Bayesian Inference (I)

Intro to Bayesian Data Analysis & Cognitive Modeling
Adrian Brasoveanu

[based on slides by Sharon Goldwater & Frank Keller]

Fall 2012 · UCSC Linguistics

1. Decision Making
    - Decision Making
    - Bayes' Theorem
    - Base Rate Neglect
    - Base Rates and Experience

2. Bayesian Inference
    - Probability Distributions

3. Making Predictions
    - ML estimation
    - MAP estimation
    - Posterior Distribution and Bayesian integration

How do people make decisions? For example,

- Medicine: Which disease to diagnose?
- Business: Where to invest? Whom to trust?
- Law: Whether to convict?
- Admissions/hiring: Who to accept?
- Language interpretation: What meaning to select for a word? How to resolve a pronoun? What quantifier scope to choose for a sentence?

In all these cases, we use two kinds of information:

- Background knowledge:

  prevalence of disease
  previous experience with business partner
  historical rates of return in market
  relative frequency of the meanings of a word
  scoping preference of a quantifier
  etc.

- Specific information about this case:

  test results
  facial expressions and tone of voice
  company business reports
  various features of the current sentential and discourse context
  etc.

Example question from a study of decision-making for medical diagnosis (Casscells et al. 1978):

Example

If a test to detect a disease whose prevalence is 1/1000 has a false-positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?

Most frequent answer: 95%

Reasoning: if false-positive rate is 5%, then test will be correct 95% of the time.

Correct answer: about 2%

Reasoning: assume you test 1000 people; only about one person actually has the disease, but the test will be positive in another 50 or so cases (5%). Hence the chance that a person with a positive result has the disease is about 1/50 = 2%.

Only 12% of subjects give the correct answer.

Mathematics underlying the correct answer: Bayes' Theorem.

To analyze the answers that subjects give, we need:

Bayes' Theorem

Given a hypothesis h and data D which bears on the hypothesis:

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

$p(h)$: independent probability of h: *prior probability*
$p(D)$: independent probability of D: *marginal likelihood / evidence*
$p(D|h)$: conditional probability of D given h: *likelihood*
$p(h|D)$: conditional probability of h given D: *posterior probability*

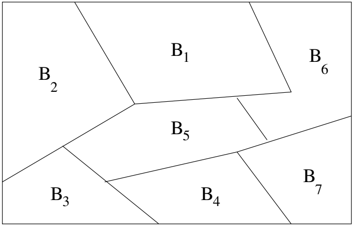
We also need the *rule of total probability.*

Theorem: Rule of Total Probability

If events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S and $p(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S:

$$p(A) = \sum_{i=1}^{k} p(A|B_i)p(B_i)$$

$B_1, B_2, \ldots, B_k$ form a *partition* of S if they are pairwise mutually exclusive and if
$B_1 \cup B_2 \cup \ldots \cup B_k = S$.



Evidence/Marginal Likelihood and Bayes' Theorem

Evidence/Marginal Likelihood

The **evidence** is also called the **marginal likelihood** because it is the likelihood $p(D|h)$ marginalized relative to the prior probability distribution over hypotheses $p(h)$:

$$p(D) = \sum_h p(D|h)p(h)$$

It is also sometimes called the **prior predictive distribution** because it provides the average/mean probability of the data D given the prior probability over hypotheses $p(h)$.

Reexpressing Bayes' Theorem

Given the above formula for the evidence, Bayes' theorem can be alternatively expressed as:

$$p(h|D) = \frac{p(D|h)p(h)}{\sum_h p(D|h)p(h)}$$

Bayes' Theorem for Data D and Model Parameters θ

In the specific case of a model with parameters θ (e.g., the bias of a coin), Bayes' theorem is:

$$p(\theta_j|D_i) = \frac{p(D_i|\theta_j)p(\theta_j)}{\sum_{j \in J} p(D_i|\theta_j)p(\theta_j)}$$

		parameter values		
data values	$\dots$	θ_j	$\dots$	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
D_i	$\dots$	$p(D_i, \theta_j)$ $= p(D_i\|\theta_j)p(\theta_j)$ $= p(\theta_j\|D_i)p(D_i)$	$\dots$	$p(D_i) = \sum_{j \in J} p(D_i\|\theta_j)p(\theta_j)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$\dots$	$p(\theta_j)$	$\dots$	

In Casscells et al.'s (1978) example, we have:

- h: person tested has the disease;
- $\overline{h}$: person tested doesn't have the disease;
- D: person tests positive for the disease.

$p(h) = 1/1000 = 0.001 \quad p(\overline{h}) = 1 - p(h) = 0.999$
$p(D|\overline{h}) = 5\% = 0.05 \quad p(D|h) = 1$ (assume perfect test)

Compute the probability of the data (rule of total probability):

$$p(D) = p(D|h)p(h) + p(D|\overline{h})p(\overline{h}) = 1 \cdot 0.001 + 0.05 \cdot 0.999 = 0.05095$$

Compute the probability of correctly detecting the illness:

$$p(h|D) = \frac{p(h)p(D|h)}{p(D)} = \frac{0.001 \cdot 1}{0.05095} = 0.01963$$

Base Rate Neglect

Base rate: the probability of the hypothesis being true in the absence of any data, i.e., $p(h)$ (the prior probability of disease).

Base rate neglect: people tend to ignore / discount base rate information, as in Casscells et al.'s (1978) experiments.

- has been demonstrated in a number of experimental situations;
- often presented as a fundamental bias in decision making.

Does this mean people are irrational/sub-optimal?

Base Rates and Experience

Casscells et al.'s (1978) study is abstract and artificial. Other studies show that

- data presentation affects performance (1 in 20 vs. 5%);
- direct experience of statistics (through exposure to many outcomes) affects performance;
  (which is why you should tweak the R and JAGS code in this class extensively and try it against a lot of simulated data sets)
- task description affects performance.

Suggests subjects may be interpreting questions and determining priors in ways other than experimenters assume.

Evidence that subjects can use base rates: diagnosis task of Medin and Edelson (1988).

Bayesian interpretation of probabilities is that they reflect *degrees of belief*, not frequencies.

- Belief can be influenced by frequencies: observing many outcomes changes one's belief about future outcomes.
- Belief can be influenced by other factors: structural assumptions, knowledge of similar cases, complexity of hypotheses, etc.
- Hypotheses can be assigned probabilities.

Bayes' Theorem, Again

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

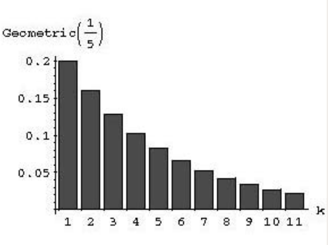
$p(h)$: *prior probability* reflects plausibility of h regardless of data.
$p(D|h)$: *likelihood* reflects how well h explains the data.
$p(h|D)$: *posterior probability* reflects plausibility of h after taking data into account.

Upshot:

- $p(h)$ may differ from the "base rate" / counting
- the base rate neglect in the early experimental studies might be due to equating probabilities with relative frequencies
- subjects may use additional information to determine prior probabilities (e.g., if they are wired to do this)

So far, we have discussed *discrete distributions*.

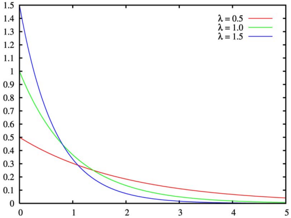
- Sample space S is finite or countably infinite (integers).
- Distribution is a *probability mass function*, defines probability of r.v. having a particular value.
- Ex: $p(Y = n) = (1 - \theta)^{n-1}\theta$ (Geometric distribution):



(Image from http://eom.springer.de/G/g044230.htm)

We will also see *continuous distributions*.

- Support is uncountably infinite (real numbers).
- Distribution is a *probability density function*, defines relative probabilities of different values (sort of).
- Ex: $p(Y = y) = \lambda e^{-\lambda y}$ (Exponential distribution):



(Image from Wikipedia)

Discrete distributions ($p(\cdot)$ is a probability mass function):

- $0 \leq p(Y = y) \leq 1$ for all $y \in S$
- $\sum_y p(Y = y) = \sum_y p(y) = 1$
- $p(y) = \sum_x p(y|x)p(x)$ (Law of Total Prob.)
- $E[Y] = \sum_y y \cdot p(y)$ (Expectation)

Continuous distributions ($p(\cdot)$ is a probability density function):

- $p(y) \geq 0$ for all y
- $\int_{-\infty}^{\infty} p(y)dy = 1$ (if the support of the dist. is $\mathbb{R}$)
- $p(y) = \int_x p(y|x)p(x)dx$ (Law of Total Prob.)
- $E[X] = \int_x x \cdot p(x)dx$ (Expectation)

Simple inference task: estimate the probability that a particular coin shows heads. Let

- θ: the probability we are estimating.
- H: hypothesis space (values of θ between 0 and 1).
- D: observed data (previous coin flips).
- n_h, n_t: number of heads and tails in D.

Bayes' Rule tells us:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \propto p(D|\theta)p(\theta)$$

How can we use this for predictions?

1. Choose θ that makes D most probable, i.e., ignore $p(\theta)$:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} p(D|\theta)$$

This is the *maximum likelihood* (ML) estimate of θ, and turns out to be equivalent to relative frequencies (proportion of heads out of total number of coin flips):

$$\hat{\theta} = \frac{n_h}{n_h + n_t}$$

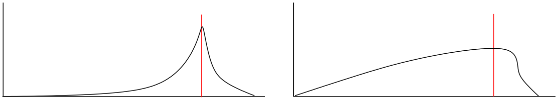
- Insensitive to sample size (10 coin flips vs 1000 coin flips), and does not generalize well (overfits).

2. Choose θ that is most probable given D:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} p(\theta|D) = \operatorname*{argmax}_{\theta} p(D|\theta)p(\theta)$$

This is the *maximum a posteriori* (MAP) estimate of θ, and is equivalent to ML when $p(\theta)$ is uniform.

- Non-uniform priors can reduce overfitting, but MAP still doesn't account for the shape of $p(\theta|D)$:



Posterior Distribution and Bayesian Integration

3. Work with the entire posterior distribution $p(\theta|D)$.

Good measure of central tendency – the expected posterior value of θ instead of its maximal value:

$$E[\theta] = \int \theta p(\theta|D) d\theta = \int \theta \frac{p(D|\theta)p(\theta)}{p(D)} d\theta \propto \int \theta p(D|\theta)p(\theta) d\theta$$

This is the *posterior mean*, an average over hypotheses. When prior is uniform (i.e., $Beta(1, 1)$, as we will soon see), we have:

$$E[\theta] = \frac{n_h + 1}{n_h + n_t + 2}$$

- Automatic smoothing effect: unseen events have non-zero probability.

Anything else can be obtained out of the posterior distribution: median, 2.5% and 97.5% quantiles, any function of θ etc.

E.g.: Predictions based on MAP vs. Posterior Mean

Suppose we need to classify inputs y as either positive or negative, e.g., indefinites as taking wide or narrow scope.

There are only 3 possible hypotheses about the correct method of classification (3 theories of scope preference): h_1, h_2 and h_3 with posterior probabilities 0.4, 0.3 and 0.3, respectively.

We are given a new indefinite y, which h_1 classifies as positive / wide scope and h_2 and h_3 classify as negative / narrow scope.

- using the MAP estimate, i.e., hypothesis h_1, y is classified as wide scope
- using the posterior mean, we average over all hypotheses and classify y as narrow scope

Casscells, W., A. Schoenberger, and T. Grayboys: 1978, 'Interpretation by Physicians of Clinical Laboratory Results', *New England Journal of Medicine* **299**, 999–1001.

Medin, D. L. and S. M. Edelson: 1988, 'Problem Structure and the Use of Base-rate Information from Experience', *Journal of Experimental Psychology: General* **117**, 68–85.