Class Notes: Parametric Estimation

2008

Parametric estimation methods assume that a sample is drawn from some known distribution, that is, $\chi^t \sim p\left(x\right)$ where $p\left(x\right)$ can be any distribution, for example Gaussian.

Parametric estimation assume a form for $p\left(x \mid \theta\right)$ and estimate θ using χ. When you have estimated the distribution parameters from the given sample, the whole distribution is known.

Maximum Likelihood Estimation

Maximum likelihood is the most commonly used method for parametric estimation.

Let us say we have an independent an identically distributed sample $\chi = \{x^t\}_{t=1}^{N}$. We assume x^t are drawn from some known probability density whit parameters θ.

$x^t \sim p\left(x \mid \theta\right)$

We want to find θ that makes sampling x^t from $p\left(x \mid \theta\right)$ as likely as possible. Given the Independence of the points the likelihood of sample χ is the product of the likelihoods of the individual points.

$$l\left(\theta\right) \equiv p\left(\chi \mid \theta\right) = \prod_{t=1}^{N} p\left(x^t \mid \theta\right)$$

In order to find θ the most likely to be drawn, we search for θ that maximizes the likelihood of the sample, denoted by $l\left(\theta \mid \chi\right)$.

It is commonly used the $\log$ of the likelihood as a trick to simplify computations without changing the value where likelihood takes its maximum.

$$\mathcal{L}\left(\theta \mid \chi\right) \equiv \log\, l\left(\theta \mid \chi\right) = \sum_{t=1}^{N} \log p\left(x^t \mid \theta\right)$$

Then, the estimated θ is expressed as

$$\theta^* = \arg\max_\theta\, \mathcal{L}\left(\theta \mid \chi\right)$$

Bernoulli Density

In a Bernoulli distribution an event occurs or it does not. The event occurs with probability p, and the nonoccurence of the event has probability $1-p$.

$$P\left(x\right) = p^x \left(1-p\right)^{1-x},\, x\,\epsilon\,\{0,1\}$$

p is the only parameter, so we want to calculate its estimator, $\hat{p}$. The log likelihood is

$$L\left(\theta \mid \chi\right) = \log \prod_{t=1}^{N} p^{x^t} \left(1-p\right)^{\left(1-x^t\right)} = \sum_t x^t \log\, p + \left(N - \sum_t x^t\right) \log\left(1-p\right)$$

Maximum Likelihood Estimator (MLE) is found by solving $\partial\mathcal{L}/dp$

$$\hat{p} = \frac{\sum_t x^t}{N}$$

Multinomial Density

Multinomial is the generalization of Bernoulli where instead of two states, the outcome of a random m event is one of K mutually exclusive states.

$p\left(x_1, x_2, \ldots, x_k\right) = \prod\limits_{i=1}^{K} p_i^{x_i}$, where $x_i$is the indicator variable of occurring state i

If we do N such independent experiments the MLE of p_i is

$\hat{p}_i = \frac{\sum_t x_i^t}{N}$

That is, the estimate for the probability of state i is the ratio of experiments with outcome of state i to the total number of experiments.

Gaussian (Normal) Distribution

Given a sample $\chi = \{x^t\}_{t=1}^N$ which Gaussian (normal) distributed with mean μ and variance σ^2 denoted by $\aleph\left(\mu, \sigma^2\right)$, its density function is

$p\left(x\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$

The $\log$ likelihood of a Gaussian sample is

$L\left(\theta \mid \chi\right) = -\frac{N}{2}\log\left(2\pi\right) - N\log\sigma - \frac{\sum_t \left(x^t-\mu\right)^2}{2\sigma^2}$

The MLE are

$\hat{\mu} = \frac{\sum_t x^t}{N}$

$\hat{\sigma}^2 = \frac{\sum_t \left(x^t-\hat{\mu}\right)^2}{N}$

The Bayes's Estimator

Sometimes, before looking at a sample, we may have some prior information on the possible value range for a parameter θ. This prior information is quite useful, especially when the sample is small. The prior information does not tell us exactly what the parameter values is, so we model this uncertainty by viewing θ as a random variable and by defining a prior density, $p\left(\theta\right)$.

The prior density, $p\left(x\right)$, tell us the likely values for θ before looking at the sample. Using Bayes's rule, after looking the sample we get the posterior density of θ by combining the prior density with the sample.

$p\left(\theta \mid \chi\right) = \frac{p(\chi|\theta)p(\theta)}{p(\chi)} = \frac{p(\chi|\theta)p(\theta)}{\int p(\chi|\theta')p(\theta')d\theta'}$

For estimating the density at x, we have

$$\begin{aligned} p\left(x \mid \chi\right) &= \int p\left(x, \theta \mid \chi\right) d\theta \\ &= \int p\left(x \mid \theta, \chi\right) p\left(\theta \mid \chi\right) d\theta \\ &= \int p\left(x \mid \theta\right) p\left(\theta \mid \chi\right) d\theta \end{aligned}$$

$p\left(x \mid \theta, \chi\right) = p\left(x \mid \theta\right)$ because knowing θ we know everything about the distribution. If we are doing a prediction in the form, $y = g\left(x \mid \theta\right)$, as in regression, then we have.

$y = \int g\left(x \mid \theta\right) p\left(\theta \mid \chi\right) d\theta$

Some times the posterior does not have a nice form and integration could not be feasible, then, using the maximum a posteriori (MAP) estimate will make the calculation easier, assuming that $p\left(\theta \mid \chi\right)$ has a narrow peak around this mode:

$\theta_{MAP} = \arg\max_{\theta} p\left(\chi \mid \theta\right)$

Then, we get ride of the integral using

$$\begin{aligned} p\left(x \mid \chi\right) &= p\left(x \mid \theta_{MAP}\right) \\ y_{MAP} &= g\left(x \mid \theta_{MAP}\right) \end{aligned}$$

If we have no prior reason to favor some values of θ, then the posterior will have the same form as the likelihood, $p\left(\chi \mid \theta\right)$, and the MAP estimate will be equivalent to the maximum likelihood estimate.

$\theta_{ML} = \arg\max_{\theta} p\left(\chi \mid \theta\right)$

Another possibility is the Bayes's estimator, defined as the expected value of the posterior density.

$\theta_{Bayes} = E\left[\theta \mid \chi\right] = \int \theta p\left(\theta \mid \chi\right) d\theta$

Suppose $x^t \sim \aleph\left(\theta, \sigma_0^2\right)$ and $\theta \sim \aleph\left(\theta, \sigma\right)$ where μ, σ, σ_0^2 are known

$p\left(\chi \mid \theta\right) = \frac{1}{\left(2\pi\right)^{N/2}} \exp\left[-\frac{\sum_t \left(x^t - \theta\right)^2}{2\sigma_0^2}\right]$

$p\left(\theta\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(\theta-\mu\right)^2}{2\sigma^2}\right]$

It can be shown that $p\left(\theta \mid \chi\right)$ is normal with

$E\left[\theta \mid \chi\right] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$