Tutorial on Bayesian Methods and the MaxEnt Principle

Wray Buntine
Heuristicrats Research, Inc.
wray@Heuristicrat.COM
http://www.Heuristicrat.COM/wray/

1678 Shattuck Avenue, Suite 310
Berkeley, CA, 94709-1631
Tel: +1 (510) 845-5810
Fax: +1 (510) 845-4405

Peter Cheeseman
Caelum Research Corp.
cheesem@ptolemy.arc.nasa.gov

NASA Ames Research Center
MS 269-2
Moffett Field, CA, 94035-1000
Tel: +1 (415) 604-4946
Fax: +1 (415) 604-3594

Sante Fe, New Mexico, June 31st, 1995.

Outline

- **Basic probability theory ....(Peter)**
- **Simple examples of Bayesian Inference...........(Peter)**
- **Types of probabilistic inference ....(Peter)**
- **Case Studies...........(Peter)**
- **Advanced Modeling...........(Wray)**
- **Graphical (probabilistic) models...........(Wray)**
- **Computation...........(Wray)**
- **Priors...........(Wray)**
- **Other views and ideas...........(Peter and Wray)**

Bayesian Inference I

- **Q1: How should a rational agent form beliefs under uncertainty?**
- **Q2: How should a rational agent make decisions under uncertainty?**
- **Initially concentrate on beliefs of a rational agent.**
- **Must Generalize logic:**
  - T or F (0 or 1) --> degree of belief (numerical).
  - degree of belief depends on particular (known) context
- **Cox's Proof shows that probability theory is the only consistant theory that generalizes logic in this way (more later!).**
- **Example probability statement:**
  - P(Clinton will win in 1996 | Bosnia-resolved-by-1996, 1995) = .4
  - .4 is degree of belief
  - "Clinton will win in 1996" is target proposition (form beliefs about it)
  - "1995" is a proposition describing the current conditioning context.

Bayesian Inference II

  - Bosnia-resolved -by-1996 is a conditioning proposition.
  - The | symbol separates the target proposition from the conditioning proposition(s).
- **Target Proposition:**
  - Can be atomic or Boolean combination of propositions.
  - Propositions can quantified–e.g. "All people in this room are older than 25 years".
- **Conditioning Proposition:**
  - Can be atomic or Boolean combination of propositions.
  - Always includes a proposition representing the context of the probability assertion (sometimes omitted).
  - Can include quantified proposition–e.g. "All people in this room employed".

Basic Probability Laws I

- **Probability Law of Excluded Middle (Negation Law):**
  P(A) = 1 - P(not A)
- **Positivity Law:**
  $0 \leq P(A) \leq 1$
- **Non-Truth Functionality:**
  - e.g. $0 \leq P(A \& B) \leq \min(P(A),P(B))$ [P(A & B) = P(A,B)]
  - The probability of the conjunction is not determined by its components (but is bounded by them).
- **Disjunction:**
  - P(A or B) = P(A) + P(B) - P(A & B)
  - If A and B mutually exclusive, then
  - P(A or B) = P(A) + P(B) (Additivie Law of probabilities)

Basic Probability Laws II

- **Multiplication Law:**
  P(A,B,C,..|I) = P(A|I)P(B|A,I)P(C|A,B,I)...
  = P(B|I)P(A|B,I)P(C|A,B,I)...
  = P(C|I)P(B|C,I)P(A|B,C,I)... etc.
- **Bayes Theorem**
  - From Multiplication Law
  P(A|I)P(B|A,I) = P(B|I)P(A|B,I)
  --> P(A|I) = P(B|I)P(A|B,I)/P(B|A,I) [Bayes Theorem]
- **Marginalization (Discrete)**
  P(A|C) = P(A,B|C) + P(A,not B|C) [B is <u>binary</u> auxilary variable]
  $P(A|C) = \Sigma_i P(A,X_i|C)$ [X_i is an i-way auxilary variable]
  $= \Sigma_i P(A|X_i,C)*P(X_i|C)$
- **Marginalization (Continuous)**
  $P(A|C) = \int P(A,x|C)\, dx$
  $= \int P(A|x,C)*f(x|C)\, dx$

Examples of Marginalization

- **Discrete**
  P(Pass-PhD|School) = P(Pass-PhD, Female|School) +
  P(Pass-PhD, Male|School)
  = P(Pass-PhD|Femal,|School)P(Female|School) +
  P(Pass-PhD|Male,School)P(Male|School)
  $P(\text{Pass-PhD}|\text{Female,USA}) = \Sigma_{\text{schools}} P(\text{Pass-PhD,School}|\text{Female,USA})$
- **Continuous**
  $P(\text{Pass-Phd}|\text{USA}) = \int P(\text{Pass-PhD,Age}|\text{USA})\, d(\text{Age})$
  $= \int P(\text{Pass-PhD}|\text{Age,USA})*f(\text{Age}|\text{USA})\, d(\text{Age})$
- **Marginalization Eliminates "Nuisance" Variables:**
  - The effect of Marginalization is to eliminate explicit dependence on the variable(s) that are marginalized away.

Probability Density Functions

- **Probabilities are numbers from 0 to 1, representing degree of belief in target proposition given conditioning informtion.**

E.g.--Q: What is probability that this rock weighs <u>exactly</u> 1 Kg.?

Ans: Zero (infinitessimal)

--> Need probability density functions!

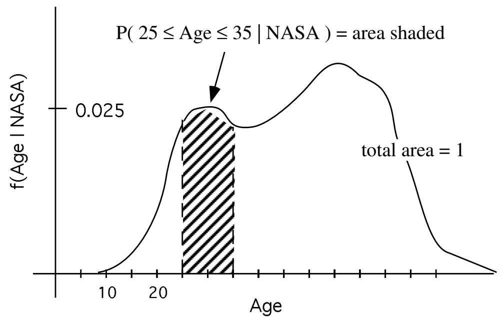
- **Definition: Probability Density Function (pdf).**
  $f(x|C)$ is a piece-wise continuous function of x s.t.
  - $f(x|C) \geq 0$
  - $\int f(x|C)\, dx = 1$ (i.e. x must have <u>some</u> value!)
- **Probabilities found by integrating pdfs over specific ranges.**
  - Example:

  $P(1\text{Kg.} \leq \text{weight(rock)} < 1.1\text{ Kg.}) = \int f(\text{weight(rock)})\, dw$

  i.e. the probability that the rock weighs between 1 and 1.1 Kg. is given by the integral of the pdf over the range. (see next slide)

PDF Example

Area under curve is required probability:



Note:

- f(x|C) can be > 1 [f(x|C) is <u>not</u> a probability.]
- f(x\C) can be regarded as the limiting result of a probabilistic histogram as the bin sizes go to zero.

Probability Notes 1

- **All Probabilities are conditional probabilities:**
  - always condition on context
  - Sometimes conditioning information understood (not explicit)--Danger!!
- **There is no such thing as <u>THE</u> probability of a proposition:**
  - As learn new conditioning information and choose to use it, the resulting conditional probability will be different than previous conditional probabilities--i.e the best estimate probability changes with new information.
  - Probability statements can refer to the <u>next</u> outcome in a series or to future values based on current evidence, but <u>not</u> to long term frequency.
- **Conditional Probability ≠ Probability of a Conditional !!**

  e.g. "Where ever there is smoke there is likely to be fire".
  - Is P(Fire | Smoke, context) = high (.9)
  - <u>Not</u> P(Smoke -> Fire | context) = high (.9); [No smoke events count as evidence!]

Probability Notes II

- **Probability is not a <u>Frequency</u> (it is a measure of belief).**
  - Can have a probability of a single event e.g. Prob. of Clinton being re-elected in 1996.
  - probability equals expected frequency in repeated trials (probability and frequency are closely related).
- **Conditioning Information can be Hypothetical.**

  e.g. "If I miss my fight, I can probably get another one today".
  - conditioning information does <u>not</u> have to be true.
  - can consider many mutually inconsistant conditioning contexts.
  - probabilistic inference is montonic--i.e. do not have to change previous beliefs if the context changes (compute new probabilities in the new context instead).
- **Odds map probabilities from [0,1] to [0,∞]--i.e.**

  Odds(A) = P(A)/P(not A)

  = P(A)/(1 - P(A)) [Only good for Binary propositions]

  To transform from Odds to probability use: P = Odds/(1 + Odds)

Alternative Forms of Bayes Theorem

- **Basic Form of Bayes theorem for a set of mutually exclusive and exhaustive hypotheses H(i), given evidence E:**

$$P(H_i|E,C) = \frac{P(H_i|C)*P(E|H_i,C)}{P(E|C)}$$

<u>posterior prob.</u> = <u>prior prob</u>. x <u>likelihood</u> / normalizing const.

Where $P(E|C) = P(E|H_i,C)*P(H_i|C)$ —i.e. marginalize over all H_i.

Note that $P(E|C)$ does not depend on H_i — it is just a normalizing constant

- **Relative version of Bayes:**
-

$$\frac{P(H_i|E,C)}{P(H_j|E,C)} = \frac{P(H_i|C)*P(E|H_i,C)}{P(H_j|C)*P(E|H_j,C)}$$

  - Eliminates the normalizing constant, but requirement that $\sum_i P(H_i|E,C) = 1$ allows the $P(H_i|E,C)$'s to be normalized.

Example of Bayesian Inference

Situation: **There are 64 coins in a box, one of these coins is double-headed (H2), the rest are ordinary (H1) . A single coin is drawn from the box.**

- **Q1: What is the probability that this coin is the double-headed coin?**

Ans: P(H2|C) = 1/64 [C is the context]

– Principle of Indifference (or more generally, Maximum Enrtopy).

New Situation: **The selected coin is flipped, and the result (R1) is "heads". [A "tails" result means that not double-headed coin]**

- **Q2: What is the new probability that this coin is the double-headed coin?**

Ans:--Use Bayes!!

-- Relative version of Bayes is easiest to use.

Double-Headed Coin Example (Cont.)

- **Relative Bayes for H1 and H2:**

$$\frac{P(H2|R1,C)}{P(H1|R1,C)} = \frac{P(H2|C)*P(R1|H2,C)}{P(H1|C)*P(R1|H1,C)}$$

P(H2|C) = 1/64 (prev. slide); P(H1|C) = 63/64 (By normalization)

P(R1|H2,C) = 1 (only possible outcome); P(R1|H1,C) = 1/2 (fair coin).

Therefore: P(H2|R1,C)/P(H1|R1,C) = (1/63)*2 = 2/63 (increased prob.)

And: P(H2|R1,C) = 2/65

<u>New Situation:</u> **The selected coin is flipped again, and the result (R2) is also "heads".**

[Note: If any flip gives "tails" then P(H2|E,C) = 0]

Want: P(H2|R1,R2,C) --> Bayes again!

Double-Headed Coin Example (Cont.)

- **Relative Bayes again:**

$$\frac{P(H2|R1,R2,C)}{P(H1|R1,R2,C)} = \frac{P(H2|C)*P(R1,R2|H2,C)}{P(H1|C)*P(R1,R2|H1,C)}$$

- P(R1,R2|H2,C) = 1 (only possibility), but what is P(R1,R2|H1,C)?

Note: In principle, P(R1,R2|H1,C) could be any value from 0 to 1/2.

Solution: Use <u>principle of maximum entropy</u> to find <u>the</u> probability that maximizes the entropy subject to any constraints (more later)!

Result: <u>Conditional Independence</u>--i.e.

P(R1,R2 |H1,C) = P(R1 |H1,C)*P(R2 |H1,C) or

P(R1 |R2,H1,C) = P(R1 |H1,C)

Double-Headed Coin Example (Cont.)

- **Two Flip (R1,R2) Conclusion:**

$$\frac{P(H2|R1,R2,C)}{P(H1|R1,R2,C)} = \frac{P(H2|C)*P(R1,R2|H2,C)}{P(H1|C)*P(R1,R2|H1,C)} = \frac{(1/64)*1}{(63/64)*(1/2)*(1/2)} = \frac{4}{63}$$

Which gives: P(H2|R1,R2,C) = 4/69

- **<u>Recursive form of Bayes</u> (when evidence is conditionally independent).**

$$\frac{P(H2|R1,R2,C)}{P(H1|R1,R2,C)} = \frac{P(H2|C)*P(R1,R2|H2,C)}{P(H1|C)*P(R1,R2|H1,C)} = \frac{P(H2|C)*P(R1|H2,C)*P(R2|H2,C)}{P(H1|C)*P(R1|H1,C)*P(R2|H2,C)}$$

$$= \frac{P(H2|R1,C)*P(R2|H2,C)}{P(H1|R1,C)*P(R2|H2,C)} = \frac{\text{Prior * Likelihood}}{\text{Prior * Likelihood}}$$

i.e. Previous posterior probability becomes the prior on the next iteration!

HIV Testing Example

Situation 1: A patient enters a clinic.

Q1: What is the probability that this patient is HIV+ ?

Ans: P(HIV+|Clinic) = .01 (answer depends on clinic, location etc.)

Note: P(HIV+|Clinic) $\neq$ P(HIV+|USA) ("The" prior probability)

Situation 2: A blood sample from the patient is tested using the ELISA test, and is found +ve (E1+).

Q2: What is the prob. that the patient is HIV+ given E1+ ?

Ans: Relative Bayes: Posterior ratio = Prior-ratio*Likelihood-ratio

$$\frac{P(HIV+|E1+,C)}{P(HIV-|E1+,C)} = \frac{P(HIV+|C)*P(E1+|HIV+,C)}{P(HIV-|C)*P(E1+|HIV-,C)} = \frac{.01 \times .98}{.99 \times .05} = .198$$

--> P(HIV+|E1+,C) = .165 (much less than 1!)

HIV Testing Example (Cont.)

Situation 3: The blood sample from the patient is tested using the ELISA test, and is found -ve (E1-).

Q3: What is the prob. that the patient is HIV+ given E1- ?

Ans: Relative Bayes: Posterior ratio = Prior-ratio*Likelihood-ratio

$$\frac{P(HIV+|E1-,C)}{P(HIV-|E1-,C)} = \frac{P(HIV+|C)*P(E1-|HIV+,C)}{P(HIV-|C)*P(E1-|HIV-,C)} = \frac{.01 \times .02}{.99 \times .95} = .00021$$

--> P(HIV+|E1+,C) = .00021 (from a prior of .01 !)

Situation 4: The blood sample from the patient is tested again using the ELISA test, and is found +ve (E2+) after the first test was +ve (E1+).

Q4: What is the prob. that the patient is HIV+ given E1+ and E2+ ?

Ans: Relative Bayes: Posterior ratio = Prior-ratio*Likelihood-ratio

HIV Testing Example (Cont.)

$$\frac{P(HIV+|E1+,E2+,C)}{P(HIV-|E1+,E2+,C)} = \frac{P(HIV+|C)*P(E1+,E2+|HIV+,C)}{P(HIV-|C)*P(E1+,E2+|HIV-,C)} = \frac{.01 \times ???}{.99 \times ???}$$

Q5: What value should be used for P(E1+,E2+|HIV+,C) and P(E1+,E2+|HIV-,C)?

Possible Answers:

Total Dependence: P(E1+,E2+ |HIV+,C) = P(E1+ |HIV+,C)
(No new Info.) P(E1+,E2+ |HIV-,C) = P(E1+ |HIV-,C)

Conditional Independence:
P(E1+,E2+ |HIV+,C) = P(E1+ |HIV+,C) * P(E2+ |HIV+,C)
P(E1+,E2+ |HIV+,C) = P(E1+ |HIV-,C) * P(E2+ |HIV-,C)

Empirically Determined Values: E.g.
P(E1+,E2+ |HIV+,C) = #(E1+,E2+ |HIV+,C)/ #(all test results|HIV+,C)

HIV Testing Example (Cont.)

Situation 5: The blood sample from the patient is tested again using the Western Blot test, and is found -ve (WB-), after an ELISA test was found +ve (E1+).

Q6: What is the prob. that the patient is HIV+ given E1+ and WB- ?

Ans: Relative Bayes: Posterior ratio = Prior-ratio*Likelihood-ratio

$$\frac{P(HIV+|E1+,WB-,C)}{P(HIV-|E1+,WB-,C)} = \frac{P(HIV+|C)*P(E1+,WB-|HIV+,C)}{P(HIV-|C)*P(E1+,WB-|HIV-,C)} = \frac{.01 \times ???}{.99 \times ???}$$

Q7: What value should be used for P(E1+,WB-|HIV+,C) and P(E1+,WB-|HIV-,C)?

Possible Answer: Assume conditional independence–i.e. result of tests depends only sample--not on the results of other tests.

HIV Testing Example (Cont.)

$$\frac{P(HIV+|E1+,WB-,C)}{P(HIV-|E1+,WB-,C)} = \frac{P(HIV+|C)*P(E1+,WB-|HIV+,C)}{P(HIV-|C)*P(E1+,WB-|HIV-,C)} = \frac{.01 \times .0001}{.99 \times .05} = .000002$$

-> P(HIV+|E1+,WB-,C) = .000002 i.e. The WB- evidence overwhelms the E1+ evidence.

<u>Summary--HIV Example:</u>

- Probabilistic inference is an update procedure---prior beliefs--> posterior
- Even though there may be a large change in <u>relative</u> probability in a Bayesian update, the absolute magnitude may still be small.
- How new evidence interacts with previous evidence depends on the domain. Whether conditional independence (maxent) applies is domain dependent.
- Priors are dependent on the <u>specific</u> context of the inference.
- Evidence is never “contraditory” (e.g. E1+ and WB-), but different pieces of evidence can swing the probability toward 0 or 1.

Types of Probabilistic Inference

- **Direct (Likelihood):**
  - Likelihood determination
  - Maximum Likelihood estimation.
- **Inductive:**
  - Posterior Probability Inference (inverse inference)
  - Maximum Posterior probability estimation
  - Abductive Reasoning
- **Projective (marginalization):**
  - eliminate nuisance variables
  - Important special case–convolution
- **Transductive:**
  - i.e Find probability of new evidence given old.
- **Probability Transformation (Re-parameterization):**

Types of Probabilistic Inference, –Direct–

Example (Likelihood):

P(Observed Intensity|Intrinsic luminousity, distance) = N(mean,var)

- Likelihood **is** the domain model (states how observables depend on the true state of the world, assumed known).
- Likelihood is usually a <u>function</u> of (conditioned on) the state of the world.

Maximum Likelihood Inference:

- Example: P(heart-attack| age) = f(age). Given that someone has had a heart-attack, what is their most likely age?
- Vary the conditioning variable(s) to find the value(s) that maximize the probability (or pdf). This value(s) is the maximum likelihood (ML) <u>estimator(s)</u>.
- Can estimate the uncertainty of the ML estimator by looking at the change in probability around the maximum as the variable(s) are varied.

Types of Probabilistic Inference, –Inductive–I

Induction ≡ P(Model | Data)

$\propto$ P(Model) * P(Data | Model) [Bayes]

Previous Examples:

- Double-Headed Coin example (Binary target variable, discrete evidence)
- HIV Testing example.

General Inductive Inference = Inverse Inference

- i.e. If know true state of the world, then can predict the data (probabilistically), <u>but</u> given the data want the true state of the world.
- e.g. X-ray crystallography, IRS audit prediction, diagnosis,....

Bayes is general Solution to Inverse Problems

- Bayes finds the posterior probability distribution over possible models given data and a prior distribution over models.

Types of Probabilistic Inference, –Inductive–II

Maximum Aposteriori Probability (MAP) Estimation:

- Picks the model(s) with maximum posterior probability
- Most posterior probability distributions have many local maxima.
- Need to search to find maximum (or local maximum)
- Need to indicate how concentrated the probability distribtion is around the maximum ("error bars").

Why find MAP estimates?

- Posterior probability distribution contains all the information from prior beliefs and data–the MAP estimate is a summary that loses information.
- The most likely posterior model is not generally the same as the mean model, and can vary depending on how the problem is parameterized.
- Hill climbing is a simple procedure for finding (local) MAP estimates.

Conclusion:

Where convenient use full posterior distribution!

Types of Probabilistic Inference, –Projective–

Project out the variable(s) of interest = marginalize over all "nuisance' variables.

Example:

$$f(\mu,\sigma| X) \longrightarrow f(\mu|X) = \int f(\mu,\sigma|X)\, d\sigma$$
$$= \int f(\mu|\sigma,x)*P(\sigma|x)\, dx$$

For a Normal: $f(\mu|X) = \dfrac{\Gamma(I/2) * S^{(I-1)}}{\sqrt{\pi} * \Gamma(I/2 - 1/2) * \{S^2 + (m - \mu)^2\}}$

——Student "T" distribution.

Where S = sample standard deviation, m = sample mean, and $\Gamma()$ is the Gamma function.

Types of Probabilistic Inference, –Transduction–

Transductive inference gives the probability of new data given old data (by marginalizing over model possibilities).

Example (Previous HIV Example):

P(WB+|E1+)

= P(WB+,HIV+|E1+) + P(WB+,HIV-|E1+)

= P(WB+|HIV+)*P(HIV+|E1+) + P(WB+|HIV-)*P(HIV-|E1+)

Where we have assumed conditional independence of evidence e.g. P(WB+|HIV+) = P(WB+|HIV+,E1+)

Can use transduction to evaluate the effect of evidence that could be obtained.

Types of Probabilistic Inference, –Probability Transformation–

Probability transformation allows a PDF in one representation to be transformed to another.

Example: Transform from Polar to Cartesian representation, i.e.

$$f(r,\theta) \rightarrow h(x,y)$$

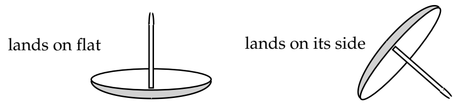
Answer:

$$f(r,\theta) = h(x,y) * \mathrm{Det}\left[\frac{d(x,y)}{d(r,\theta)}\right];$$

I.E. Multiply by the Jacobian to transform correctly.

Thumb-Tack Example

We toss a thumbtack N times with probability θ of it landing on its flat



Direct Inference: If know θ, what is the probability that will get n "flats" in N trials?

Ans: From logic get <u>Binomial Distribution</u>:

$$P(n|\theta,N) = \frac{n!\,(N-n)!}{N!} * \theta^{n} (1 - \theta)^{(N-n)}$$

Thumb-Tack Example II

Inductive Inference: Given number of "sides" n, and total number of trials N, what is θ ?

Ans: Use Bayes to invert the binomial distribution:

$f(\theta|x) \propto \pi(\theta) * l(x|\theta)$. Use (conjugate) prior dist $\pi(\theta) \propto \theta^{\alpha} (1-\theta)^{\alpha}$

$$\text{Then } f(\theta|x) = \beta(\theta|x) = \frac{\Gamma(N + 2\alpha)}{\Gamma(n + \alpha) * \Gamma(N-n+\alpha)} * \theta^{(n+\alpha-1)} (1-\theta)^{(N-n+\alpha-1)}$$

Note: The beta distribution gives the posterior distribution on the unknown parameter θ, but it is very similar in form to the binomial distribution.

Thumb-Tack Example III

Transductive Inference: Given n previous "flats" in N trials, what is the probability of getting r "flats" in R trials?

Ans: Marginalize over θ—i.e.

$$P(r|n,N,R) = \int P(r|R,\theta) * f(\theta|n,N)\, d\theta$$

$$= \frac{n! * (r + R)! * (N + n - R - r)! * N!}{r! * (n-r)! * R! * (N-R)! * (N + n)!}$$

This is the beta-binomial distribution (independent of θ, but still dependent on the conditionally independent trials model).

Summary of Probabilistic Inference

- **General method for reasoning under uncertainty.**
- **Simplest generalization of classical (binary) logic**
  - allows degrees of belief (not just 0 or 1)
  - explicitly conditions belief on specific known evidence
- **Probabilistic Inference computes degrees of belief**
  (does not make decisions--this requires Decision Theory).
- **Bayesian Inference provides a way of computing beliefs given particular evidence.**
  - No such thing as "the" probability of a proposition.
  - Probabilities are not frequencies, but these are closely related.
  - Evidence can be hypothetical