EE126: Probability and Random Processes

Lecture 23: Bayesian Estimation

Abhay Parekh

UC Berkeley

April 19, 2011

Logistics

- Turn in your take-home if you attempted it
- I'll hand out the exams at the end of class
- Mean 44.25, $\sigma = 19.45$, median $= 43$
- Regrades until Thursday
- HW due Friday; next HW due Wed.

MC Results: Common Method

1. Write a recurrence in terms of the state just before or just after the state of interest
2. Show that the resulting set of equations has a unique solution.

- C-K: Just before i: $\pi_i = \sum_j \pi_j p_{ji}$
- First recurrent state from i: Just after i: $a_i = \sum_j p_{ij} a_i$.
- Mean time to absorbtion from i: Just after i: $\mu_i = 1 + \sum_j p_{ij} \mu_j$
- Mean time from i to s: Just after i: $t_i = 1 + \sum_j p_{ij} t_j$

Steady State Convergence

Any MC with a single aperiodic recurrent class must converge in the sense that

1. For each state j:

$$\lim_{n\to\infty} r_{ij}(n) = \pi_j, \quad \text{for all } i$$

2. π_j are given by the system of equations:

$$\begin{aligned} \pi_j &= \sum_{k=1}^{m} \pi_k p_{kj}, \quad j = 1, 2, ..., m \\ 1 &= \sum_{k=1}^{m} pi_k \end{aligned}$$

3. $\pi_j = 0$ for all transient states j.
   $\pi_j > 0$ for all recurrent states j.

First Recurrent State: Absorbtion

Consider a MC with some transient states. Let s be a recurrent state. We want the prob that s is the first recurrent state visited when $X_0 = i$. Let this be a_i.

- Make each recurrent state absorbing. Then the MC has transient and absorbing states.
- Let T be the set of transient states.

$$a_i = \sum_{j=1}^{m} p_{ij} a_j, i \in T, \quad a_s = 1, \quad a_i = 0, \text{if } i \text{ is not } s \text{ but recurrent}$$

This system of equations always has a unique solution.

Expected Time to Absorbtion

The expected times to absorbtion $\mu_1, ..., \mu_m$ are the unique solution to the equations

$$\mu_i = 0, \ \ i \text{ recurrent}$$

$$\mu_i = 1 + \sum_{j=1}^{m} p_{ij}\mu_j \quad i \text{ transient}$$

Return times

Mean First Passage Time and Recurrence Times

Consider a MC with a single recurrent class and let s be a particular recurrent state.

- The Mean First Passage time is the expected time for a recurrent state s to be reached from some state i:
  The mean first passage times t_i to reach s starting from i are given by

$$t_s = 0, \;\; t_i = 1 + \sum_{j=1}^{m} p_{ij} t_j, \;\; \text{for all } i \neq s$$

- The mean recurrence time is the expected time that a state s takes to return to itself.
  The mean recurrence time t_s^* of state s is given by

$$t_s^* = 1 + \sum_{j=1}^{m} p_{sj} t_j$$

Inference

So far we've studied problems that have

- well defined models (Poisson Process, Markov Chain, etc)
- clearly defined variable (Bias of a coin, time to reach destination etc)

Often, things aren't so clear-cut! We have to answer questions based on a whole lot of noisy data!

1. Model Inference: What does the traffic on the internet look like? Is it Poisson? Must infer from the data...Recommendation Engines. What if the data fits more than model - which one to pick?
2. Variable Inference: Suppose a flip a coin 5 times. What can I infer about the value of the bias?

Inferring from observed data (which frequently include noise) is obviously a very important problem. But how to do this is not as clear-cut as the problems we have been studying...

Two ways to treat unknowns

- The goal of the inference is to come up with the best estimate of the VALUE of the unknown: Example: The bias of the coin is 0.75.
- The goal of the inference is to come up with the best estimate of the DISTRIBUTION of the unknown: Example: The bias of the coin is uniformly distributed between $[0.75, 0.85]$.

The first approach follows CLASSICAL statistics and the second approach follows BAYESIAN statistics.

Example: Bias of a Coin

Bob tosses a coin 3 times and gets three heads. What should he estimate the probability of Heads, h, to be? To keep things simple, assume that $h \in \{0, 0.1, 0.2, ..., 1\}$. Let X be the number of heads in 3 tosses.

1. Find the value, h^*, that maximizes $P(X = 3|h)$. Thus we estimate the bias to be 1.
2. Find $P(h|X = 3)$.

$$\frac{P(h, X = 3)}{P(X = 3)} = \frac{P(X = 3|h)P(h)}{\sum_h P(X = 3|h)P(h)}.$$

This gives us a probability distribution that depends on $P(h)$:

Two Basic Questions

1. Point estimate or distribution?
2. Which conditional should we use: $P(h|X = 3)$ or $P(X = 3|h)$?

Comparison between conditionals

$$P(h|X=3) = P(X=3|h)\frac{P(h)}{\sum_h P(X=3|h)P(h)}.$$

- If h is uniformly distributed then:

$$P(h|X=3) = P(X=3|h)\frac{1}{\sum_h P(X=3|h)}.$$

  Since denominator does not depend on h the two approaches are pretty much the same.

- $P(h=0.5) = 0.9, P(h=1) = .1$:

$$P(h|X=3) = \begin{cases} \frac{(0.5)^3(0.9)}{(0.5)^3(0.9)+(1)(0.1)} = 0.529, & \text{p=0.5;} \\ \frac{(1)(0.1)}{(0.5)^3(0.9)+(1)(0.1)} = 0.471, & \text{p=1;} \\ 0, & \text{o.w.} \end{cases}$$

  Now the difference is significant!

If you know $P(h)$ you should use it!

Distribution or Point Estimate?

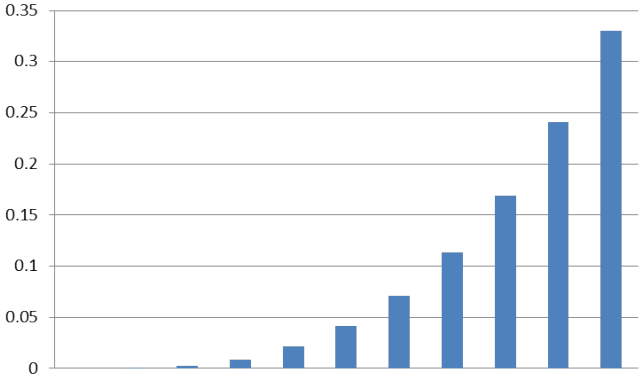
Suppose $P(h)$ is uniform, i.e. $P(h) = \frac{1}{11}$, $h = 0, 0.1, 0.2, ..., 1$

$$P(h, X = 3) = P(X = 3|p)P(h) = \frac{p^3}{11}$$

$$P(X = 3) = \sum_{i=1}^{10} P(X = 3|p = \frac{1}{i})\frac{1}{11} = \frac{1}{11}(1+0.9^3+0.8^3+...+0.1^3)$$

So

$$P(p|X = 3) = \frac{p^3}{3.025}$$



Point estimate of $h = 1$ has an error of 0.67. Distribution tells us that $P(h \in \{0.8, 0.9, 1\} \geq 0.74$

Bayesian Estimation

If we are trying to estimate a parameter Θ and have observations $X = (X_1, X_2, ..., X_n)$:

- The distribution of p_θ (f_Θ) is assumed to be known. This is called the **Prior Distribution**.
- A complete solution to the estimation problem is provided by $p_{\theta|X}$ (or $f_{\Theta|X}$). This is called the **Posterior Distribution**.

Bayesian Estimation and Priors

- What if we don't have a good initial estimate of the prior? Can we learn as make observations? Otherwise this approach could be very inaccurate!
- Estimate at the end of each observation. Start with say a uniform prior. After making an observation, calculate the posterior distribution. Then use this posterior as the prior for the next observation!

Learning by Updating Prior

Suppose $P(h)$ actually distributed so that $P(h = 1) = 1.0$, but the possible values are $h = 0, 0.1, ..., 0.9, 1$.

- Start with a uniform prior. Observe a head. The posterior is:

$$P(h|H) = \frac{P(H|h)P(h)}{\sum_h P(H|h)P(h)} = \frac{h}{5.5}$$

- Set $P(h) = h/5.5$. Observe another H. Now the posterior is:

$$P(h|H) = \frac{P(H|h)P(h)}{\sum_h P(H|h)P(h)} = \frac{hh/5.5}{\sum_h h^2/5.5} = \frac{h^2}{3.85}$$

- Set $P(h) = h^2/3.85$. Observe another H. Now the posterior is

$$P(h|H) = \frac{P(H|h)P(h)}{\sum_h P(H|h)P(h)} = \frac{hh^2/3.85}{\sum_h h^3/3.855} = \frac{h^3}{3.025}$$

  This is what we got before from direct calculation.

The method converges to the correct distribution as more observations are made.

Example

Bob is usually late for his meetings with Alice. In fact he is late by a random amount X uniformly distributed on $[0, \Theta]$ where Θ is unknown. Alice models the prior of Θ to be uniform in $[0, 1]$.
Suppose Bob is late by x for the next meeting.
$f_{X|\Theta}(x|\theta) = \frac{1}{\theta}$ if $x \in [0, \theta]$ and 0 otherwise.
So Alice finds the posterior: It is non-zero as long as $x \leq \theta \leq 1$ and is equal to

$$f_{\Theta|X}(\theta|x) = \frac{1/\theta}{\int_x^1 \frac{1}{\theta'} d\theta'} = \frac{1}{\theta|\ln x|}$$

For the next meeting, she uses the posterior as as the new prior.
The new prior is **not** uniform.

Example: Inference of a Common Mean

We make n measurements of an unknown quantity θ and make n independent observations:

$$X_i = \theta + W, i = 1, 2, ..., n$$

where W is noise distributed $N(0, \sigma^2)$ and independent of θ. We start with the prior $\theta \sim N(x_0, \sigma^2)$.
Then

$$f_\Theta(\theta) = c_1 \exp\{-\frac{(\theta - x_0)^2}{2\sigma^2}\}$$

and

$$f_{X_1,...,X_n|\theta}(x_1, ..., x_n|\theta) = c_2 \exp\{-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}\}$$

So one can calculate the posterior. We get that $\theta|X_1, ..., X_n \sim N(m, v)$ where

$$m = \frac{x_0 + ... + x_n}{n+1}, \quad v = \frac{\sigma^2}{n+1}$$

So the posterior has the same kind of distribution as the prior!

Recursive Calculation of Common Mean

Since the prior and posterior have are normally distributed we just need to keep track of two numbers: mean and variance. Suppose n observations have been made and we now make the $n+1^{st}$ one. The new posterior has mean

$$m_{n+1} = \frac{(m_n/v_n) + (x_{n+1}/\sigma^2)}{(1/v_n)(1+\sigma^2)}$$

and variance

$$v_{n+1} = \frac{1}{(1/v_n)(1/\sigma^2)}.$$

This allows for easy updating of the estimate. Unfortunately, most of the time, finding the posterior distribution is computationally intensive.

Maximum a Posteriori (MAP) Rule

Suppose we are forced to pick a point estimate instead of a distribution.
Then as we have seen, the Bayesian is going to choose the rule:

$$\hat{\theta} = \text{argmax}_\theta p_{\Theta|x}(\theta|x) \quad (\Theta \text{ discrete})$$

$$\hat{\theta} = \text{argmax}_\theta f_{\Theta|x}(\theta|x) \quad (\Theta \text{ continuous})$$

Example: Suppose there are two possible values: θ_1 and θ_2. Then MAP picks θ_1 iff $p_{\Theta|X)}(\theta_1|x) \geq p_{\Theta|X}(\theta_1|x)$, I.e.,

$$\frac{P(X = x|\theta = \theta_1)}{P(X = x|\theta = \theta_2} \geq \frac{P(\theta_2)}{P(\theta_1)}$$

Notice that $P(X = x)$ always cancels out...

How good is the MAP Rule?

The MAP rule can be pretty good esp. when θ is discrete. Let $g_M(X)$ be the MAP rule.
Suppose there is another procedure that produces an estimate $g_o(X)$ given observations X. Let I_o be a boolean r.v. that is 1 when the procedure is correctly estimates θ and zero o.w.

$$E[I_o|X] = P(g(X) = \theta|X) \leq P(\Theta = \hat{\theta}|X) = E[I_m|X]$$

Iterated expectations:

$$E[I_o] \leq E[I_m]$$

So the MAP rule maximizes the probability of estimating, over all rules. All bets are off when θ is continuous...

Example: Binary Hypothesis Testing

- Coins of two types: bias is either p_1 or p_2. Equally likely to pick either.
- Toss the selected coin n times and get k heads. Which coin did we pick?

MAP picks θ_1 iff $p_{\Theta|X)}(\theta_1|x) \geq p_{\Theta|X}(\theta_1|x)$, I.e.,

$$\frac{P(X = x|\theta = \theta_1)}{P(X = x|\theta = \theta_2} = \frac{p_1^k(1-p_1)^{n-k}}{p_2^k(1-p)^{n-k}} \geq \frac{P(\theta_2)}{P(\theta_1)} = 1$$

This is a threshold rule: Assuming that $p_1 < p_2$, there is an integer k^* such that you pick θ_1 as long as $k \leq k^*$ and pick θ_2 otherwise.

$$P(\text{error}) = \mathrm{P}(\Theta = \theta_1, \mathrm{X} > \mathrm{k}^*) + \mathrm{P}(\Theta = \theta_2, \mathrm{K} \leq \mathrm{k}^*)$$

Of all threshold rules, the MAP-based threshold minimizes the $P(\text{error})$.

Example: Matched Filter

The transmitter transmits two types of messages: $\Theta = 1$ if it sends $A = (a_1, a_2, ..., a_n)$ and $\Theta = 2$ if it sends $B = (b_1, b_2, ..., b_n)$. We require that $\sum_i a_i^2 = \sum_i b_i^2$ (equal energy).

$$X_i = S_i + W_i$$

where S_i is the signal sent and $W_i \sim N(0, 1)$ and are iid. Under $\theta = 1$: The X_i are independent rvs $N(a_i, 1)$.
Result:

$$\text{Pick } \Theta = 1 \text{ if } \sum_{i=1}^{n} a_i x_i > \sum_{i=1}^{n} b_i x_i$$

(Pick $\Theta = 2$ otherwise.).
What does this mean? Project the received vector on A and B. Pick the message for the magnitude is the greatest.

Issue with Point Estimators such as MAP

1. Sometimes the maximum of a distribution is not typical.
2. Example: Higher dimensional data: Say θ is a point in $\Re^n$ and the posterior is a joint gaussian of n independent standard normals. Maximum of the distribution is at the origin (the mean). For $n = 1000$, 90% of the probability is in a shell of radius 31.6 and thinkness 2.8.
3. So use with care.

What about picking $E[\Theta|X]$ as a point estimator?