國立清華大學
NATIONAL TSING HUA UNIVERSITY

EE 306001
Probability

Lecture 26: Introduction to Statistics

李祈均

If interested in a single answer:

- Hypothesis testing (discrete unknown)
  - Pick θ that has maximum a posteriori probability (MAP)

$$p_{\Theta|X}(\theta^*|x) = \max_\theta p_{\Theta|X}(\theta|x)$$

  - Minimizes probability of error; used in hypothesis testing

- Estimation case (continuous unknown)
  - Pick θ that has maximum a posteriori probability (MAP)

$$f_{\Theta|X}(\theta^*|x) = \max_\theta f_{\Theta|X}(\theta|x)$$

  - The point at the maximum point of the density function
- OR?
  - Use conditional expectation (LMS)

$$\theta^* = E[\Theta|X = x] = \int \theta\, f_{\Theta|X}(\theta|x)d\theta$$

  - Average of the density function (center of gravity)

Least mean square estimation (LMS estimator)

Try to minimize the following (find θ^*):

$$E[(\Theta - \theta^*)^2]$$

- In the absence of any observation, $\theta^* = E[\Theta]$

$$E[(\Theta - E[\Theta])^2] \leq E\left[\left(\Theta - \hat{\theta}\right)^2\right], \text{for all } \hat{\theta}$$

- For any given value x of X, $E[(\Theta - \theta^*)^2|X = x]$ is minimized $\theta^* = E[\Theta|X = x]$

$$E[(\Theta - E[\Theta])^2|X = x] \leq E\left[\left(\Theta - \hat{\theta}\right)^2|X = x\right], \text{for all } \hat{\theta}$$

- Out of all estimators $g(X)$ of Θ based on X, the mean squared estimation error $E[(\Theta - g(X))^2]$ is minimized when $g(X) = E[\Theta|X]$

Example

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter, θ, is unknown and is modeled as the value of a random variable, Θ, which is uniformly distributed between zero and one hour.

Now, assume Juliet is late by x on their first date, how should Romeo use this information to update the distribution of Θ

- First note the prior distribution pdf is

$$f_\Theta(\theta) = \begin{cases} 1, \text{if } 0 \le \theta \le 1 \\ 0, \text{otherwise} \end{cases}$$

The conditional pdf of the observation (data given parameter):

$$f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{\theta}, \text{if } 0 \le x \le \theta \\ 0, \text{otherwise} \end{cases}$$

Now, we can directly use Bayes rule ($f_{\Theta|X}(\theta|x)$), and note that $f_\Theta(\theta) f_{X|\Theta}(x|\theta)$ is nonzero only if $0 \le x \le \theta \le 1$

- The new posterior probability distribution function is:

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)}{\int_0^1 f_{\Theta}(\theta')\, f_{X|\Theta}(x|\theta') d\theta'}$$

$$= \frac{1/\theta}{\int_x^1 1/\theta' d\theta'} = \frac{1}{\theta|\log x|}, \text{if x} \le \theta \le 1$$

$$f_{\Theta|X}(\theta|x) = \begin{cases} \dfrac{1}{\theta|\log x|}, \text{if x} \le \theta \le 1 \\ 0, \text{otherwise} \end{cases}$$

Okay so with this, can we come up with a single point estimate?

$$f_{\Theta|X}(\theta|x) = \begin{cases} \dfrac{1}{\theta|\log x|}, \text{if } \mathrm{x} \leq \theta \leq 1 \\ 0, \text{otherwise} \end{cases}$$

First, lets think about this function,

- Any given x, $f_{\Theta|X}(\theta|x)$ is decreasing with θ over the range $[x, 1]$
  - So what's best estimate if we are going for MAP (maximum a-posteriori estimation)?
  - $\theta = x$

How about if we have another point estimate?

- Instead of MAP, let's try conditional expectation

$$E[\Theta|X = x] = \int_x^1 \theta \frac{1}{\theta|\log x|} d\theta$$

$$= \frac{1 - x}{|\log x|}$$

Note:

- Before the first date:
- Compute the probability that Juliet is going to be late by $X = x_1$, we can now use LMS:

$$X \sim \mathrm{uniform}(0, E[\Theta]) = \mathrm{uniform}(0, 0.5)$$

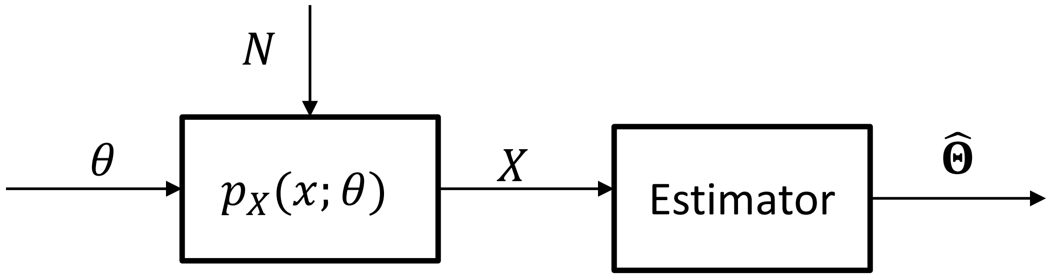
- After the first date:

Now, if we use LMS estimator $\theta_{LMS} = \frac{1-x}{|\log x|}$

- Then to compute the probability that Juliet is going to be late by $X = x_2$, we can now assume:

$$X \sim \mathrm{uniform}(0, E[\Theta|X = x]) = \mathrm{uniform}\left(0, \frac{1 - x_1}{|\log x_1|}\right)$$

Classical statistics



Classical statistics
θ Is a parameter (not a r.v.)

θ: the unknown, nothing random about it, it's just a number

$p_X(x;\theta)$: this distribution depends on θ, however, it is **NOT** conditional pdf (conditional pdf is for r.v.s)!

You can imagine this could just be a parameter to describe the distribution X (somehow depends on θ) – maybe a normal distribution with mean θ

Data X could be a vector $[X_1, X_2, \ldots, X_n]$, and θ could be a vector of parameters too!

Desired probability of an estimator

- This estimator, $\widehat{\Theta}_n$, is random
- Unbiased: $E\left[\widehat{\Theta}_n\right] = \theta$
- Consistent: $\widehat{\Theta}_n \to \theta$ (convergence in probability)
- Small mean square error (MSE)
  - $E\left[\left(\widehat{\Theta} - \theta\right)^2\right] = var\left(\widehat{\Theta} - \theta\right) + \left(E\left[\widehat{\Theta} - \theta\right]\right)^2 = var\left(\widehat{\Theta}\right) + (\text{bias})^2$

Estimator you already know in classical sense

- Sample mean
  - This is an estimator with very good property
  - Very easy and good estimator for mean of a distribution
- ML estimate
  - Maximum data likelihood
  - Observable evidence is the KING
- They don't necessary coincide

Confidence interval

- Idea: you want to know how much you can trust a given estimate (say, for example: you estimate the mean to be 2.37)

- Can we construct an interval to say the likely value of ‘true thetas’?

Confidence interval

- Design a $1 - a$ confidence interval
  - $[\widehat{\Theta}_n^-, \widehat{\Theta}_n^+]$
  - Such that: $P(\widehat{\Theta}_n^- \widehat{\Theta}_n^+) \geq 1 - a \text{ for all } \theta$
  - Often this $a = 0.05, 0.025, \text{or } 0.01$

- Note: this interval is 'random' – uppercase rvs!
  - So what you are doing exactly is to construct two other random variables as estimators!

- Interpretation (subtle)
  - Say you have an interval, and given the data you observe, you realize the value of the interval (uppercase -> lowercase)
  - Say it’s between 1.97 – 2.56 ($a = 0.05$)
  - Can you say:
    - With probability 0.95, the true theta falls in that interval (1.97, 2.56)?
    - Nope, probability statement is associated with randomness statement
    - θ is a number, the two realized intervals are numbers, so it’s either you are ‘IN’ the interval or ‘NOT’
  - Proper way to state this:
    - the interval, that's being constructed by our procedure, should have the property that, with probability 95%, it's going to fall on top of the true value of theta

- Imagine this procedure as experiment
  - You do it once on a day, seeing the data, construct the interval, and yes, the true theta is in
  - You do it on another day, seeing different data, construct the interval, and yes, the true theta is in
  - You do it on another way, seeing different data, but this time, nope!
  - 95% of the days when I use this procedure to construct the confidence interval, I got it right!

- it's a statement about the distribution of these random confidence intervals, how likely are they to fall on top of the true theta
  - It's a statement about probabilities associated with a confidence interval (intervals are random variables)
  - Not about the theta!

Example: polling

Consider the polling problem, where we wish to estimate the fraction θ of voters who support a particular candidate for office

We collect n independent sample voter responses, where each X_i is a Bernoulli random variable, with $X_i = 1$ if the i^{th} voter supports the candidate

We estimate θ with sample mean $\widehat{\Theta}_n$ and construct a confidence interval based on normal approximation and different ways of estimating unknown variances

- For example: 684 out of a sample of $n = 1200$ voters support the candidate, so that $\widehat{\Theta}_n = 0.57$

Case I:

Using the unbiased sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_n\right)^2$$

$$= \frac{1}{1199}\left(684 * \left(1 - \frac{684}{1200}\right)^2 + (1200 - 684) * \left(0 - \frac{684}{1200}\right)^2\right)$$

$$\approx 0.245$$

Now, we can use this to find the CI interval

Assume $\widehat{\Theta}_n$ is normal random variable, with mean θ, variance $\hat{S}_n^2/n$, the 95% CI is the following: ($\hat{S}_n = 0.245$)

$$\left[\widehat{\Theta}_n - 1.96\frac{\hat{S}_n}{\sqrt{n}}, \widehat{\Theta}_n + 1.96\frac{\hat{S}_n}{\sqrt{n}}\right] = [0.542, 0.598]$$

Case II: if we use conservative estimate:

$$\sigma \leq \frac{1}{2}, \text{upperbounded in Bernoulli}$$

$$\left[\widehat{\Theta}_n - 1.96\frac{1/2}{\sqrt{n}}, \widehat{\Theta}_n + 1.96\frac{1/2}{\sqrt{n}}\right]$$

This is a tad bit wider bound than the previous

As you can see in this case, they are not that much different, also as n get larger, they are essentially the same!

Some logistics about project presentation

- Each team, 10 minutes + 5 Q/A
- Need to cover:
  - Your data collection in detail (how much data, how do you collect, show evidence of your data collection)
  - What distribution do you use to model the event of interest and why?
  - Derive ML estimator (or other estimation if you use) for the parameters of the distribution
  - Evaluation: split your data into 80/20 (cross validation) and do your own testing
    - How accurate is your model?
    - Can it be better?
    - If it is not very good, what happened?
    - What is the take home message