Sta 250 Lecture Notes

Dmitriy Izyumin

October 8, 2013

Announcement: The Statistics Seminar this week (Thursday at 4:10PM) is "Bayes at scale." We're about to start our Bayesian unit, so this is good timing.

The goal of this lecture is to cover the following four topics:

Maximum likelihood
Bootstrap
Monte Carlo
Markov Chains

Preliminaries

Data $x \in X, x_j \in X_j$
Parameter θ
"Marginal" pdf for $x_1 : p(x_1) = \int_{X_2} p(x_1, x_2|\theta)dx_2$
"Conditional" pdf for $x_1|x_2, \theta : p(x_1|x_2, \theta) = \frac{p(x_1,x_2|\theta)}{p(x_2|\theta)} = \frac{p(x_1,x_2|\theta)}{\int p(x_1,x_2|\theta)dx_1}$
So we have $p(x_1, x_2|\theta) = p(x_2|\theta)p(x_1|x_2, \theta) = p(x_1|\theta)p(x_2|x_1, \theta)$

In this class we will need to be familiar with marginalizing and conditioning, and to know when it is appropriate. In higher dimensions, we have:

Independent: $\prod p(x_i|\theta)$
General: $\prod p(x_i|x_{[0:i-1]}, \theta) = p(x_1|\theta)p(x_2|x_1, \theta)p(x_3|x_2, x_1, \theta) = p(x_1, x_2, x_3|\theta)$
Markov: $\prod p(x_i|x_{i-1}, \theta)$

Example

Suppose $Y_{ij}|\lambda_i \sim \text{Poi}(e_{ij}\lambda_i)$ and are independent for $i = 1, \cdots, k$ and $j = 1, \cdots, n$.
$\lambda_i \sim \text{Gamma}(\alpha, \beta)$
Observations: $\{y_i\}$
Unknowns: $\{\lambda_i, \alpha, \beta\}$
Model: $\prod_{i=1}^{k} p(\lambda_i|\alpha, \beta) \prod_{j=1}^{n} p(y_{ij}|\lambda_i) = p(y, \lambda|\alpha, \beta), p(y|\alpha, \beta) = \int p(y, \lambda|\alpha, \beta)d\lambda$

Maximum Likelihood

Bayesian statistics is based on maximum likelihood estimation, so it is important to understand the concepts involved here. An estimate $\hat{\theta}$ is said to be the MLE of θ if
$\hat{\theta} = \arg\max p(y|\theta) = \arg\max_\theta p(\text{data}|\text{parameter})$, 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 almost always work on the log scale in this class.

Useful Properties of the MLE

1. Let $y_i \sim P(y|\theta_0)$, independent. Then $\hat{\theta_N} \to_P \theta_0$, i.e. θ_0 is the true value of the parameter and the MLE converges to the true parameter as $n \to \infty$.

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

Confidence Intervals

We say $C^{1-\alpha}(y)$ 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 $y_1, \cdots, y_n \sim N(\mu, \sigma^2)$, independent. 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 $\alpha/2$ 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. How often does this happen? All the time.
What does the MLE converge to?

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

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

If the model is true, then $V_1(\theta) = J_1(\theta)$ and we get the usual result. When 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 a single observation.

The Bootstrap

The bootstrap is a very general method used to obtain standard errors for parameter estimates. Let $Y_1, \cdots, Y_n \sim F$, independent (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. If 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.

R Pseudocode

```
n<-250
mu<-c(-1,-0.5)

'bootstrap'<-function(data,f,B=200){
  #assumes scalar estimates
  #and that data can be sampled fom
  est_vec <- rep(NA,B)
  for(b in 1:B){
    #Resamle dataset:
    x_star <-sample(data,replace=TRUE)
    #Compute estimate
    est_vec[b] <- f(x_star)
  }
  #Return bootstrap distribution
  return(Est_vec)
}

#Perform bootstrap:
breps=bootstrap(data=x,f=mean,B=500)
#Plot:
hist(breps)
#estimate SE:
sd(breps)
```

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 $\hat{\theta}_b^* : b = 1, \cdots, n$.

Some sample Python code is available on the website. The code concerns functions used for random number sampling.