STA250 Lecture Notes

Jiaying Huang

October 10, 2013

Things we will cover today:
Maximum likelihood
Bootstrap
Monte Carlo
Markov Chains

Notice
The seminar this Thursday(Octorber 10th at 4:10pm) will talk about "Bayes at Scale" in Statistics department.

Basic Bayes Theorem Data $x \in X, x_j \in X_j$
Parameter $\theta intheta$
Joint pdf $P(x_1, x_2|\theta)$ "Marginal" pdf for $x_1 : p(x_1) = \int_{X_2} p(x_1, x_2|\theta)dx_2$
"Conditional" pdf for $x_1|x_2 : p(x_1|x_2, \theta) = \frac{p(x_1,x_2|\theta)}{p(x_2|\theta)} = \frac{p(x_1,x_2|\theta)}{\int p(x_1,x_2|\theta)dx_1}$

$$\Rightarrow p(x_1, x_2|\theta) = p(x_2|\theta)p(x_1|x_2, \theta) = p(x_1|\theta)p(x_2|x_1, \theta)$$

Extension for higher dimension: $P(x_1, x_2, ..., x_n|\theta)$
(1) Independent: $\prod p(x_i|\theta)$
(2) General: $\prod p(x_i|x_{[0:i-1]}, \theta) = p(x_1|\theta)p(x_2|x_1, \theta)...p(x_n|x_{n-1}, x_2, x_1, \theta) = p(x_1, x_2, ..., x_n|\theta)$
(3) Markov: $\prod p(x_i|x_{i-1}, \theta)$

Example
Suppose $Y_{ij}|\lambda_i \sim$ Poisson$(e_{ij}\lambda_i)$ and are independent for $i = 1, ..., k$ and $j = 1, ..., n$.
$\lambda_i \sim$ Gamma(α, β)

, λ_i are independent
Observations: $\{y_i\}$
Unknowns: $\{\lambda_i, \alpha, \beta\}$
Model: $\prod_{i=1}^{k} p(\lambda_i|\alpha,\beta) \prod_{j=1}^{n_i} p(y_{ij}|\lambda_i) = p(y,\lambda|\alpha,\beta), p(y|\alpha,\beta) = \int p(y,\lambda|\alpha,\beta)d\lambda$

Maximum Likelihood(MLE)
An estimate $\hat{\theta}$ is said to be the MLE of θ if
$\hat{\theta} = \arg\max_\theta p(y|\theta) = \arg\max_\theta p(\text{data}|\text{parameter})$,
i.e. value of the parameter that makes the observed data "most likely."
In practice we use $\hat{\theta_n} = \arg\max_\theta L_n(\theta)$, where $L_n(\theta) = log(y_1, \cdots, y_n|\theta)$.
In this class we will usually use log scale to do the work.

Properties of the MLE

1. Let $y_i \sim P(y|\theta_0)$, iid. Then $\hat{\theta_N} \rightarrow_P \theta_0$,
   That is to say, θ_0 is the true value of the parameter and the MLE converges to the true parameter as $n \rightarrow \infty$.

2. Also $\sqrt{n}(\hat{\theta_n} - \theta_0) \rightarrow_{distribution} N(0, I_1^{-1}(\theta))$, where $I_1(\theta) = E[-\frac{\delta^2}{\delta\theta^2} log(p(y|\theta))|\theta]$.

Example Suppose $y_1, \cdots, y_n \sim_{iid} N(\mu, \sigma^2)$.
$P(y_1, ...y_n|\mu,\sigma^2) = \prod_{i=1}^{n} p(y_i|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{s\sigma^2}}$ // Take log and take derivatives:

$$\hat{\mu} = 1/n \sum y_i = \bar{y}$$

$$\hat{\sigma^2} = 1/n \sum (y_i - \bar{y}) = \frac{n-1}{n} S^2, where S^2 is sample variance$$

Suppose my data generate from N(0,1) is $\mu_0 = 0, \sigma_0^2 = 1$, then as $n is close to infinite$, $\mu is close to 0 and \sigma^2 is close to 1$

Confidence Intervals
$C^{1-\alpha}(y)$ is a $100(1-\alpha)\%$ CI for θ if $P(\theta \in C^{1-\alpha}(y)) = 1-\alpha$, for all $\theta \in \Theta$.
i.e Under repeated sampling of datasets, $100(1-\alpha)\%$ of intervals will contain the true value of the parameter.
Example

Suppose $y_1, \cdots, y_n \sim_{iid} N(\mu, \sigma^2)$. To estimate μ we use $\hat{\mu} = \bar{y}$.
A $100(1-\alpha)\%$ CI for μ turns out to be

$$\bar{x} \pm t_{n-1,1-\alpha/2} * \sigma/\sqrt{(n)}$$

, where $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$ quartile of the t-distribution with $n-1$ degrees of freedom.

Model Misspecification

We use a density $p(y|\theta)$ to model our data, but what happens if the data comes from a different density, say, g? In other words, suppose your model is wrong and the data comes from a different density. And it happens all the time.
How will MLE perform in this case?

1. $\hat{\theta_n} \to \theta^*$, where θ^* generates the member of $p(y|\theta)$ that is "closest" to g.

2. $\sqrt{n}(\hat{\theta_n} - \theta^*) \to_d N(0, J_1^{-1}(\theta^*)V_1(\theta^*)J_1^{-1}(\theta^*))$, where $V_1(\theta) = Var[\frac{\delta}{\delta\theta} log(p(y|\theta)\|\theta]$ and $J_1(\theta) = E[-\frac{\delta^2}{\delta\theta^2} log(p(y|\theta)|\theta]$

If the model is true, then $V_1(\theta) = J_1(\theta)$ and we get the usual result.
If the model is wrong, we have extra J terms on either side, which leads to the so-called "sandwich estimate for the variance of $\hat{\theta}$".
IMPORTANT: Note the subscript 1 above. these values are based on per unit information.
$V_n(\theta) = var[\frac{\delta}{\delta\theta} log P(y_1, ..., y_n|\theta)|\theta]$, If not iid, we have to diverge on $V_n(\theta)$.

The Bootstrap

The bootstrap is a general method used to obtain standard errors for parameter estimates.
Let $Y_1, \cdots, Y_n \sim_{iid} F$, (pdf f, cdf F).
We want to estimate some population quantity $\theta = T(F)$ (for example if we are interested in the mean, T would be integration). We are going to use the plug-in estimate: $\hat{\theta_n} = T(\hat{F_N}) = t(X_n)$, where $\hat{F_n}$ is the empirical distribution (cdf) of the data $X_n = (Y_1, \cdots, Y_n)$, which places mass 1/n on each of the data points.

For example, we are interested in the population median $\theta = F^{-1}(0.5)$, then the plug-in estimate is the sample median $\hat{\theta_n} = \hat{F}_n^{-1}(0.5)$. Once we have an estimate $\hat{\theta}_n$, we want to estimate its distribution, or specifically its standard error.

Idea: Resample from the empirical distribution to approximate the distribution of $\hat{\theta}_n$ under the true model.

Algorithm

```
for(b in 1:B){
  #Resamle dataset with replacement size n
  bdata<-sample(data,replace=TRUE)
  #Compute estimate of $\theta$ for the bootstrap dataset
  est_vec[b] <- f(bdata)
```

NOTE: The size of the bootstrap data set is the same as the size of the original dataset

To estimate the standard error of $\hat{\theta}$, we use the standard deviation of the bootstrap estimates $\hat{\theta}_b^* : b = 1, \cdots, n$.