CS 5135/6035 Learning Probabilistic Models

Lecture 13: Introduction to Bayesian Estimation

Gowtham Atluri

October 16, 2018

Reading Material

- Larry Wasserman, Lecture Notes 14 Bayesian Inference
  http://www.stat.cmu.edu/~larry/=stat705/Lecture14.pdf
- David Barber, Bayesian Reasoning and Machine Learning
  - Chapter 9. Learning as Inference
    http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf

Topics

- Bayes' rule
  - Bayesian statistics
- Bayesian parameter estimation
  - Introduction
  - Properties
  - Why? Why not?
- Examples
  - Discrete parameters
  - Continuous parameters

Bayes' Rule

Definition

If A and B are events in F, and $P(B) > 0$, then the conditional probability of A given B, written $P(A|B)$, is

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

Theorem (Bayes' Rule)

If A and B are events in F, then Bayes' Rule states

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Application to Down Syndrome screening

If a pregnant woman was tested for Down syndrome and it was positive, what is the probability that the child will have Down syndrome?

- Let D indicate a child with Down syndrome and D^c the opposite.
- Let '+' indicate a positive test result and $-$ a negative result.

$$\begin{aligned} \text{sensitivity} &= P(+|D) &&= 0.94 \\ \text{specificity} &= P(-|D^c) &&= 0.77 \\ \text{prevalence} &= P(D) &&= 1/1000 \end{aligned}$$

$$\begin{aligned} P(D|+) &= \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|D^c)P(D^c)} = \frac{0.94 \cdot 0.001}{0.94 \cdot 0.001 + 0.23 \cdot 0.999} \\ &\approx 1/250 \\ P(D|-) &\approx 1/10,000 \end{aligned}$$

Bayes' Rule

- Bayes' Rule applied to a partition of sample space of $A = \{A_1, A_2, \ldots\}$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}$$

- Bayes' Rule also applies to probability density (or mass) functions, e.g.

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

where the integral plays the role of the sum in the previous statement.

Bayesian statistics

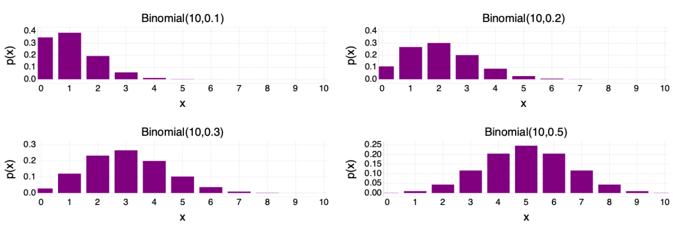
- **Parameter estimation:**
$$p(\theta|y, M)$$
where M is a model with parameter (vector) θ and y is data assumed to come from model M with true parameter θ_0.
- **Hypothesis testing/model comparison:**
$$p(M_j|y, \mathcal{M})$$
where $\mathcal{M}$ is a set of models with $M_j \in \mathcal{M}$ for $i = 1, 2, \ldots$ and y is data assumed to come from some model $M_0 \in \mathcal{M}$.
- **Prediction:**
$$p(\tilde{y}|y, M)$$
where $\tilde{y}$ is unobserved data and y and $\tilde{y}$ are both assumed to come from M.

Parameter Estimation

- E.g.: 20 apples from a large lot of apples were inspected and 3 were found to be damaged. If four apples are randomly sampled from the lot, find the probability that at least one apple in the sample of four is defective.
- To answer this probabilitic inference question we need to know the probability distribution $p(x=0), p(x=1), \ldots$.
- Parameter estimation invovles estimation of parameters given a parametric model and observed data drawn from it.
- Parametric Model:
$$p(x) = \binom{n}{x} a^x (1-a)^{n-x}, \quad x = 0, 1, 2, \ldots n$$
- Observed data: 3 out of 20 appples were found to be damaged.
- Parameter that needs to be estimated: a

Parameter Estimation

- Why is it non-trivial?
  - This (3 out of 20) can be a result several Binomial distributions, which one would have generated this.

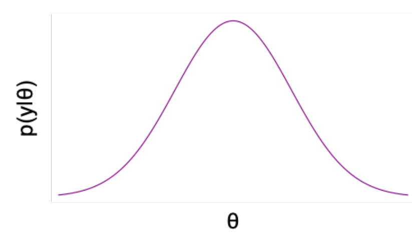


Approaches for parameter estimation - I

Maximum Likelihood Estimation (MLE)

- Parameters are assumed to be fixed but unknown
- ML solution seeks the solution that best explains the dataset y

$$\hat{\theta}_{MLE} = argmax_\theta p(y|\theta)$$

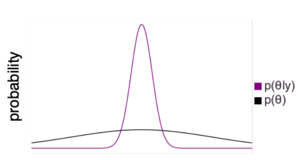


Approaches for parameter estimation - II

Bayesian Estimation

- Parameters are assumed to be random variables with some known a priori distribution $p(\theta)$
- Prior distribution is either a belief or prior knowledge
- Bayesian methods seek to estimate the posterior density $p(\theta|y)$

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



Terminology	Notation
Posterior	$p(\theta\|y)$
Prior	$p(\theta)$
Model	$p(y\|\theta)$
Prior predictive distribution (marginal likelihood)	$p(y)$

Bayesian estimation: Why and Why not?

Why do a Bayesian analysis?

- Incorporate prior belief or existing knowledge via $p(\theta)$
- Coherent with rules of probability, i.e. everything follows from specifying $p(\theta|y)$
- Captures uncertainty in the parameter estimates
- Interpretability of results, e.g. the probability the parameter is in (L, U) is 95%

Why not do a Bayesian analysis?

- Need to specify $p(\theta)$
- Computational cost of evaluating the likelihood function
- Does not guarantee coverage

Bayesian estimation: update posterior with new data

- Bayes' Rule provides a formula for updating from prior beliefs to our posterior beliefs based on the data we observe, i.e.

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

- Suppose we gather $y_1, \ldots, y_n$ sequentially (and we assume y_i independent conditional on θ), then we have

$$\begin{aligned} p(\theta|y_1) &\propto p(y_1|\theta)p(\theta) \\ p(\theta|y_1, y_2) &\propto p(y_1, y_2|\theta)p(\theta) \\ p(\theta|y_1, y_2) &\propto p(y_2|\theta)p(y_1|\theta)p(\theta) \\ p(\theta|y_1, y_2) &\propto p(y_2|\theta)p(\theta|y_1) \end{aligned}$$

  and

$$p(\theta|y_1, \ldots, y_i) \propto p(y_i|\theta)p(\theta|y_1, \ldots, y_{i-1})$$

So Bayesian learning is

$$p(\theta) \to p(\theta|y_1) \to p(\theta|y_1, y_2) \to \cdots \to p(\theta|y_1, \ldots, y_n).$$

Example: Betting on random coin flips

Consider this hypothetical situation:

- Bets are placed on the result of a coin flip
- You have been watching the results of this coin flip
- How do you determine if this is a fair coin?
- All you have is observations of the earlier coin flips

To approach this problem...

- Identify what needs to be estimated: probability of heads.
- Determine the model.
- A prior needs to be selected $p(\theta)$.
- Determine the Likelihood $p(y|\theta)$
- Compute the posterior distribution $p(\theta|y)$ using Bayes' rule.

Learning the bias of a coin

- Consider a set of samples $y = \{y_1, y_2, \ldots, y_n\}$ expressing the results of tossing a coin.

$$y_n = \begin{cases} 1 & \text{if on toss } n \text{ the coin comes up heads} \\ 0 & \text{if on toss } n \text{ the coin comes up tails} \end{cases}$$

Our aim is to estimate the probability θ that the coin will be a head, $\theta = p(y = 1)$.

- if $p(y = 1)$ deviates from 0.5, we conclude that the coin is biased.

Note that...

- The data points are binary.
- Only one parameter needs to be estimated - θ
  - Let's try the Bayesian approach!

Bayesian Estimation

- We would like to determine the probability distribution over possible values of θ.
- We first express our belief or prior knowledge of the distribution of θ as $p(\theta)$
- Write the likelihood function $p(y|\theta)$
- We compute the posterior distribution of θ

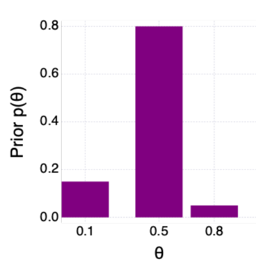
$$\begin{aligned} p(\theta|y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n, \theta)}{p(y_1, \ldots, y_n)} &= \frac{p(y_1, \ldots, y_n|\theta)p(\theta)}{p(y_1, \ldots, y_n)} \\ &\propto p(y_1, \ldots, y_n|\theta)p(\theta) \end{aligned}$$

Defining The Prior $p(\theta)$

To avoid complexities resulting from continuous variables, we'll consider a discrete θ with only three possible states, $\theta \in \{0.1, 0.5, 0.8\}$.

Specifically, we assume

- $p(\theta = 0.1) = 0.15$
- $p(\theta = 0.5) = 0.8$
- $p(\theta = 0.8) = 0.05$



The prior indicates that we believe the coin is more likely to be fair.

Likelihood $p(y|\theta)$

From our experience with MLE:

$$\begin{aligned} p(y|\theta) &= p(y_1, \ldots, y_n|\theta) \\ &= p(y_1|\theta)\ldots p(y_n/\theta) \text{ (assuming i.i.d.)} \\ &= \prod_{i=1}^{n} p(y_i|\theta) \end{aligned}$$

Because each coin toss is a Bernoulli trial, the probability of each sample y_i is

$$p(y_i|\theta) = \theta^{\mathbb{I}[y_i=H]}(1-\theta)^{\mathbb{I}[y_i=T]}$$

The likelihood then is

$$p(y|\theta) = \prod_{i=1}^{n} \theta^{\mathbb{I}[y_i=H]}(1-\theta)^{\mathbb{I}[y_i=T]}$$

Posterior $p(\theta|y)$

$$
\begin{aligned}
p(\theta|y_1,\dots,y_n) &\propto p(y_1,\dots,y_n|\theta)p(\theta) \\
&\propto \left(\prod_{i=1}^{n} \theta^{\mathbb{I}[y_i=H]}(1-\theta)^{\mathbb{I}[y_i=T]}\right)\left(p(\theta)\right) \\
&\propto p(\theta)\theta^{\sum_{i=1}^{n}\mathbb{I}[y_i=H]}(1-\theta)^{\sum_{i=1}^{n}\mathbb{I}[y_i=T]} \\
&\propto p(\theta)\ \theta^{N_H}(1-\theta)^{N_T}
\end{aligned}
$$

N_H is the number of occurrences of heads. N_T is the number of tails.

$N_H = \sum_{i=1}^{n}\mathbb{I}[y_i = H]$

$N_T = \sum_{i=1}^{n}\mathbb{I}[y_i = T]$

Coin Posterior

- For an experiment we have 2 heads and 8 tails
  - $N_H = 2$ and $N_T = 8$
- Our prior distribution is

$$p(\theta = 0.1) = 0.15,\ p(\theta = 0.5) = 0.8,\ p(\theta = 0.8) = 0.05$$

- Our posterior equation is

$$p(\theta|y_1,\dots,y_n) \propto p(\theta)\ \theta^{N_H}(1-\theta)^{N_T}$$

- Our posterior for $\theta = 0.1$ is

$$p(\theta = 0.1|y_1,\dots,y_n) \propto p(\theta = 0.1)0.1^{N_H}(1-0.1)^{N_T}$$

Similarly,

$$p(\theta = 0.5|y_1,\dots,y_n) \propto p(\theta = 0.5)0.5^{N_H}(1-0.5)^{N_T}$$

$$p(\theta = 0.8|y_1,\dots,y_n) \propto p(\theta = 0.8)0.8^{N_H}(1-0.8)^{N_T}$$

Coin Posterior

$$
\begin{aligned}
p(\theta = 0.1|y_1,\dots,y_n) &\propto 6.4\times 10^{-4} \\
p(\theta = 0.5|y_1,\dots,y_n) &\propto 7.8\times 10^{-4} \\
p(\theta = 0.8|y_1,\dots,y_n) &\propto 8.2\times 10^{-8}
\end{aligned}
$$

We know that

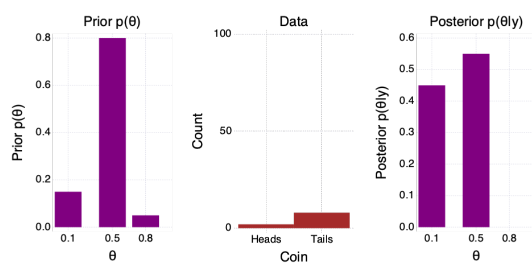
$$p(\theta = 0.1|y_1,\dots,y_n) + p(\theta = 0.5|y_1,\dots,y_n) + p(\theta = 0.8|y_1,\dots,y_n) = 1$$

By dividing each of the values by
$p(\theta = 0.1|y_1,\dots,y_n) + p(\theta = 0.5|y_1,\dots,y_n) + p(\theta = 0.8|y_1,\dots,y_n)$,

We get,

$$
\begin{aligned}
p(\theta = 0.1|y_1,\dots,y_n) &= 0.45 \\
p(\theta = 0.5|y_1,\dots,y_n) &= 0.55 \\
p(\theta = 0.8|y_1,\dots,y_n) &= 0
\end{aligned}
$$

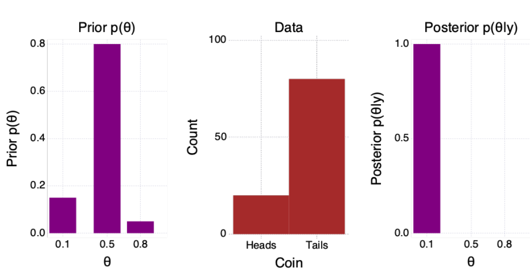
Coin Posterior ($N_H = 2, N_T = 8$)



- Most likely state is $\theta = 0.5$; $\theta = 0.1$ is also appreciable.
- *aposteriori* is close to the *apriori* belief than to the observed data.

Coin Posterior ($N_H = 20, N_T = 80$)

Repeating the above with $N_H = 20$, $N_T = 80$, the posterior changes to



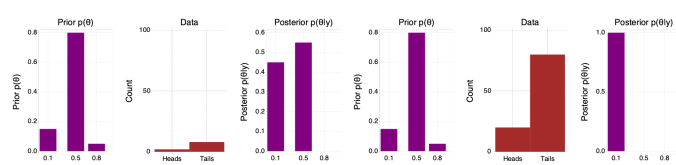
- The posterior belief in $\theta = 0.1$ dominates.
  - Heads are unlikely to result from this coin.
- *aposteriori* is very different from our *apriori* belief

Posteriori Effect

- Posterior is a compromise between the data and the prior

$$p(\theta|y_1,\dots,y_n) \propto p(y_1,\dots,y_n|\theta)p(\theta)$$

- In both cases $N_T/N_H = 4$
  - only when $N_H = 20, N_T = 80$ we are much more confident that $\theta = 0.1$



Continuous Parameters

- In the previous example, we considered only discrete values of θ. Here will consider continuous values.
- We first examine the case of a 'flat' prior $p(\theta) = k$
  - for some constant k.
- Any probability density function must hold

$$\int_0^1 p(\theta) d\theta = 1 \implies k\theta|_0^1 = 1 \implies k = 1$$

- Priori $p(\theta) = 1$.
- Our posterior equation is

$$\begin{aligned} p(\theta|y_1, \ldots, y_n) &\propto p(\theta)\ \theta^{N_H}(1-\theta)^{N_T} \text{ (using a flat prior } p(\theta) = 1) \\ &\propto \theta^{N_H}(1-\theta)^{N_T} \\ &= \frac{1}{c}\theta^{N_H}(1-\theta)^{N_T} \end{aligned}$$

Continuous Parameters

- Our posterior is $\quad p(\theta|y_1, \ldots, y_n) = \frac{1}{c}\theta^{N_H}(1-\theta)^{N_T}$
  - The constant c ensures that $\int_0^1 p(\theta|y_1, \ldots, y_n) = 1$.
- We recognize, the part $\theta^{N_H}(1-\theta)^{N_T}$ is the same as the functional form of Beta distribution

$$Beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad \text{for } 0 < x < 1$$

- Comparing the corresponding exponents in the posterior and Beta dist., we get
  - $\alpha = N_H + 1$ and $\beta = N_T + 1$
- From this we can say that the Posterior has the form $Beta(N_H + 1, N_T + 1)$

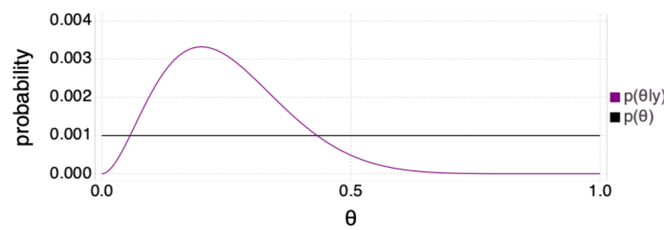
Coin posterior - using a continuous prior

- From an experiment we have $N_H = 2$ and $N_T = 8$
- Prior distribution is $p(\theta) = 1$
- Likelihood is $\theta^{N_H}(1-\theta)^{N_T}$
- Posterior $p(\theta|y_1, \ldots, y_n) = \frac{1}{c}\theta^{N_H}(1-\theta)^{N_T}$
  - Same as $Beta(N_H + 1, N_T + 1)$.
- We can compute the probabilities $p(\theta|y_1, \ldots, y_n)$ directly from the pdf $Beta(N_H + 1, N_T + 1)$

Computing the posterior probabilities in Julia

Posterior Distribution for our observations $N_H = 2$ and $N_T = 8$

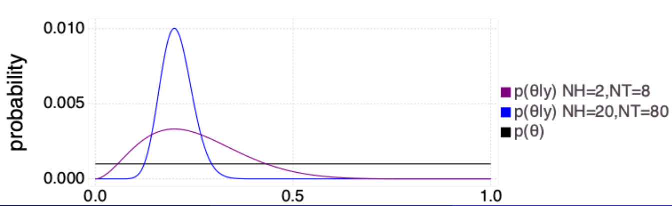
```
x = collect(0:0.001:1);
prior = ones(length(x)); prior = prior./sum(prior);
d = Beta(3,9);
posterior = pdf.(d,x); posterior = posterior./sum(posterior);
myplot = plot(layer(x=x,y=posterior,Geom.line),
              layer(x=x,y=prior,Geom.line));
```



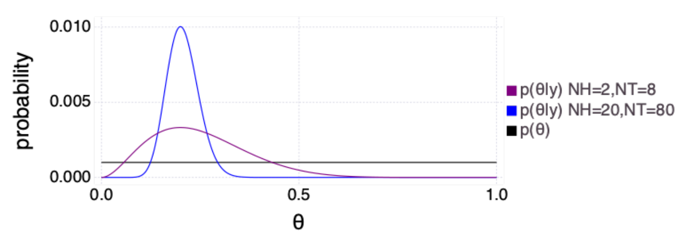
Computing the posterior probabilities in Julia

Posterior Distribution for our observations $N_H = 20$ and $N_T = 80$

```
x = collect(0:0.001:1);
prior = ones(length(x)); prior = prior./sum(prior);
d1 = Beta(3,9);
d2 = Beta(21,81);
posterior1 = pdf.(d1,x); posterior1 = posterior1./sum(posterior1);
posterior2 = pdf.(d2,x); posterior2 = posterior2./sum(posterior2);
myplot = plot(layer(x=x,y=posterior1,Geom.line),
              layer(x=x,y=posterior2,Geom.line),
              layer(x=x,y=prior,Geom.line));
```



Coin Posterior



- Maximum of the posterior is at $\theta = 0.2$
  - for both cases $N_H = 2, N_T = 8$ and $N_H = 20, N_T = 80$
  - because $N_T/N_H = 4$
- The posterior is much narrower for $N_H = 20, N_T = 80$