Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2014/2015

Lesson 12 20 March 2015

Plausible Reasoning with Bayes Rule

For two proposition A and B, Bayes Rule tells us that

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

Bibliography : E. T. Jaynes, Probability Theory: The Logic Of Science, Cambridge University Press, 2003 (available on the course web site).

Plausible Reasoning

What is truth?

Logical truth: Logical consistency with a set of axioms and postulates.

Plausibility. A state of knowledge of the world. Plausible truth concerns the ability to predict and explain the world.

In the real world, new information is continually arriving. Plausible reasoning assimilates new information as it arrives.

"Plausibility" represents the degree to which a statement can be believed. This can be represented by likelihood or probability.

Plausible reasoning seeks to validate or invalidate hypotheses using uncertain or unreliable information.
Plausible reasoning can be used to reason about the truth of single hypothesis (H or ¬H) or choose from a number of competing hypotheses $\{H_i\}$.

Bayes Rule gives us a technique for using evidence to support hypotheses.
Let H be an hypothesis and let E be evidence for the hypotheses.

Then Bayes rule tells us

$$P(H, E) = P(H \mid E)\, P(E) = P(E \mid H)\, P(H)$$

so that $$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

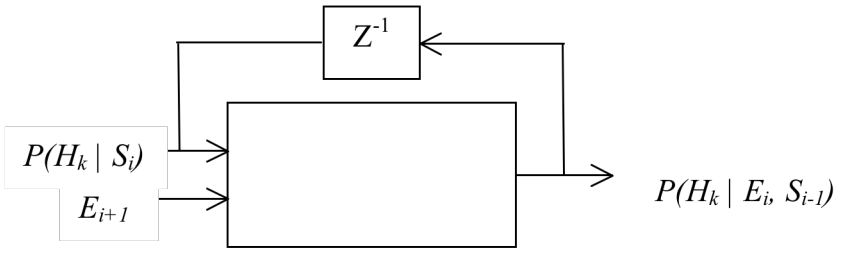
We would like to apply this technique recursively, as new data arrives and use this to reason over multiple hypotheses:

Assume that we have K hypotheses, H_k and we seek to accumulate evidence to select the most likely hypothesis. For multiple evidence we could write:

$$P(H_k \mid E_1,...,E_i) = \frac{P(E_1,...,E_i \mid H_k)P(H_k)}{P(E_1,...,E_i)}$$

The problem is how to incrementally estimate $P(E_1, ..., E_i)$ and $P(E_1, ..., E_i \mid H_k)$

To simplify the notation, let us define $S_i=\{E_1,..., E_i\}$ as a body of previous composed of i observations, and E_{i+1} as a new observation.



Z^{-1} is a memory buffer

For each new observation, E_{i+1} The problem is how to estimate $P(H_k \mid E_{i+1}, S_i)$ using the previous evidence S. This can be written:

$$P(H_k \mid E_{i+1}, S_i) = \frac{P(E_{i+1}, S_i \mid H_k) P(H_k)}{P(E_{i+1}, S_i)}$$

The index, i, and the accumulated evidence S_i are then updated:

$$S_{i+1} \leftarrow S_i \cup E_i \; ; \quad i \leftarrow i+1;$$

For a recursive formulation we need to factor this equation into the estimate at the previous time step, i, and an update formula, $u(E_{i+1}, S_i, H_k)$:

$$P(H_k \mid E_{i+1}, S_i) = u(E_{i+1}, S_i, H_k) \cdot P(H_k \mid S_i)$$

To determine $u(E_{i+1}, S_i, H_k)$ we will follow the development as laid out by E. T. Jaynes. Jaynes derives a "product rule" that says

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

Applying this rule to evidence reasoning we see that

$$P(E, H \mid S) = P(E \mid H, S) P(H \mid S) = P(H \mid E, S) P(E \mid S)$$

From this we can derive update formula as:

$$P(H_k \mid E_{i+1}, S_i) = \frac{P(E_{i+1} \mid H_k, S_i)}{P(E_{i+1} \mid S_i)} P(H_k \mid S_i)$$

The Plausible Reasoning System of E. T. Jaynes

E. T. Jaynes derives a consistent theory for Plausible Reasoning from Evidence by starting from a set of desired criteria. He starts by defining 3 terms:

H = Some hypothesis to be tested
S = Prior (background) Information (Jaynes calls this B for Background)
E = Evidence, A newly provided observation. (Jaynes calls this D for Data)

Prior information is NOT the same as "a-priori" information. Prior information informs the plausibility of H before (prior to) including the new data. S can be used to represent "experience" with similar problems. For example, when sampling balls from an urn, S can represent the results of all previous samples.

Jaynes argues that until E is assimilated into the reasoning it is only raw data. The degree that E acts as evidence for H depends on the plausibility of E and the plausibility of H.

With this approach we may reason about the truth of single hypothesis (H or ¬H) or choose from a number of competing hypotheses $\{H_k\}$.

Jayne's desired criteria for a system of Plausible Reasoning:

Jaynes derives his system to meet a set of desired criteria

1) Degrees of plausibility are represented by real numbers, where a greater number represents more plausible.
2) The system should have qualitative correspondence with (human) common sense.
3) The solution should be consistent: If a conclusion can be reached by more than one way (sequence of inferences) then all ways must lead to the same result.
4) All available evidence must be included. Evidence can not be arbitrarily ignored. (reasoning must be non-ideological - no cherry picking as is popular with certain politicians and salesemen).
5) Equivalent states of knowledge should be represented by equivalent plausibility.

1, 2, and 3 are "structural" criteria for the system. 4, and 5 are "interface" conditions.

For example, consider rule 4 - Politicians, salesman and propagandists tend to cite favorable evidence and ignore contradictory evidence. Jaynes shows that this leads to errors.

Rule 5 leads to a formal requirement that in the absence of any prior knowledge or evidence, all hypotheses are equally plausible, because they are interchangeable. For N equally plausible hypotheses: $\forall k : P(H_k) = \frac{1}{K}$

From these desired properties, Jaynes derives two rules for plausible reasoning: The Product Rule and the Sum Rule.

Notation.

Let A, B and C be statements.
Let A, B = A∧B represent the conjunction of A and B. Both A and B are true.
Let A+B = A∨B represent the disjunction of A and B. A or B (or both) are true.
Let ¬A represent Negation. A is NOT true.
P(A) is the plausibility of A. P(A) > P(B) means that A is more plausible than B.
Conditional relation: A | B : A given B.
So P(A|B) is the plausibility of A given B.

For example: P(A | B, C) : the plausibility of A given B and C.
P(A+B | C, D) : the plausibility of A or B given C and D.

P(A | C) > P(B | C) says that given that C is true, A is more plausible than B.

We also note that P(A | B) + P(¬A | B) = 1.

Jaynes uses his desired properties to derive two fundamental rules:

The Sum Rule

$$P(A+B \mid C) = P(A \mid C) + P(B \mid C) - P(A,B \mid C)$$

The Product Rule

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

For new evidence E, and previous evidence S:
The product rule tells us:

$$P(E,H \mid S) = P(E \mid H,S)P(H \mid S) = P(H \mid E,S)P(E \mid S)$$

From this we can derive

$$P(H \mid E,S) = \frac{P(E \mid H,S)}{P(E \mid S)} P(H \mid S)$$

In statistics, the term $P(H \mid E,S)$ is called the posterior probability. Posterior does NOT mean "later" but means "after assimilation of the E".

the term $L(E) = \frac{P(E \mid H,S)}{P(E \mid S)}$ is the "likelihood" of E given H and S.

Two cases can be considered (1) reasoning about a single hypothesis, and (2) reasoning over multiple hypotheses.

Single Hypothesis - the binary case.

The binary case is a realistic and valuable model for many practical problems.
In this case we seek to choose between H or ¬H given E and S.

The product rule tells us that $P(H \mid E,S) = P(H \mid S)\frac{P(E \mid H,S)}{P(E \mid S)}$

but also $P(\neg H \mid E,S) = P(\neg H \mid S)\frac{P(E \mid \neg H,S)}{P(E \mid S)}$

If we take the ratio we obtain:

$$\frac{P(H \mid E,S)}{P(\neg H \mid E,S)} = \frac{P(H \mid S)}{P(\neg H \mid S)} \frac{P(E \mid H,S)}{P(E \mid \neg H,S)}$$

This is known as the "odds" of the hypothesis (côte en Français) :

$$O(H \mid E,S) = \frac{P(H \mid E,S)}{P(\neg H \mid E,S)}$$

Odds are widely used in gambling to express the estimated frequency of occurrence of an outcome compared to an alternative.

From the product rule we can derive:

$$O(H \mid E,S) = O(H \mid S)\frac{P(E \mid H,S)}{P(E \mid \neg H,S)}$$

That is, the posteriori odds are equal to the prior odds multiplied by the likelihood ratio.

It is convenient to convert the likelihood ratio to a logarithmic scale so that we can add likelihood as evidence is obtained.
Jaynes defines this as a function e() that he refers to as evidence for the hypothesis

$$e(H \mid E,S) \equiv 10 Log_{10} O(H \mid E,S)$$

so that $$e(H \mid E,S) = e(H \mid S) + 10 Log_{10}\left[\frac{P(E \mid H,S)}{P(E \mid \neg H,S)}\right]$$

The use of base 10 allows us express the effect of new evidence in "decibels", a commonly used engineering convention for exponential scales.

When the our "data" is actually several different observations:

$E = E_1, E_2, E_3 \ldots$

Then the evidence accumulates as a sum of log likelihoods:

$$e(H \mid DB) = e(H \mid B) + 10 Log_{10}\left[\frac{P(E_1 \mid H,S)}{P(E_1 \mid \neg H,S)}\right] + 10 Log_{10}\left[\frac{P(E_2 \mid H,S)}{P(E_2 \mid \neg H,S)}\right] + \ldots$$

and for independent observations i=1, 2, …

$$e(H \mid E,S) = e(H \mid S) + 10\sum_i Log_{10}\left[\frac{P(E_i \mid H,S)}{P(E_i \mid \neg H,S)}\right]$$

In most cases, the probability of getting E_2 is not influenced by knowledge of E_1.

$$P(E_2 \mid E_1,H,S) = P(E_2 \mid H,S)$$

We say that E_1 and E_2 are "independent". Note that this says nothing about their causal relation, but only that they are logically independent with respect to our problem. For two logically related data, if we know one, then we know the other, whether or not they are causal.

The utility of log likelihood expressed in base 10 can be seen from the following table comparing log likelihood, odds and probabilities.

e	O	p
0	1:1	1/2
3	2:1	2/3
6	4:1	4/5
10	10:1	10/11
20	100:1	100/101
30	1000:1	0.999
40	10^4:1	0.9999
-e	1/O	1-p

The scale of e() is measured in dB (decibels) and can easily accommodate exponential numbers.

Jaynes observes the 1db was defined as the smallest discernable change in sound intensity. Interestingly, 1dB seems to be the smallest discernable change in likelihood given new evidence.

The binary (single hypothesis) case is simple and provides an elegant solution. Unfortunately, it is NOT extensible to multiple hypotheses.

Multiple Hypotheses.

In the case of multiple hypotheses, we have K mutually independent hypotheses, H_k. We assume that one of these is correct, and all others are false. Thus

$$\sum_k P(H_k \mid S) = 1$$

Given a new observation, E,

$$P(H'_k \mid E,S) \leftarrow \frac{P(E \mid H_k,S)}{P(E \mid S)} P(H_k \mid S)$$

We note that for this to work. $P(E \mid S) = \sum_k P(E \mid H'_k,S)$

Depending on how we compute $P(E \mid H'_k,S)$ and $P(E \mid S)$ this may not be true.

Alternatively, we can note that because the term *P(E|S)* is common to all hypotheses, it can be ignored. We can then define the relative Likelihood for each hypothesis as

$$L(H'_k \mid E,S) = P(E \mid H_k,S)P(H_k \mid S)$$

The difference is that likelihoods do not sum to 1.
The relative likelihoods can be normalized to provide probabilities by dividing by the sum of likelihoods.

$$P(H'_k \mid E,S) = \frac{L(H'_k \mid E,S)}{\sum_k L(H'_k \mid E,S)}$$

This is technique is used in found in many software tools for Bayesian reasoning.

Example: Classifying text from word frequency.

For our example, we can use Jayne's plausible reasoning to recursively recognize classes of text from the frequency of words.

Assume that we have k classes of text, C_k and that we have samples of N_k words for each class, $\{w_n^k\}$. We can use word frequency to estimate the probability of a word given a class of text. Allocate K hash tables, $h_k()$ and for each class of words count the word frequency in the samples $\{w_n^k\}$.

$\forall w \in \{w_n^k\}: h_k(w) \leftarrow h_k(w) + 1$

Then $P(w \mid C_k) = \frac{1}{N_k} h_k(w)$ and $P(w) = \frac{1}{N} h(w)$ where $h(w) = \sum_k h_k(w)$ and $N = \sum_k N_k$

we can use the number of words in the training corpus to estimate $P(H_k) = \frac{N_k}{N}$

A "probe" is a sample of text to be classified, $\{w\}$

We can use word counts to recursively estimate the class for the probe.

At i=0 we observe the first word of the probe, $E_1 = w_1$.
Initially S_0={} is empty, so that

$$P(H_k \mid E_1) = \frac{h_k(w_1)}{h(w_1)}$$

We then update the accumulated evidence, S_i, and index i. $S_1 \leftarrow w_1$; $i \leftarrow 1$;

For word 2, and for the following words:

$$L(H_k \mid w_{i+1}, S_i) = P(w_{i+1} \mid S_i, H_k) P(H_k \mid S_i)$$

But how do we compute : $P(w_{i+1} \mid S_i, H_k)$?

We can incrementally compute a new hash $h_s(w)$ for each new word in the probe

$h_s(w_{i+1}) \leftarrow h_s(w_{i+1i}) + 1$ and $i \leftarrow i+1$. Note that at each cycle $N_s = i$

For $P(w_{i+1} \mid S_i, H_k)$ we combine the hash from the probe with the hash for each class.

$$P(w_{i+1} \mid S_i, H_k) = \frac{1}{(N_k + i)}(h_k(w_{i+1}) + h_p(w_{i+1}))$$

we then obtain:

$$L(H_k \mid w_{i+1}, S_i) = P(w_{i+1} \mid S_i, H_k)P(H_k \mid S_i)$$

To return the set of likelihoods to a probability we normalize by dividing by the sum of likelihoods.

$$P(H'_k \mid w_{i+1}, S_i) = \frac{L(H'_k \mid w_{i+1}, S_i)}{\sum_k L(H'_k \mid w_{i+1}, S_i)}$$

There remains a problem. What happens if we encounter a rare word. If the word has never been encountered in any corpus, then the recursive update is:

$$P(w_{i+1} \mid S_i, H_k) = 0$$

This can be detected and the word can be rejected. However, if the word was in one of the training corpuses, but not the others, all other word classes will have their probability set to 0. Forever. This can cause an error. One solution is to replace 0 by the probability that the word is in the class C_k, but that it was not in the training corpus, $\{W_n^k\}$. We can estimate this as

for example: we can estimate $P((w \in C_k) \wedge (w \notin \{w_n^k\})) = \frac{1}{N_k}$

This is equivalent a minimum value of P_{min} for $P(w \mid H_k)$.

if $P(w \mid H_k) < P_{min}$ then $P(w \mid H_k) = P_{min}$.

An alternative is to assimilate the new word into the probe before updating the probability.

$h_s(w_{i+1}) \leftarrow h_s(w_{i+1}) + 1$ and $N_s = i+1$

This serves to assure that the first time the word is encountered, the update factor is $\frac{1}{N_k + i}$.