Computational Cognitive Science

Lecture 9: Bayesian Estimation

Chris Lucas
(Slides adapted from Frank Keller's)

School of Informatics
University of Edinburgh
clucas2@inf.ed.ac.uk

17 October, 2017

Reading: Griffiths and Yuille, 2006.

Cognition as Inference

The story of probabilistic cognitive modeling so far:

- models define probabilities that correspond to some aspect of human behavior;
- example: $P(R_i = A|i)$, the probability of assigning category A to item i in the GCM;
- models have parameters that determine these probability distributions (e.g., scaling factor c in the CGM);
- maximum likelihood estimation is a way of setting these parameters: we *infer* probability distributions from data.

So are probabilities just technical devices? Or do they have a *cognitive status* in our model?

Cognition as probabilistic inference

Many recent models assume that probabilities and estimation are cognitively real – we estimate and represent something like probabilities. Why?

- people act as if they have degrees of belief or certainty
- humans must deal constantly with ambiguous and noisy information
- experimental evidence: People exploit and combine noisy information in an adaptive, graded way

Cognition as probabilistic inference

People act as if they have degrees of belief or certainty. Example:

- Alice has a coin that might be two-headed.
- Alice flips the coin four times, it comes up HHHH.

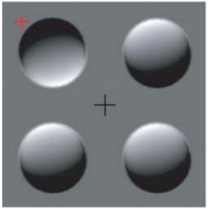
Consider the following bets:

- Would you take an even bet the coin will come up heads on the next flip?
- Would you bet 8 pounds against a profit of 1 pound?
- Would you bet your life against a profit of 1 pound?

Cognition as probabilistic inference

Humans must deal constantly with ambiguous and noisy information, e.g.,

- Visual ambiguity
- Linguistic ambiguity
- Ambiguous causes


[1]

"Infant Pulled from Wrecked Car Involved in Short Police Pursuit"[2]

[1] "A common light-prior for visual search, shape, and reflectance judgments" (Adams, 2007)

[2] Language Log – http://languagelog.ldc.upenn.edu/nll/?p=4441

Cognition as probabilistic inference

People exploit and combine noisy information in an adaptive, graded way, e.g.,

- Estimating motor forces and visual patterns from noisy data
- Combining visual and motor feedback
- Learning about cause and effect in unreliable systems
- Learning about the traits, beliefs and desires of others from their actions
- Language learning

Cognition as probabilistic inference

How do people represent and exploit information about probabilities?
Intuitively:

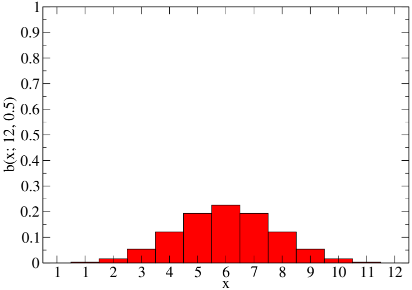
- our inferences depend on observations, but also on *prior beliefs*;
- as more observations accrue, estimates become more reliable;
- when observations are unreliable, prior beliefs are used instead.

Today we will discuss the mathematics behind these intuitions.

Distributions

Let's recap the distinction between discrete and continuous distributions. *Discrete distributions:*

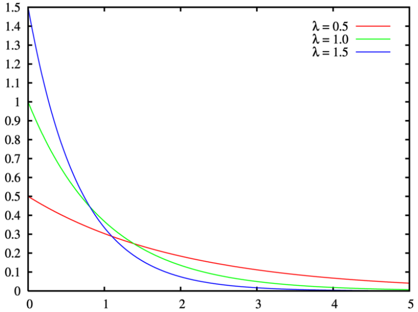
- sample space S is finite or countably infinite (e.g., integers);
- distribution is a *probability mass function,* defines probability of a random variable taking on a particular value;
- example: $P(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ (binomia):



Distributions

We have also seen examples of *continuous distributions:*

- sample space is uncountably infinite (real numbers);
- distribution is a *probability density function,* defines the probabilities if intervals of the random variable;
- example: $p(x|\theta) = \frac{1}{\theta} e^{-x/\theta}$ (exponential):



Note: Griffiths and Yuille denote density functions with $p(\cdot)$; another convention is $f(\cdot)$.

Discrete vs. Continuous

Discrete distributions:

- $P(X = x) \geq 0$ for all $x \in S$
- $\sum_{x \in S} P(x) = 1$
- $P(Y) = \sum_{x \in S} P(Y|x)P(x)$ Law of Total Probability
- $\mathbb{E}[X] = \sum_{x \in S} x \cdot P(x)$ Expectation

Discrete vs. Continuous

Discrete distributions:

- $P(X = x) \geq 0$ for all $x \in S$
- $\sum_{x \in S} P(x) = 1$
- $P(Y) = \sum_{x \in S} P(Y|x)P(x)$ Law of Total Probability
- $\mathbb{E}[X] = \sum_{x \in S} x \cdot P(x)$ Expectation

Continuous distributions:

- $p(x) \geq 0$ for all $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} p(x)dx = 1$
- $p(y) = \int p(y|x)p(x)dx$ Law of Total Probability
- $\mathbb{E}[X] = \int x \cdot p(x)dx$ Expectation

Bayes Rule

In its general form, the inference task consists of determining the *probability of a hypothesis given some data.* Notation:

- h: the hypothesis we are interested in;
- H or $\mathcal{H}$: the hypothesis space (set of all possible hypotheses);
- y: observed data (note we use y rather than d);

According to Bayes rule:

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

Bayes Rule

In its general form, the inference task consists of determining the *probability of a hypothesis given some data.* Notation:

- h: the hypothesis we are interested in;
- H or $\mathcal{H}$: the hypothesis space (set of all possible hypotheses);
- y: observed data (note we use y rather than d);

According to Bayes rule:

$$P(h|y) = \frac{P(y|h)P(h)}{P(y)} \quad \text{likelihood}$$

Bayes Rule

In its general form, the inference task consists of determining the *probability of a hypothesis given some data.* Notation:

- h: the hypothesis we are interested in;
- H or $\mathcal{H}$: the hypothesis space (set of all possible hypotheses);
- y: observed data (note we use y rather than d);

According to Bayes rule:

$$P(h|y) = \frac{P(y|h)P(h)}{P(y)} \quad \text{prior}$$

Bayes Rule

In its general form, the inference task consists of determining the *probability of a hypothesis given some data.* Notation:

- h: the hypothesis we are interested in;
- H or $\mathcal{H}$: the hypothesis space (set of all possible hypotheses);
- y: observed data (note we use y rather than d);

According to Bayes rule:

$$P(h|y) = \frac{P(y|h)P(h)}{P(y)} \quad \text{posterior}$$

Bayes Rule

In its general form, the inference task consists of determining the *probability of a hypothesis given some data.* Notation:

- h: the hypothesis we are interested in;
- H or $\mathcal{H}$: the hypothesis space (set of all possible hypotheses);
- y: observed data (note we use y rather than d);

According to Bayes rule:

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

We can compute the denominator using the law of total probability:

$$P(y) = \sum_{h' \in \mathcal{H}} P(y|h')P(h')$$

Comparing Two Hypotheses

Example: a box contains two coins, one that comes up heads 50% of the time, and one that comes up heads 90% of the time.

You pick one of the coins, flip it 10 times and observe HHHHHHHHHH. Which coin was flipped? What if you had observed HHTHTHTTHT?

Comparing Two Hypotheses

Example: a box contains two coins, one that comes up heads 50% of the time, and one that comes up heads 90% of the time.

You pick one of the coins, flip it 10 times and observe HHHHHHHHHH. Which coin was flipped? What if you had observed HHTHTHTTHT?

Let θ be the probability that the coin comes up heads. So we have two hypotheses: h_0: $\theta = 0.5$ and h_1: $\theta = 0.9$.

The probability of a sequence y with N_H heads and N_T tails is:

$$P(y|\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

A single flip has a *Bernoulli distribution* (special case of the binomial dist.).

Comparing Two Hypotheses

We can compare the probabilities of the two hypotheses directly by computing the *odds:*

$$\frac{P(h_1|y)}{P(h_0|y)} = \frac{P(y|h_1)}{P(y|h_0)}\frac{P(h_1)}{P(h_0)}$$

Comparing Two Hypotheses

We can compare the probabilities of the two hypotheses directly by computing the *odds:*

$$\frac{P(h_1|y)}{P(h_0|y)} = \frac{P(y|h_1)}{P(y|h_0)} \frac{P(h_1)}{P(h_0)} \quad \text{likelihood ratio}$$

Comparing Two Hypotheses

We can compare the probabilities of the two hypotheses directly by computing the *odds:*

$$\frac{P(h_1|y)}{P(h_0|y)} = \frac{P(y|h_1)}{P(y|h_0)} \frac{P(h_1)}{P(h_0)} \quad \text{prior odds}$$

Comparing Two Hypotheses

We can compare the probabilities of the two hypotheses directly by computing the *odds:*

$$\frac{P(h_1|y)}{P(h_0|y)} = \frac{P(y|h_1)}{P(y|h_0)} \frac{P(h_1)}{P(h_0)} \quad \text{posterior odds}$$

Comparing Two Hypotheses

We can compare the probabilities of the two hypotheses directly by computing the *odds:*

$$\frac{P(h_1|y)}{P(h_0|y)} = \frac{P(y|h_1)}{P(y|h_0)} \frac{P(h_1)}{P(h_0)}$$

We get posterior odds of 357:1 in favor of h_1 for HHHHHHHHHH and 165:1 in favor of h_0 for HHTHTHTTHT.

Comparing Infinitely Many Hypotheses

Let's now assume that θ, the probability of the coin coming up heads, can be anywhere between 0 and 1.

Now we have infinitely many hypotheses, but Bayes rule still applies:

$$p(\theta|y) = \frac{P(y|\theta)p(\theta)}{P(y)}$$

where the probability of the data is:

$$P(y) = \int_0^1 P(y|\theta)p(\theta)d\theta$$

This gives us a probability density function for theta θ given our data. What do we do with it?

Maximum Likelihood Estimation

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

$$\hat{\theta} = \arg\max_{\theta} P(y|\theta)$$

This is the *maximum likelihood* (ML) estimate of θ.

Problem: The ML estimate often generalizes poorly. It also fails to take the shape of the posterior distribution into account.

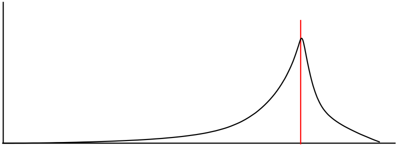
Maximum a Posteriori Estimation

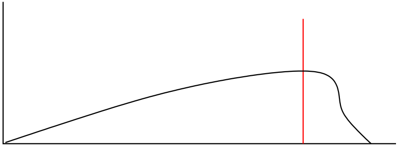
2. Choose the θ that is most probable given y:

$$\hat{\theta} = \arg\max_{\theta} p(\theta|y) = \arg\max_{\theta} P(y|\theta)p(\theta)$$

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

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





Bayesian Integration

3. Instead of maximizing, take the expected value of θ:

$$\mathbb{E}[\theta] = \int_0^1 \theta p(\theta|y)d\theta = \int_0^1 \theta \frac{P(y|\theta)p(\theta)}{P(y)} d\theta \propto \int_0^1 \theta P(y|\theta)p(\theta)d\theta$$

This is the *posterior mean,* the average over all hypotheses.

For our coin flip example with uniform $p(\theta)$, the posterior is:

$$p(\theta|y) = \frac{(N_H + N_T + 1)!}{N_H! N_T!} \theta^{N_H} (1-\theta)^{N_T}$$

This is known as the *beta distribution.*

Bayesian Integration

3. Instead of maximizing, take the expected value of θ:

$$\mathbb{E}[\theta] = \int_0^1 \theta p(\theta|y) d\theta = \int_0^1 \theta \frac{P(y|\theta)p(\theta)}{P(y)} d\theta \propto \int_0^1 \theta P(y|\theta)p(\theta) d\theta$$

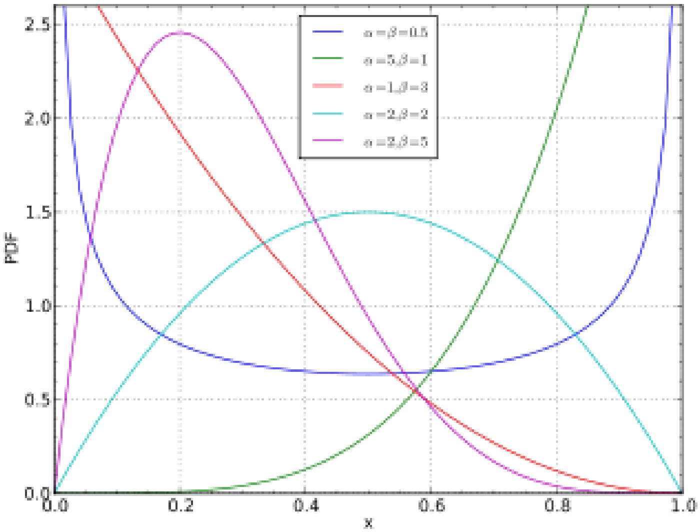
This is the *posterior mean,* the average over all hypotheses.

For our coin flip example with uniform $p(\theta)$, the posterior is:

$$p(\theta|y) = \frac{(N_H + N_T + 1)!}{N_H! N_T!} \theta^{N_H} (1-\theta)^{N_T} = \text{beta}(N_H + 1, N_T + 1)$$

This is known as the *beta distribution.*

Beta Distribution



Maximum Likelihood Estimate

Using the beta distribution, the ML estimate (equivalent to the MAP estimate with a uniform prior) works out as:

$$\hat{\theta} = \frac{N_H}{N_H + N_T}$$

This is a *relative frequency estimate:* it's simply the frequency of heads over the total number of coin flips.

This estimate is insensitive to sample size: if we get 10 heads and 0 tails then we are as certain about θ as if we get 100 heads and 0 tails. This explains the overfitting.

Posterior Mean

Let's compare this with the *posterior mean,* which for the beta distribution works out as:

$$\mathbb{E}[\theta] = \frac{N_H + 1}{N_H + N_T + 2}$$

This is the average over all values of θ. It pays attention to sample size (compare $\mathbb{E}[\theta]$ for 10 heads and 0 tails vs. 100 heads and 0 tails), and is less prone to overfitting.

We can think of this as adding *pseudocounts* to the relative frequency estimate. This is called *smoothing.*

Note that we are still assuming a uniform prior!

Choosing a Prior

Let's assume we want to use a non-uniform prior. We could again use the beta distribution:

$$p(\theta) = \text{beta}(V_H + 1, V_T + 1)$$

where $V_H, V_T > -1$ encodes our belief about likely values of θ.

This distribution has a mean of $(V_H + 1)/(V_H + VT + 2)$ and becomes concentrated around the mean as $V_H + V_T$ increases.

For example, $V_H = V_T = 1000$ puts a strong prior on $\theta = 0.5$.

The parameters that govern the prior distribution are called *hyperparameters.* (Here, V_H and V_T are hyperparameters.)

Choosing a Prior

Using the beta$(V_H + 1, V_T + 1)$ prior, the posterior distribution becomes:

$$p(\theta|y) = \frac{(N_H + N_T + V_H + V_T + 1)!}{(N_H + V_H)!(N_T + V_T)!}\theta^{N_H+V_H}(1-\theta)^{N_T+V_T}$$

which is beta$(N_H + V_H + 1, N_T + V_T + 1)$. The MAP estimate of this posterior is then:

$$\hat{\theta} = \frac{N_H + V_H}{N_H + N_T + V_H + V_T}$$

and the posterior mean becomes:

$$\mathbb{E}[\theta] = \frac{N_H + V_H + 1}{N_H + N_T + V_H + V_T + 2}$$

Choosing a Prior

Returning to our example, if we use a beta-prior with $V_H = V_T = 1000$, and our data consists of a sequence of 10 heads and 0 tails, then:

$$\mathbb{E}[\theta] = \frac{N_H + V_H + 1}{N_H + N_T + V_H + V_T + 2} = \frac{1011}{2012} \approx 0.5025$$

So we retain our belief that $\theta = 0.5$, even though we've seen strong evidence to the contrary. This would change had we seen 100 heads rather than 10.

Compare this to the maximum likelihood estimate, which is:

$$\hat{\theta} = \frac{N_H}{N_H + N_T} = 1$$

Conjugate Priors

The likelihood was Bernoulli distributed, and the prior beta distributed. This ensured the posterior was also beta distributed.

This is because the beta distribution is a *conjugate prior* for the Bernoulli distribution.

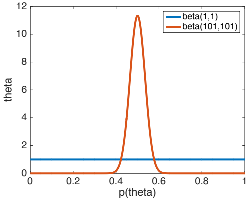
Using a conjugate prior can make the computation of the posterior tractable (e.g., by ensuring that there is an analytic solution).

Likelihood:		Conjugate Prior:	
	Bernoulli		beta
	binomial		beta
	multinomial		Dirichlet
	normal		normal

Bayesian Decision Theory

4. Keep $p(\theta|y)$ around for when we need to make a decision, or use Bayesian decision theory to define an estimator.

For example, under a beta(1,1) prior, our expectations for θ are the same having seen (1) no data; or (2) 100 heads and 100 tails. Are these data equivalent if we're considering a bet that 10 of the next 10 flips will come up heads?



Challenge: Compute the expected probability of winning the bet under each of the two data sets.

Summary

- Cognitive tasks can be modeled as probabilistic inference;
- using Bayes rule, inference can be broken down into posterior, likelihood, and prior distributions;
- standard techniques such as maximum likelihood estimation or MAP generate point estimates of the parameters;
- Bayesian techniques instead use averaging (Bayesian integration) over all parameter values;
- this makes them less prone to overfitting and allows the use of informative priors;
- the prior distribution is typically chosen to be conjugate with the likelihood distribution.

References

Griffiths, T. L. & Yuille, A. (2006). A primer on probabilistic inference. *Trends in Cognitive Sciences, 10*(7).