Lecture Four: Bayesian Approach for Statistical Inference

June 29, 2011

Motoya Machida (mmachida@tntech.edu)
Conditional Probability

Let $A$ and $B$ be two events where $P(B) \neq 0$. Then the conditional probability of $A$ given $B$ can be defined as

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

The idea of “conditioning” is that “if we have known that $B$ has occurred, the sample space should have become $B$.” It is often the case that one can use $P(A|B)$ to find

$$P(A \cap B) = P(A|B)P(B).$$
Law of Total Probability

Let $A$ and $B$ be two events. Then we can write the probability $P(A)$ as

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

In general, suppose that we have a sequence $B_1, B_2, \ldots, B_n$ of mutually disjoint events satisfying $\bigcup_{i=1}^n B_i = \Omega$, where “mutual disjointness” means that $B_i \cap B_j = \emptyset$ for all $i \neq j$. (The events $B_1, B_2, \ldots, B_n$ are called “a partition of $\Omega$.”) Then for any event $A$ we have

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A|B_i)P(B_i).$$
Bayes Rule

Let $A$ and $B_1, \ldots, B_n$ be events such that the $B_i$'s are mutually disjoint, $\bigcup_{i=1}^{n} B_i = \Omega$ and $P(B_i) > 0$ for all $i$. Then

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

is called Bayes rule.
Joint Distribution

When a pair \((X, Y)\) of random variables is considered, a \textit{joint density function} \(f(x, y)\) is used to compute probabilities constructed from the random variables \(X\) and \(Y\) simultaneously. Given the joint density function \(f(x, y)\), the density function for each of \(X\) and \(Y\) is called the \textit{marginal density functions}, denoted by \(f_X(x)\) and \(f_Y(y)\). If \(X\) and \(Y\) are discrete,

\[
f_X(x) = \sum_y f(x, y) \quad \text{and} \quad f_Y(y) = \sum_x f(x, y).
\]

If \(X\) and \(Y\) are continuous,

\[
f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, dx.
\]
Conditional Probability Distributions

Suppose that two random variables $X$ and $Y$ has a joint density function $f(x, y)$. If $f_X(x) > 0$, then we can define the conditional density function $f_{Y|X}(y|x)$ given $X = x$ by

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}.$$

Similarly we can define the conditional density function $f_{X|Y}(x|y)$ given $Y = y$ by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

if $f_Y(y) > 0$. Then, clearly we have the following relation

$$f(x, y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y).$$
Bayes Rules for Probability Densities

Assume $f_X(x) > 0$. We obtain

$$f_{Y|X}(y|x) = \begin{cases} 
\frac{f_{X|Y}(x|y)f_Y(y)}{\sum_y f_{X|Y}(x|y)f_Y(y)} & \text{if } X \text{ and } Y \text{ are discrete;} \\
\frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y) \, dy} & \text{if } X \text{ and } Y \text{ are continuous.}
\end{cases}$$
Bayesian uses the concept of prior belief about the parameter $\theta$ of interest. Then the uncertainty of $\theta$ changes according to the data $D$. Here Bayesian interprets $\theta$ as a random variable, and the prior belief is given in the form of probability density $\pi(\theta)$ of $\theta$. In a Bayesian model we will investigates the posterior density $\pi(\theta | D)$ of $\theta$. 
Bayesian Method

Let $f(x; \theta)$ be a density function with parameter $\theta \in \Omega$. In a Bayesian model the parameter space $\Omega$ has a distribution $\pi(\theta)$, called a prior distribution. Furthermore, $f(x; \theta)$ is viewed as the conditional distribution of $X$ given $\theta$. By the Bayes’ rule the conditional density $\pi(\theta | x)$ can be derived from

$$
\pi(\theta | x) = \begin{cases} 
\frac{\pi(\theta) f(x; \theta)}{\sum_{\theta \in \Omega} \pi(\theta) f(x; \theta)} & \text{if } \Omega \text{ is discrete;} \\
\frac{\pi(\theta) f(x; \theta)}{\int_{\Omega} \pi(\theta) f(x; \theta) \, d\theta} & \text{if } \Omega \text{ is continuous.}
\end{cases}
$$
The distribution $\pi(\theta \mid x)$ is called the *posterior distribution*. Whether $\Omega$ is discrete or continuous, the posterior distribution $\pi(\theta \mid x)$ is “proportional” to $\pi(\theta)f(x; \theta)$ up to the constant. Thus, we write

$$
\pi(\theta \mid x) \propto \pi(\theta)f(x; \theta).
$$

It is often the case that both the prior density function $\pi(\theta)$ and the posterior density function $\pi(\theta \mid x)$ belong to the same family of density function $\pi(\theta; \eta)$ with parameter $\eta$. Then $\pi(\theta; \eta)$ is called *conjugate* to $f(x; \theta)$. 
A random sample

\[ X_1, \ldots, X_n \]

is regarded as independent and identically distributed (iid) random variables governed by an underlying probability density function \( f(x; \theta) \). A value \( \theta \) represents the characteristics of this underlying distribution, and is called a parameter. A point estimate is a “best guess” for the true value \( \theta \). Suppose that the underlying distribution is the normal distribution with \( (\mu, \sigma^2) \). Then the sample mean

\[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

is in some sense a best guess of the parameter \( \mu \).
Maximum Likelihood Estimate (MLE)

Having observed a random sample \((X_1, \ldots, X_n) = (x_1, \ldots, x_n)\) from an underlying pdf \(f(x; \theta)\), we can construct the likelihood function

\[
L(\theta, x) = \prod_{i=1}^{n} f(x_i; \theta),
\]

and consider it as a function of \(\theta\). Then the maximum likelihood estimate (MLE) \(\hat{\theta}\) is the value of \(\theta\) which “maximizes” the likelihood function \(L(\theta, x)\). It is usually easier to maximize the log likelihood

\[
\ln L(\theta, x) = \sum_{i=1}^{n} \ln f(x_i; \theta).
\]
Example: Bernoulli Trials

Let $X$ be a random variable taking value only on 0 and 1. An experiment observing such a variable is called *Bernoulli trial*. It is determined by the parameter $\theta$ (which represents the probability that $X = 1$). Suppose that $x = (x_1, \ldots, x_n)$ are observed from $n$ independent Bernoulli trials. Then the joint density function is given by

$$f(x; \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

By solving the equation

$$\frac{\partial \ln L(\theta, x)}{\partial \theta} = \left( \sum_{i=1}^n x_i \right) \frac{1}{\theta} - (n - \sum_{i=1}^n x_i) \frac{1}{1 - \theta} = 0,$$

we obtain $\theta^* = \frac{\sum_{i=1}^n x_i}{n}$. In fact, $\theta^*$ maximizes $\ln L(\theta, x)$, and therefore, it is the MLE of $\theta$. 
Beta Distributions

When $X_1, \ldots, X_n$ are independent uniform random variables on $[0, 1]$, the pdf of the $k$-th order statistic $X_{(k)}$ becomes

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k}, \quad 0 \leq x \leq 1,$$

which is known as the beta distribution with parameters $\alpha_1 = k$ and $\alpha_2 = n - k + 1$. In general, the density function

$$f(x; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1-1} (1-x)^{\alpha_2-1}.$$

is defined for $\alpha_1 > 0$ and $\alpha_2 > 0$. 
Example: Bernoulli Trials

Consider independent $n$ Bernoulli trials. Let

$$\pi(\theta; \eta_1, \eta_2) \propto \theta^{\eta_1} (1 - \theta)^{\eta_2}$$

be a prior density of beta distribution. Given the data $\mathbf{x} = (x_1, \ldots, x_n)$, the posterior density is calculated as

$$\pi(\theta | \mathbf{x}) = \pi(\theta; \eta_1 + \sum_{i=1}^{n} x_i, \eta_2 + n - \sum_{i=1}^{n} x_i)$$

The expected value of posterior density becomes

$$E(\theta | \mathbf{x}) = \frac{\eta_1 + \sum_{i=1}^{n} x_i + 1}{\eta_2 + n + 2}$$
Exponential Conjugate Family

Suppose that the pdf has the form

\[ f(x; \theta) = \exp \left[ n c_0(\theta) + \sum_{j=1}^{m} c_j(\theta) k_j(x) + h(x) \right], \]

and that a prior distribution is given by

\[ \pi(\theta; \eta_0, \eta_1, \ldots, \eta_m) \propto \exp \left[ c_0(\theta) \eta_0 + \sum_{j=1}^{m} c_j(\theta) \eta_j \right]. \]

Then we obtain the posterior density

\[ \pi(\theta | x) = \pi(\theta; \eta_0 + n, \eta_1 + k_1(x), \ldots, \eta_m + k_m(x)). \]

Thus, the family of \( \pi(\theta; \eta_0, \eta_1, \ldots, \eta_m) \) is conjugate to \( f(x; \theta) \), and the parameter \( (\eta_0, \eta_1, \ldots, \eta_m) \) of prior distribution is called the hyperparameter.
Dirichlet Distribution

Let $\alpha_1, \ldots, \alpha_k$ be parameters. Then we can define the pdf

$$
\pi(\theta; \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1}
$$

over the simplex

$$
Q = \{ \theta \in [0, 1]^k : \theta_1 + \cdots + \theta_k = 1 \}.
$$

This is called the Dirichlet distribution.
Suppose that \( \theta_j \) represents the probability for the \( j \)-th outcome for each \( j = 1, \ldots, m \), and \( n \) outcomes are repeatedly obtained according to the probability \( \theta_1, \ldots, \theta_k \). Then we can count for each \( j \) the number \( m_j \) of the \( j \)-th outcome, and the chance of getting the values \((m_1, \ldots, m_k)\) is given by the multinomial distribution

\[
f(m_1, \ldots, m_k; \theta) = \frac{n!}{m_1! \cdots m_k!} \prod_{j=1}^{k} \theta_j^{m_j}
\]

over

\[
\mathcal{Z}^\dagger = \{(m_1, \ldots, m_k) : m_1 + \cdots + m_k = n\}
\]
Let $\pi(\theta; \alpha_1, \ldots, \alpha_k)$ be the prior density for $\theta$ with hyperparameter $(\alpha_1, \ldots, \alpha_k)$. Given the data from the multinomial distribution with parameter $\theta$, we can obtain the posterior density

$$
\pi(\theta \mid m_1, \ldots, m_k) \propto \theta_1^{\alpha_1+m_1-1} \cdots \theta_1^{\alpha_k+m_k-1}
$$

which is the Dirichlet distribution $\pi(\theta; \alpha_1 + m_1, \ldots, \alpha_k + m_k)$. 