COMP4620/8620: Advanced Topics in AI
Foundations of Artificial Intelligence

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

ANU

3 BAYESIAN PROBABILITY THEORY

- Uncertainty and Probability
- Frequency Interpretation: Counting
- Objective Interpretation: Uncertain Events
- Subjective Interpretation: Degrees of Belief
- Kolmogorov's Axioms of Probability Theory
- Bayes and Laplace Rule
- How to Determine Priors
- Discussion

Bayesian Probability Theory: Abstract

The aim of probability theory is to describe uncertainty. There are various sources and interpretations of uncertainty. I compare the frequency, objective, and subjective probabilities, and show that they all respect the same rules. I derive Bayes' and Laplace's famous and fundamental rules, discuss the indifference, the maximum entropy, and Ockham's razor principle for choosing priors, and finally present two brain-teasing paradoxes.

Uncertainty and Probability

The aim of probability theory is to describe uncertainty.

Sources/interpretations for uncertainty:

- Frequentist: probabilities are relative frequencies.
  (e.g. the relative frequency of tossing head.)
- Objectivist: probabilities are real aspects of the world.
  (e.g. the probability that some atom decays in the next hour)
- Subjectivist: probabilities describe an agent's degree of belief.
  (e.g. it is (im)plausible that extraterrestrians exist)

3.1 Frequency Interpretation: Counting: Contents

- Frequency Interpretation: Counting
- Problem 1: What does Probability Mean?
- Problem 2: Reference Class Problem
- Problem 3: Limited to I.I.D

Frequency Interpretation: Counting

- The frequentist interprets probabilities as relative frequencies.
- If in a sequence of n independent identically distributed (i.i.d.) experiments (trials) an event occurs $k(n)$ times, the relative frequency of the event is $k(n)/n$.
- The limit $\lim_{n\to\infty} k(n)/n$ is defined as the probability of the event.
- For instance, the probability of the event head in a sequence of repeatedly tossing a fair coin is $\frac{1}{2}$.
- The frequentist position is the easiest to grasp, but it has several shortcomings:

What does Probability Mean?

- What does it **mean** that a property holds with a certain probability?
- The frequentist obtains probabilities from physical processes.
- To scientifically reason about probabilities one needs a math theory. Problem: how to define random sequences?
- This is much more intricate than one might think, and has only been solved in the 1960s by Kolmogorov and Martin-Löf.

Problem 1: Frequency Interpretation is Circular

- Probability of event E is $p := \lim_{n\to\infty} \frac{k_n(E)}{n}$,
  $n = \#$ i.i.d. trials, $k_n(E) = \#$ occurrences of event E in n trials.

- Problem: Limit may be anything (or nothing):
  e.g. a fair coin can give: Head, Head, Head, Head, ... $\Rightarrow p = 1$.

- Of course, for a fair coin this sequence is "unlikely".
  For fair coin, $p = 1/2$ with "high probability".

- But to make this statement rigorous we need to formally know what "high probability" means. Circularity!

Problem 2: Reference Class Problem

- Philosophically and also often in real experiments it is hard to justify the choice of the so-called reference class.
- For instance, a doctor who wants to determine the chances that a patient has a particular disease by counting the frequency of the disease in "similar" patients.
- But if the doctor considered everything he knows about the patient (symptoms, weight, age, ancestry, ...) there would be no other comparable patients left.

Problem 3: Limited to I.I.D

- The frequency approach is limited to a (sufficiently large) sample of i.i.d. data.
- In complex domains typical for AI, data is often non-i.i.d. and (hence) sample size is often 1.
- For instance, a single non-i.i.d. historic weather data sequences is given. We want to know whether certain properties hold for this **particular** sequence.
- Classical probability non-constructively tells us that the set of sequences possessing these properties has measure near 1, but cannot tell which objects have these properties, in particular whether the single observed sequence of interest has these properties.

3.2 Objective Interpretation: Uncertain Events: Contents

- Objective Interpretation: Uncertain Events
- Kolmogorov's Axioms of Probability Theory
- Conditional Probability
- Example: Fair Six-Sided Die
- Bayes' Rule 1

Objective Interpretation: Uncertain Events

- For the objectivist probabilities are real aspects of the world.
- The outcome of an observation or an experiment is not deterministic, but involves physical random processes.
- The set Ω of all possible outcomes is called the sample space.
- It is said that an event $E \subset \Omega$ occurred if the outcome is in E.
- In the case of i.i.d. experiments the probabilities p assigned to events E should be interpretable as limiting frequencies, but the application is not limited to this case.
- The Kolmogorov axioms formalize the properties which probabilities should have.

Kolmogorov's Axioms of Probability Theory

Axioms 3.1 (Kolmogorov's axioms of probability theory)
Let Ω be the sample space. Events are subsets of Ω.

- If A and B are events, then also the intersection $A \cap B$, the union $A \cup B$, and the difference $A \setminus B$ are events.
- The sample space Ω and the empty set $\{\}$ are events.
- There is a function p which assigns nonnegative reals, called probabilities, to each event.
- $p(\Omega) = 1$, $p(\{\}) = 0$.
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.
- For a decreasing sequence $A_1 \supset A_2 \supset A_3 ...$ of events with $\bigcap_n A_n = \{\}$ we have $\lim_{n\to\infty} p(A_n) = 0$.

The function p is called a probability mass function, or, probability measure, or, more loosely probability distribution (function).

Conditional Probability

Definition 3.2 (Conditional probability) If A and B are events with $p(A) > 0$, then the probability that event B will occur under the condition that event A has occured is defined as

$$p(B|A) := \frac{p(A \cap B)}{p(A)}$$

- $p(\cdot|A)$ (as a function of the first argument) is also a probability measure, if $p(\cdot)$ satisfies the Kolmogorov axioms.
- One can "verify the correctness" of the Kolmogorov axioms and the definition of conditional probabilities in the case where probabilities are identified with limiting frequencies.
- But the idea is to take the axioms as a starting point to avoid some of the frequentist's problems.

Example: Fair Six-Sided Die

- Sample space: $\Omega = \{1,2,3,4,5,6\}$
- Events: Even$= \{2,4,6\}$, Odd$= \{1,3,5\}$ $\subseteq \Omega$
- Probability: $p(6) = \frac{1}{6}$, $p(\text{Even}) = p(\text{Odd}) = \frac{1}{2}$
- Outcome: $6 \in E$.
- Conditional probability: $p(6|\text{Even}) = \frac{p(6 and\ \text{Even})}{p(\text{Even})} = \frac{1/6}{1/2} = \frac{1}{3}$

Bayes' Rule 1

Theorem 3.3 (Bayes' rule 1) If A and B are events with $p(A) > 0$ and $p(B) > 0$, then $p(B|A) = \frac{p(A|B)p(B)}{p(A)}$

Bayes' theorem is easily proven by applying Definition 3.2 twice.

3.3 Subjective Interpretation: Degrees of Belief: Contents

- Subjective Interpretation: Degrees of Belief
- Cox's Axioms for Beliefs
- Cox's Theorem
- Bayes' Famous Rule

Subjective Interpretation: Degrees of Belief

- The subjectivist uses probabilities to characterize an agent's degree of belief in something, rather than to characterize physical random processes.

- This is the most relevant interpretation of probabilities in AI.

- We define the plausibility of an event as the degree of belief in the event, or the subjective probability of the event.

- It is natural to assume that plausibilities/beliefs $\mathrm{Bel}(\cdot|\cdot)$ can be repr. by real numbers, that the rules qualitatively correspond to common sense, and that the rules are mathematically consistent. $\Rightarrow$

Cox's Axioms for Beliefs

Axioms 3.4 (Cox's (1946) axioms for beliefs)

- The degree of belief in event B (plausibility of event B), given that event A occurred can be characterized by a real-valued function $\text{Bel}(B|A)$.
- $\text{Bel}(\Omega \setminus B|A)$ is a twice differentiable function of $\text{Bel}(B|A)$ for $A \neq \{\}$.
- $\text{Bel}(B \cap C|A)$ is a twice continuously differentiable function of $\text{Bel}(C|B \cap A)$ and $\text{Bel}(B|A)$ for $B \cap A \neq \{\}$.

One can motivate the functional relationship in Cox's axioms by analyzing all other possibilities and showing that they violate common sense [Tribus 1969].

The somewhat strong differentiability assumptions can be weakened to more natural continuity and monotonicity assumptions [Aczel 1966].

Cox's Theorem

Theorem 3.5 (Cox's theorem) Under Axioms 3.4 and some additional denseness conditions, $\text{Bel}(\cdot|A)$ is isomorphic to a probability function in the sense that there is a continuous one–to-one onto function $g : \mathbb{R} \to [0,1]$ such that $p := g \circ \text{Bel}$ satisfies Kolmogorov's Axioms 3.1 and is consistent with Definition 3.2.

Only recently, a loophole in Cox's and other's derivations have been exhibited [Paris 1995] and fixed by making the mentioned "additional denseness assumptions".

Conclusion: Plausibilities follow the same rules as limiting frequencies.

Other justifications: Gambling / Dutch Book / Utility theory

Bayes' Famous Rule

Let D be some possible data (i.e. D is event with $p(D) > 0$) and $\{H_i\}_{i \in I}$ be a countable complete class of mutually exclusive hypotheses (i.e. H_i are events with $H_i \cap H_j = \{\}$ $\forall i \neq j$ and $\bigcup_{i \in I} H_i = \Omega$).

Given: $p(H_i)$ = a priori plausibility of hypotheses H_i (subj. prob.)

Given: $p(D|H_i)$ = likelihood of data D under hypothesis H_i (obj. prob.)

Goal: $p(H_i|D)$ = a posteriori plausibility of hypothesis H_i (subj. prob.)

Theorem 3.6 (Bayes' rule) $$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{\sum_{i \in I} p(D|H_i)p(H_i)}$$

Proof sketch: From the definition of conditional probability and

$$\sum_{i \in I} p(H_i|...) = 1 \quad \Rightarrow \quad \sum_{i \in I} p(D|H_i)p(H_i) = \sum_{i \in I} p(H_i|D)p(D) = p(D)$$

Proof of Bayes Rule

$p(A \cup B) = p(A) + p(B)$ if $A \cap B = \{\}$, since $p(\{\}) = 0$.

$\Rightarrow$ for finite I by induction: $\sum_{i \in I} p(H_i) = p(\bigcup_i H_i) = p(\Omega) = 1$.

$\Rightarrow$ for countably infinite $I = \{1, 2, 3, ...\}$ with $S_n := \bigcup_{i=n}^{\infty} H_i$:

$$\sum_{i=1}^{n-1} p(H_i) + p(S_n) \;=\; p(\bigcup_{i=1}^{n-1} H_i \cup \bigcup_{i=n}^{\infty} H_i) \;=\; p(\Omega) \;=\; 1$$

$S_1 \supset S_2 \supset S_3 ....$

Further, $\omega \in \Omega \;\Rightarrow\; \exists n : \omega \in H_n \;\Rightarrow\; \omega \notin H_i \forall i > n \;\Rightarrow\; \omega \notin S_i \;\forall i > n$

$\Rightarrow\; \omega \notin \bigcap_n S_n \;\Rightarrow\; \bigcap_n S_n = \{\}$ (since ω was arbitrary).

$$\Rightarrow \quad 1 \;=\; \lim_{n \to \infty} \sum_{i=1}^{n-1} p(H_i) + p(S_n) \;=\; \sum_{i=1}^{\infty} p(H_i) \;=\; \sum_{i \in I} p(H_i)$$

Proof of Bayes Rule (ctnd)

By Definition 3.2 of conditional probability we have

$$p(H_i|D)p(D) \;=\; p(H_i \cap D) \;=\; p(D|H_i)p(H_i)$$

Summing over all hypotheses H_i gives

$$\sum_{i\in I} p(D|H_i)p(H_i) \;=\; \sum_{i\in I} p(H_i|D)\cdot p(D) \;=\; 1\cdot p(D)$$

$$\Rightarrow \quad p(H_i|D) \;=\; \frac{p(D|H_i)p(H_i)}{p(D)} \;=\; \frac{p(D|H_i)p(H_i)}{\sum_{i\in I} p(D|H_i)p(H_i)}$$

■

3.4 Determining Priors: Contents

- How to Choose the Prior?
- Indifference or Symmetry Principle
- Example: Bayes' and Laplace's Rule
- The Maximum Entropy Principle ...
- Occam's Razor — The Simplicity Principle

How to Choose the Prior?

The probability axioms allow relating probabilities and plausibilities of different events, but they do not uniquely fix a numerical value for each event, except for the sure event Ω and the empty event $\{\}$.

We need new principles for determining values for at least some basis events from which others can then be computed.

There seem to be only 3 general principles:

- The principle of indifference — the symmetry principle
- The maximum entropy principle
- Occam's razor — the simplicity principle

Concrete: How shall we choose the hypothesis space $\{H_i\}$ and their prior $p(H_i)$.

Indifference or Symmetry Principle

Assign same probability to all hypotheses:

$p(H_i) = \frac{1}{|I|}$ for finite I

$p(H_\theta) = [\text{Vol}(\Theta)]^{-1}$ for compact and measurable Θ.

$\Rightarrow p(H_i|D) \propto p(D|H_i) \stackrel{\wedge}{=}$ classical Hypothesis testing (Max.Likelihood).

Example: $H_\theta =$Bernoulli(θ) with $p(\theta) = 1$ for $\theta \in \Theta := [0,1]$.

Problems: Does not work for "large" hypothesis spaces:

(a) Uniform distr. on infinite $I = \mathbb{N}$ or noncompact Θ not possible!

(b) Reparametrization: $\theta \rightsquigarrow f(\theta)$. Uniform in θ is not uniform in $f(\theta)$.

Example: "Uniform" distr. on space of all (binary) sequences $\{0,1\}^\infty$:

$p(x_1...x_n) = (\frac{1}{2})^n \; \forall n \forall x_1...x_n \Rightarrow p(x_{n+1} = 1|x_1...x_n) = \frac{1}{2}$ always!

Inference so not possible (No-Free-Lunch myth).

Predictive setting: All we need is $p(x)$.

Example: Bayes' and Laplace's Rule

Assume data is generated by a biased coin with head probability θ, i.e. H_θ :=Bernoulli(θ) with $\theta \in \Theta := [0,1]$.

Finite sequence: $x = x_1 x_2 ... x_n$ with n_1 ones and n_0 zeros.

Sample infinite sequence: $\omega \in \Omega = \{0,1\}^\infty$

Basic event: $\Gamma_x = \{\omega : \omega_1 = x_1, ..., \omega_n = x_n\}$ = set of all sequences starting with x.

Data likelihood: $p_\theta(x) := p(\Gamma_x | H_\theta) = \theta^{n_1}(1-\theta)^{n_0}$.

Bayes (1763): Uniform prior plausibility: $p(\theta) := p(H_\theta) = 1$

$$(\int_0^1 p(\theta)\, d\theta = 1 \text{ instead } \sum_{i \in I} p(H_i) = 1)$$

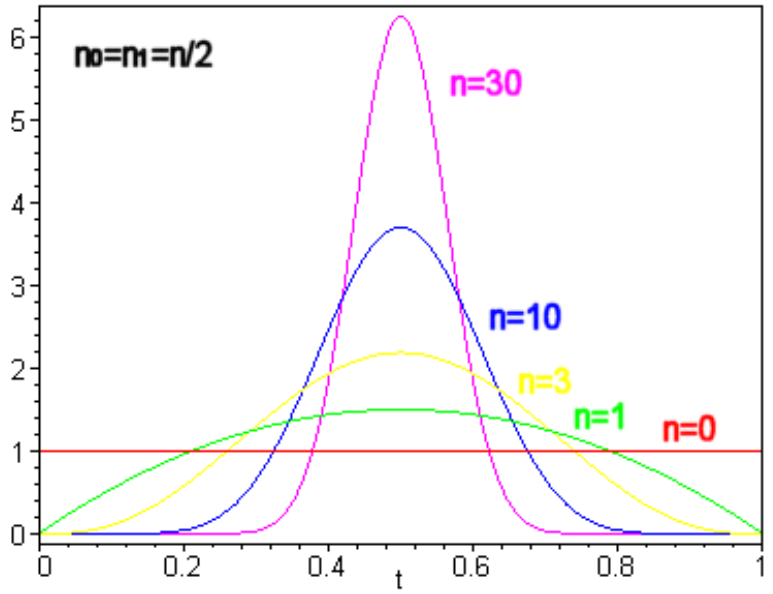
Evidence: $p(x) = \int_0^1 p_\theta(x) p(\theta)\, d\theta = \int_0^1 \theta^{n_1}(1-\theta)^{n_0}\, d\theta = \frac{n_1! n_0!}{(n_0+n_1+1)!}$

Example: Bayes' and Laplace's Rule

Bayes: Posterior plausibility of θ after seeing x is:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{(n+1)!}{n_1!n_0!}\theta^{n_1}(1-\theta)^{n_0}$$

.



Laplace: What is the probability of seeing 1 after having observed x?

$$p(x_{n+1} = 1|x_1...x_n) = \frac{p(x1)}{p(x)} = \frac{n_1+1}{n+2}$$

Laplace believed that the sun had risen for 5000 years = 1'826'213 days, so he concluded that the probability of doomsday tomorrow is $\frac{1}{1826215}$.

The Maximum Entropy Principle ...

... is based on the foundations of statistical physics.

... chooses among a class of distributions the one which has maximal entropy.

The class is usually characterized by constraining the class of all distributions.

... generalizes the symmetry principle.

... reduces to the symmetry principle in the special case of no constraint.

... has same limitations as the symmetry principle.

Occam's Razor — The Simplicity Principle

- Only Occam's razor (in combination with Epicurus' principle) is general enough to assign prior probabilities in *every* situation.
- The idea is to assign high (subjective) probability to simple events, and low probability to complex events.
- Simple events (strings) are more plausible a priori than complex ones.
- This gives (approximately) justice to both Occam's razor and Epicurus' principle.

this prior will be quantified and discussed later

3.5 Discussion: Contents

- Probability Jargon
- Applications
- Outlook
- Summary
- Exercises
- Literature

Probability Jargon

Example: (Un)fair coin: $\Omega = \{\text{Tail,Head}\} \simeq \{0,1\}$. $p(1) = \theta \in [0,1]$:

Likelihood: $p(1101|\theta) = \theta \times \theta \times (1-\theta) \times \theta$

Maximum Likelihood (ML) estimate: $\hat{\theta} = \arg\max_\theta p(1101|\theta) = \frac{3}{4}$

Prior: If we are indifferent, then $p(\theta) =$const.

Evidence: $p(1101) = \sum_\theta p(1101|\theta)p(\theta) = \frac{1}{20}$ (actually $\int$)

Posterior: $p(\theta|1101) = \frac{p(1101|\theta)p(\theta)}{p(1101)} \propto \theta^3(1-\theta)$ (BAYES RULE!).

Maximum a Posterior (MAP) estimate: $\hat{\theta} = \arg\max_\theta p(\theta|1101) = \frac{3}{4}$

Predictive distribution: $p(1|1101) = \frac{p(11011)}{p(1101)} = \frac{2}{3}$

Expectation: $\mathbb{E}[f|...] = \sum_\theta f(\theta)p(\theta|...)$, e.g. $\mathbb{E}[\theta|1101] = \frac{2}{3}$

Variance: $\text{Var}(\theta) = \mathbb{E}[(\theta - \mathbb{E}\theta)^2|1101] = \frac{2}{63}$

Probability density: $p(\theta) = \frac{1}{\varepsilon}p([\theta, \theta+\varepsilon])$ for $\varepsilon \to 0$

Applications

- Bayesian dependency networks
- (Naive) Bayes classification
- Bayesian regression
- Model parameter estimation
- Probabilistic reasoning systems
- Pattern recognition
- ...

Outlook

- Likelihood functions from the exponential family
  (Gauss, Multinomial, Poisson, Dirichlet)
- Conjugate priors
- Approximations: Gaussian, Laplace, Gradient Descent, ...
- Monte Carlo simulations: Gibbs sampling, Metropolis-Hastings,
- Bayesian model comparison
- Consistency of Bayesian estimators

Summary

- The aim of probability theory is to describe uncertainty.
- Frequency interpretation of probabilities is simple,
  but is circular and limited to i.i.d.
- Distinguish between subjective and objective probabilities.
- Both kinds of probabilities satisfy Kolmogorov's axioms.
- Use Bayes rule for getting posterior from prior probabilities.
- But where do the priors come from?
- Occam's razor: Choose a simplicity biased prior.
- Still: What do objective probabilities really mean?

Exercise 1 [C25] Envelope Paradox

- I offer you two closed envelopes, one of them contains twice the amount of money than the other. You are allowed to pick one and open it. Now you have two options. Keep the money or decide for the other envelope (which could double or half your gain).
- Symmetry argument: It doesn't matter whether you switch, the expected gain is the same.
- Refutation: With probability $p = 1/2$, the other envelope contains twice/half the amount, i.e. if you switch your expected gain increases by a factor 1.25=1/2*2+1/2*1/2.
- Present a Bayesian solution.

Exercise 2 [C15-45] Confirmation Paradox

(i) $R \to B$ is confirmed by an R-instance with property B

(ii) $\neg B \to \neg R$ is confirmed by a $\neg B$-instance with property $\neg R$.

(iii) Since $R \to B$ and $\neg B \to \neg R$ are logically equivalent,
$R \to B$ is also confirmed by a $\neg B$-instance with property $\neg R$.

Example: Hypothesis (o): All ravens are black (R=Raven, B=Black).

(i) observing a Black Raven confirms Hypothesis (o).

(iii) observing a White Sock also confirms that all Ravens are Black,
since a White Sock is a non-Raven which is non-Black.

This conclusion sounds absurd.

Present a Bayesian solution.

More Exercises

3. [C15] Conditional probabilities: Show that $p(\cdot|A)$ (as a function of the first argument) also satisfies the Kolmogorov axioms, if $p(\cdot)$ does.

4. [C20] Prove Bayes rule (Theorem 3.6).

5. [C05] Assume the prevalence of a certain disease in the general population is 1%. Assume some test on a diseased/healthy person is positive/negative with 99% probability. If the test is positive, what is the chance of having the disease?

6. [C20] Compute $\int_0^1 \theta^n (1-\theta)^m \, d\theta$ (without looking it up)

Literature (from easy to hard)

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, MA, 2003.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Pre02] S. J. Press. *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. Wiley, 2nd edition, 2002.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall / CRC, 1995.

[Fel68] W. Feller. *An Introduction to Probability Theory and its Applications*. Wiley, New York, 3rd edition, 1968.

[Szé86] G. J. Székely. *Paradoxes in Probability Theory and Mathematical Statistics*. Reidel, Dordrecht, 1986.