Maximum Likelihood Estimation
(part 1)

November 21, 2016

Overview – What the heck are we doing?!

- ➔ High Level: what is statistical inference?
  - ◆ It's the process of making a statement about how data is generated in the world.
  - ◆ Think of the data that we observe in the world as a product of some data generation process (DGP) that is fundamentally unknown and likely highly convoluted.
- ➔ As data scientists, we try to use observed data to learn something about the DGP and in turn, represent the DGP with a statistical model.

Finding the statistical model....

First you have to select the probability distribution to represent the DGP

Selecting the probability distribution....

To select the correct probability distribution:

- ➔ Look at the variable in question.
- ➔ **Review the descriptions of the probability distributions.**
- ➔ Select the distribution that characterizes the variable.
- ➔ If historical data are available, use distribution fitting to select the distribution that best describes your data.

Descriptions of Probability Distributions...

→ **Go to this site:**

https://docs.oracle.com/cd/E12825_01/epm.111/cb_user/frameset.htm?apas04.html

Discrete	Continuous
• Binomial • Poisson • Geometric	• Beta • Gamma • Normal • Uniform

Finding the statistical model....continued

- Once we have selected a probability distribution to represent the data generation process, we aren't done:
  - Have not specified a unique distribution, but just a family of distributions.
  - Why just a family?
    - Because we leave one (or more) parameters as unknown.

- **The goal of statistical inference is then to use observed data to make a statement about parameters that govern our model.**

Let's throw in some (BASIC) notation....

- Let's call the model, or probability distribution, that we choose:

  $f(\cdot)$ $\Rightarrow$ this probability distribution is going to depend on a parameter θ (or vector of parameters, $\theta = \theta 1, \theta 2, ..., \theta k$) that characterize the distribution.

- The set Ω of all possible values of a parameter or the vector of parameters is called the parameter space.
- We then observe some data, drawn from this distribution:

$$X \sim f(x|\theta)$$

Let's throw in some (BASIC) notation....

$$X \sim f(x|\theta)$$

The random variables X1, ... ,Xn are independent and identically distributed because they are drawn independently from the same DGP.

➔ Remember, **the goal is to use the observed data x to learn about θ.**

- Knowing this parameter specifies a particular distribution from the family of distributions we have selected to represent the data generation process. In the end, we hope that θ is a substantively meaningful quantity that teaches us something about the world...

Where things start to diverge...

So far, we've just set up the general inferential goal. Now, we can introduce different theories of actually achieving said goal.

- ➔ We focus on two general approaches to inference and estimation:
  - frequentist / maximum likelihood **versus** Bayesian
    - The two are distinguished by their sources of variability, the mathematical objects involved, and estimation and inference.
    - It is important to keep track of the sources of randomness in each of these paradigms since different estimators are used for random variables as opposed to constants.

- First, let's restate the goal of inference:
  - it's to estimate the probability that the parameter governing our assumed distribution is θ conditional on the sample we observe, denoted as x.

We denote this probability as $\xi(\theta \mid x)$.

$$\underbrace{\xi(\theta|\mathbf{x})}_{\text{posterior}} \propto \underbrace{f_n(\mathbf{x}|\theta)}_{\text{likelihood}} \underbrace{\xi(\theta)}_{\text{prior}}$$

Maximum Likelihood vs. Bayesian Estimation

MLE

- The parameters in the frequentist setting (likelihood theory of inference) are unknown constants.
- Therefore, we can ignore $\xi(\theta)$ and just focus on the likelihood since everything we know about the parameter based on the data is summarized in the likelihood function.
- The likelihood function is a function of θ: it conveys the relative likelihood of drawing the sample observations you observe given some value of θ.

Bayesian

The parameters are latent random variables, which means that there is some variability attached to the parameters. This variability is captured through one's prior beliefs about the value of θ and is incorporated through the prior, $\xi(\theta)$.

The focus of Bayesian inference is estimating the posterior distribution of the parameter, $\xi(\theta|x)$.

The posterior distribution of θ, $\xi(\theta|x)$, is

Maximum Likelihood vs. Bayesian Estimation

MLE

- The parameters in the frequentist setting (likelihood theory of inference) are unknown constants.
- Therefore, we can

Bayesian

The parameters are latent random variables, which means that there is some variability attached to the parameters. This variability is captured through one's prior beliefs about the value of θ and is incorporated through the prior, $\xi(\theta)$.

The focus of Bayesian inference is estimating the posterior distribution of the parameter, $\xi(\theta|x)$.

The posterior distribution of θ, $\xi(\theta|x)$, is the distribution of the parameter conditional upon the observed data and provides some sense of (relative) uncertainty regarding our estimate for θ.

Maximum Likelihood vs. Bayesian Estimation

MLE

- The parameters in the frequentist setting (likelihood theory of inference) are unknown constants.
- Therefore, we can

Bayesian

The parameters are latent random variables, which means that there is some variability attached to the parameters. This variability is captured through one's prior beliefs about the value of θ and is incorporated through the prior, $\xi(\theta)$.

The focus of Bayesian inference is estimating the posterior distribution of the parameter, $\xi(\theta|x)$.

The posterior distribution of θ, $\xi(\theta|x)$, is the distribution of the parameter conditional upon the observed data and provides some sense of (relative) uncertainty regarding our estimate for θ.

Recap...

As a result of the differences in philosophies, the estimation procedure and the approach to inference differ between frequentists and Bayesians.

Specically, under the frequentist framework, we use the likelihood theory of inference where the maximum likelihood estimator (MLE) is the single point summary of the likelihood curve.

It is the point which maximizes the likelihood function.

In contrast, the Bayesian approach tends to focus on the posterior distribution of θ and various estimators, such as the posterior mean (PM) or maximum a posteriori estimator (MAP), which summarize the posterior distribution.

Introduction to Maximum Likelihood Estimation
PART 2

. . . Coming soon . . . maybe

- What is Likelihood and the MLE?
- Examples of Analytical MLE Derivations

Source:

http://www.konstantinkashin.com/notes/stat/Maximum_Likelihood_Estimation.pdf

2.2.3 Gamma Distribution

For the gamma distribution, θ is the scale parameter and α is the shape parameter. We seek the conditions for the maximum likelihood estimates of (θ, α).

The likelihood function for a gamma distribution is the following:

$$L(\alpha, \theta|\mathbf{x}) = f_n(\mathbf{x}|\alpha, \theta) = \frac{1}{\Gamma^n(\alpha) \cdot \theta^{n\alpha}} \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\alpha}\right)$$

Taking the log, we obtain:

$$\ell(\alpha, \theta) = -n \cdot \log(\Gamma(\alpha)) - n\alpha \cdot \log(\theta) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

Taking the derivative of the log likelihood with respect to θ and setting it equal to 0:

$$\frac{\partial \ell(\alpha, \theta)}{\partial \theta} = -\frac{n\alpha}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i = 0$$

Solving for θ, we obtain the following MLE for θ:

$$\hat{\theta}|\alpha = \frac{\sum_{i=1}^{n} x_i}{\alpha \cdot n} = \frac{1}{\alpha} \bar{x}_n$$

Plugging this back into the log likelihood function, taking its derivative with respect to α, and setting the result equal to 0:

$$\frac{dL(\alpha, \hat{\theta|\alpha})}{d\alpha} = -\frac{n \cdot \Gamma'(\alpha)}{\Gamma(\alpha)} - n \cdot \log\left(\frac{1}{\alpha} \bar{x}_n\right) + \sum_{i=1}^{n} \log(x_i) = 0$$

$$\frac{dL(\alpha, \hat{\theta|\alpha})}{d\alpha} = -\frac{n \cdot \Gamma'(\alpha)}{\Gamma(\alpha)} + n \cdot \log(\alpha) - n \cdot \log(\bar{x}_n) + \sum_{i=1}^{n} \log(x_i) = 0$$

Solving for α as far as we can (the answer remains in terms of the digamma function), we obtain the following condition for the MLE of α:

$$\log(\alpha) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \log(\bar{x}_n) - \frac{\sum_{i=1}^{n} \log(x_i)}{n}$$

Therefore, the MLE values of α and θ must satisfy the following 2 equations (there is no unique solution) are:

$$\log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = \log(\bar{x}_n) - \frac{\sum_{i=1}^{n} \log(x_i)}{n}$$

$$\hat{\theta} = \frac{1}{\hat{\alpha}} \bar{x}_n$$

Other sources

https://docs.oracle.com/cd/E12825_01/epm.111/cb_user/frameset.htm?apas03.html

https://docs.oracle.com/cd/E12825_01/epm.111/cb_user/frameset.htm?apas04.html