# Classical statistics VS Bayesian statistics

Ning Tian

September 4, 2017

*The main difference between the two statistics is that the former regards $\theta$ unknown, and the latter regards $\theta$ as a random variable having an unknown distribution.*

## 1 Maximum likelihood (ML)

Suppose $X = (X_1, \dots, X_n)$ is a random sample from a pdf $f_{\theta_0}$, where $\theta_0 \in \Delta$ is unknown. Suppose we observe $x = (x_1, \dots, x_n)$ and we want to estimate $\theta_0$.

Consider the likelihood function given by

$$L(\theta; x) = \mathbb{P}_\theta(X = x) \tag{1}$$

A maximum likelihood estimate for $\theta$ is the $\hat{\theta}$ that maximizes the likelihood function. For the continuous case,

$$L(\theta; x) = \prod_{i=1}^{n} f_\theta(x_i) \tag{2}$$

The mle is to obtain $\hat{\theta}$ by letting $l'(\theta; x) = 0$. If $l(\hat{\theta}; x)$ reaches its maximum, we say $\hat{\theta}$ is the maximum likelihood estimate.

## 2 Maximum a posteriori (MAP)

Let $\Theta$ be a random variable with pdf $r$. That is, $\Theta \sim r(\theta)$. Here, $r$ is called the prior pdf for $\Theta$; we do not really know the true pdf for $\Theta$, and this is a subjective assignment or guess based on our present knowledge or ignorance.

We think of $f(x_1; \theta) = f(x_1|\theta)$ as the conditional pdf of a random variable $X_1$ given $\Theta = \theta$. The joint pdf of $X_1$ and $\Theta$ is thus given by $f(x_1|\theta)r(\theta)$. In other words, we first generate $\Theta = \theta$ and then generate $X_1$ with pdf $f_\theta$. Similarly, $L(x; \theta) = L(x|\theta) = \prod_{i=1}^{n} f_\theta(x_i)$, and the joint pdf of $X$ and $\Theta$ is given by $j(x, \theta) = L(x|\theta)r(\theta)$. What we are interested in is to update our knowledge or belief about the distribution of $\Theta$ after the observation of $X = x$; more precisely, using Bayes' theorem we consider

$$s(\theta|x) = \frac{j(x,\theta)r(\theta)}{f_X(X)} = \frac{L(x,\theta)r(\theta)}{f_X(X)} = \frac{L(x,\theta)r(\theta)}{\int_{\theta\in\Delta} L(x,\theta)r(\theta)d\theta} \tag{3}$$

We call $s$ the posterior pdf. Thus 'prior' refers to our knowledge of the distribution of $\Theta$ prior to our observation of $X$ and 'posterior' refers to our knowledge after our observation of $X$.

The idea of MAP is to find $\theta$ which maximizes $s(\theta|x)$. Here, we only consider the maximization of $L(x,\theta)r(\theta)$ in that $\theta$ cannot change $\int_{\theta\in\Delta} L(x,\theta)r(\theta)d\theta$.

## 3 Example 1

Consider a case like $Y = H\Theta + V$. Here, $\Theta \sim N(\theta_0, P_0)$, $V \sim (0, R)$. Therefore, $Y|\Theta \sim N(H\theta, R)$, and we have

$$\begin{aligned} s(\theta|y) &\propto L(y|\theta)r(\theta) \\ &\propto \frac{1}{|R||P_0|} exp\left(-(y-H\theta)^\top R^{-1}(y-H\theta)\right) exp\left((\theta-\theta_0)^\top P_0^{-1}(\theta-\theta_0)\right) \end{aligned} \tag{4}$$

The MAP is to find $\hat{\theta}$ to minimize $(y-H\theta)^\top R^{-1}(y-H\theta) + (\theta-\theta_0)^\top P_0^{-1}(\theta-\theta_0)$.

## 4 Bayes' estimator

The function $\mathcal{L}(\theta,\theta')$ is called a loss function. Here, we consider a squared loss function $\mathcal{L}(\theta,\theta') = |\theta-\theta'|^2$. For such a squared loss function $\mathcal{L}(\theta,\theta')$, a Bayes' estimate of $\theta$ is a decision function $\delta(x)$ which minimizes $\mathbb{E}\left[\mathcal{L}\left(\Theta,\delta(x)\right)|X=x\right]$. Here, $\delta(x) = \mathbb{E}(\Theta|X=x)$. The Bayes' estimator is $\delta(X) = \mathbb{E}(\Theta|X)$.

Here is a brief proof about why $\delta(x) = \mathbb{E}(\Theta|X=x)$ is the minimizer in respect to the squared loss function.

*Proof.* Consider the scalar case. We assume the estimate of $\theta$ is $z$. We can thus write $\mathbb{E}\left[\mathcal{L}\left(\Theta,\delta(x)\right)|X=x\right]$ as $\mathbb{E}\left[(\Theta-z)^2|X=x\right]$.

$$\begin{aligned} \mathbb{E}\left[(\Theta-z)^2|X=x\right] &= \mathbb{E}\left[\Theta^2|X=x\right] - 2z\mathbb{E}\left(\Theta|X=x\right) + z^2 \\ &= \mathbb{E}\left[\Theta^2|X=x\right] - \mathbb{E}^2\left[\Theta|X=x\right] + \left(z-\mathbb{E}\left(\Theta|X=x\right)\right)^2 \end{aligned} \tag{5}$$

It can be easily seen that $\mathbb{E}\left(\Theta|X=x\right)$ is the minimizer. □

For different loss function, we can obtain different minimizers. In fact, the MAP minimizer represents the Bayes' estimator under a certain loss function corresponding to the minimizer as mode. However, for Gaussian distribution, the mode is the same with the mean. Therefore, the MAP estimate is the same with the Bayes' estimate.

# 5 Example 2

Consider the example as follows:

$$Y = X + W, \tag{6}$$

where $X \sim N(0, P_0)$, $W \sim N(0, Q)$, and $x$, $y$ are both scalars. We are going to compute the distribution of $X|Y$ from two aspects. One is from MAP, and the other from the joint Gaussian distribution. First, we consider the MAP method. From the distribution of $X$ and $W$, we have

$$Y|X = x \sim N(x, P_0) \tag{7}$$

According to the Bayes' rule, we have

$$\begin{aligned} p_{X|Y}(x|y) \propto p_{Y|X}(y|x)p_X(x) \propto exp\left((y-x)^2 P_0^{-1}\right) exp\left(x^2 P_0^{-1}\right) \\ \propto exp\left(\left(x - \frac{yP_0}{P_0+Q}\right)^2 \left(\frac{P_0 Q}{P_0+Q}\right)^{-1}\right) \end{aligned} \tag{8}$$

It can be observed that $X|Y = y \sim N\left(\frac{yP_0}{P_0+Q}, \frac{P_0 Q}{P_0+Q}\right)$. The MAP minimizer will be $\frac{yP_0}{P_0+Q}$, which is equal to the Bayes' estimate under the squared loss function.

Next, we consider the joint Gaussian distribution.

$$\left[\begin{array}{c} X \\ Y \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} P_0 & P_0 \\ P_0 & P_0+Q \end{array}\right]\right) \tag{9}$$

Therefore, $X|Y = y \sim N\left(\frac{yP_0}{P_0+Q}, \frac{P_0 Q}{P_0+Q}\right)$, which is the same with that obtained based on the MAP.

Furthermore, we can find that from

$$\left[\begin{array}{c} Y \\ X \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} P_0+Q & P_0 \\ P_0 & P_0 \end{array}\right]\right) \tag{10}$$

we have $Y|X = x \sim N\left(x, P_0\right)$, which is the same with our assumption.