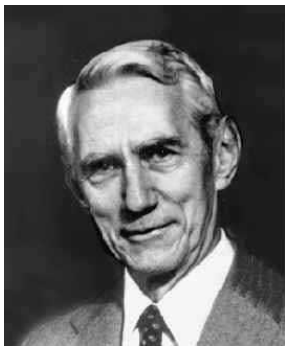
6. Minimum Description Length

6.1 Some History

Jorma Rissanen
IBM Research Division, Almaden Research Center

Introduced MDL in 1978, which triggered a large body of research in the communities of statistics, mathematics, machine learning, and philosophy, etc.

William of Occam (1285 – 1349)
Born in the village of Occam in Surrey (England), was the most influential philosopher of the century and a theologian.
The medieval rule of parsimony, or principle of economy, frequently used by Occam came to be known as:
Occam's razor: *plurality should not be assumed without necessity*
(or, in modern English, *keep it simple, stupid*)



Claude E Shannon (1916 – 2001)
MIT, Ph.D. in Mathematics, 1940
MIT, Master in electrical engineering

The father of "Information Theory"
Shannon's theorem (1948) concerns information entropy and gives the theoretical maximum rate at which error-free bits can be transmitted over a noisy channel.

Thomas Bayes (1702 – 1761)
Was a British mathematician and Presbyterian minister.
Known for having formulated Bayes theorem.

Bayes theorem was originally used to prove the existence of God: without assuming the existence of God, the operation of the universe is extremely unlikely; therefore, since the operation of the universe is a fact, it is very likely that God exists.

6.2 Some Background: Entropy Encoding

An **entropy encoding** is a coding scheme that assigns codes to symbols so as to match code lengths with the probabilities of the symbols. Typically, entropy encoders are used to compress data by replacing symbols represented by equal-length codes with symbols represented by codes proportional to the negative logarithm of the probability. Therefore, the most common symbols use the shortest codes.

According to Shannon's theorem, the optimal code length for a symbol is $-\log_b P$, where b is the number of symbols used to make output codes and P is the probability of the input symbol. In case of binary data, the optimal code length is $-\log_2 P$.

Three of the most common entropy encoding techniques are Huffman coding, Range encoder, and arithmetic encoding.

Therefore, the expected length for transmitting one message is:

$$\sum_i - P_i \log_2 P_i$$

6.3 The MDL Principle

Choose h_{MDL} such that

$$h_{MDL} = \text{argmin} \{ Length_{C1}(h) + Length_{C2}(D|h) \}$$

where $Length_C(x)$ is the description length of x under encoding C.

- ✓ Note that MDL is an operational form of Occam's razor, which states that *one should not increase, beyond what is necessary, the number of entities required to explain anything.*

- ✓ The following is an Example:
  H = decision trees D = training data labels
  $Length_{C1}(h)$ = the number of bits used to describe tree h.
  $Length_{C2}(D|h)$ = the number of bits used to describe $D|h$.
  Note $Length_{C2}(D|h)$ need only to describe exceptions. It is 0 if all examples are perfectly classified by h.

- ✓ h_{MDL} trades off tree size for training errors. It might select a shorter hypothesis that makes a few errors over a longer hypothesis that perfectly classifies the training data and therefore provides a way of dealing with overfitting.

- ✓ Implicitly, we can think of a code as defining a prior probability distribution on our hypothesis space. This gives a nice Bayesian explanation for preferring shorter trees over larger trees. We simply have a prior probability distribution that prefers shorter trees.

6.4 Understanding MDL

MDL is motivated by interpreting h_{MAP} in the light of basic concepts from information theory. Recall in Bayesian learning, we are interested in finding the most probable hypothesis $h \in H$ given the observed data D. Any such maximally probable hypothesis is called a *maximum a posteriori* (MAP) hypothesis.

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\ P(D|h)P(h)$$

Since *log* is a monotonic function, this is equivalent to

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\ log_2 P(D|h) + log_2 P(h)$$

Or, alternatively

$$h_{MAP} = \underset{h \in H}{\text{argmin}}\ - log_2 P(D|h) - log_2 P(h) \qquad (1)$$

☺ Aside: A joke on the origin of logs (there are other versions of the same joke)

There's an old joke well known among mathematicians about logarithms. After the flood waters receded, Noah commanded the animals to go forth and multiply. The snakes went up to Noah and told him they couldn't multiply because they were adders. So Noah built them a piece of wooden furniture with a flat top and four legs. The adders could now multiply because they had a log table.

- ✓ Equation (1) can be interpreted as a statement that short hypotheses are preferred. Recall that optimal code uses $-log_2 P$ bits for an event with probability P:
- ✓ $-log_2 P(h)$ is the description length of h under the optimal encoding
- ✓ $-log_2 P(D|h)$ is the description length of D given h under the optimal encoding

Therefore, MDL prefer hypothesis that minimizes

$$length(h) + length(misclassifications)$$

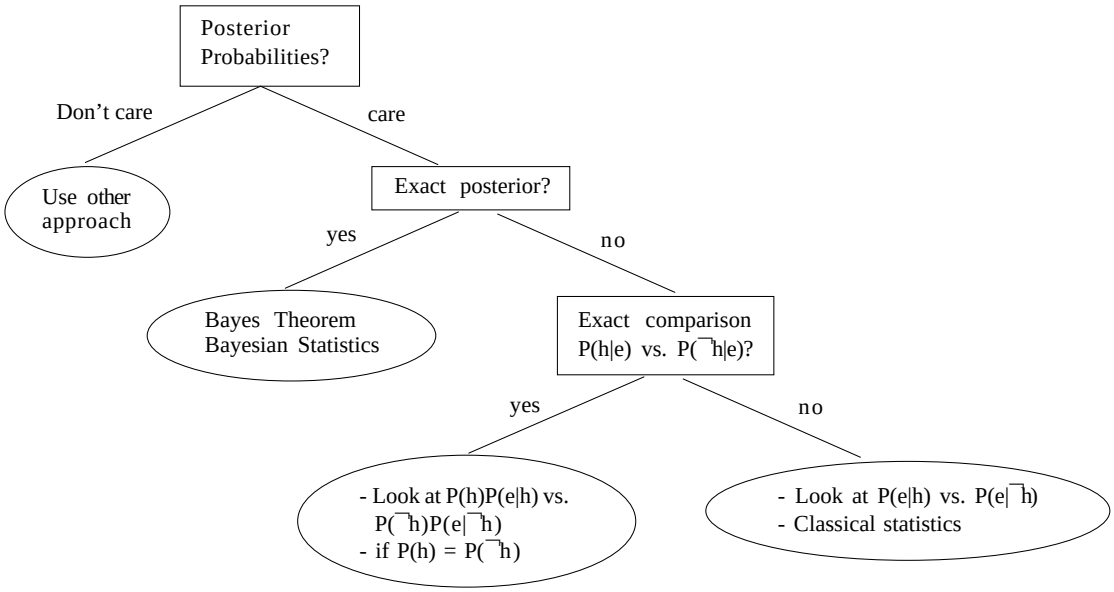
- ✓ If a representation of hypotheses is chosen so that the size of h is $-log_2(h)$, and if a representation for exceptions is chosen so that the encoding length of D given h is $-log_2(D|h)$, then the MDL principle produces MAP hypotheses.
- ✓ However, to show that we have such a representation, we must know all the priors $p(h)$ and $P(D|h)$.

6.5 Some Observations about MDL

If all hypotheses have the same description length, we then correspondingly have a uniform prior probability distribution which has maximal entropy. Then we will focus on how well a hypothesis classifies data, and this is the case typically studied by classical statisticians.

However, if some hypotheses have much shorter descriptions than others, the MDL becomes significant. MDL may prefer a hypothesis that makes more errors but much shorter than others.

7. Probabilistic Views of Learning



References:

T. Mitchel, Machine Learning, 1997

J.Rissanen, Modeling by shortest data description. *Automatica*, vol. 14 (1978), pp. 465-471

O. Schulte, CMPT-882 "Machine Learning" overheads, 2004

8 References

- T. Mitchel, Machine Learning, 1997 J.Rissanen
- Modeling by shortest data description. Automatica, vol. 14 (1978), pp. 465-471
- O. Schulte, CMPT-882 "Machine Learning" overheads, 2004
- `http://www.abelard.org/briefings/bayes.htm`
- `http://en.wikipedia.org/wiki/Main_Page`