Maximum likelihood estimation and Maximum a posteriori estimation

Pattern Recognition and Machine Learning, Jul-Nov 2019

Indian Institute of Technology Madras

August 28, 2019

Maximum likelihood estimation

For univariate case, the probability density function is given by,

$$P(x|W_i) \sim \mathcal{N}(x|\bar{\theta}_i)$$
$$\text{Parameter vector, } \bar{\theta}_i = \{\mu_i, \sigma_i^2\}$$
$$\text{Data points, } \mathcal{D} = \{x_{1i}, x_{2i}, \cdots, x_{Ni}\}$$
$$\text{Parameter vector to be estimated, } \hat{\bar{\theta}}_i = \{\hat{\mu}_i, \hat{\sigma_i^2}\}$$

Assuming conditional independence of the data points
Likelihood function,

$$\mathcal{L}(\bar{\theta}) = \prod_{k=1}^{N} P(x_k|\bar{\theta}) \tag{1}$$

Maximizing the likelihood is same as maximizing the log likelihood.
Log likelihood,

$$\ln(\mathcal{L}(\bar{\theta})) = \sum_{k=1}^{N} \ln(P(x_k|\bar{\theta}))$$

$$\bar{\nabla}_\theta = -\frac{d}{d\bar{\theta}} \ln(\mathcal{L}(\bar{\theta})) = \bar{0}$$

$$\implies \begin{bmatrix} \frac{\partial}{\partial\theta_1} \ln(\mathcal{L}(\bar{\theta})) \\ \frac{\partial}{\partial\theta_2} \ln(\mathcal{L}(\bar{\theta})) \end{bmatrix} = \begin{bmatrix} \bar{0} \\ \bar{0} \end{bmatrix}$$

$$\hat{\theta}_1 = \frac{1}{N} \sum_{k=1}^{N} x_k = \mu$$

$$\hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^{N} (x_k - \hat{\theta}_1)^2 = \sigma^2$$

Aside:

$$\frac{d}{dM}\ln|M| = M^{-1}$$

$$\frac{d}{dM}\bar{x}^t M^{-1}\bar{x} = -M^{-1}\bar{x}\bar{x}^t M^{-1}$$

Sample covariance matrix estimation:

$$\ln p(x_n|\bar{\theta}) = \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|C| - \frac{1}{2}(\bar{x}_n - \bar{\mu})^t C^{-1}(\bar{x}_n - \bar{\mu})$$

$$\Delta_c = \sum_{k=1}^{N} C^{-1} = \sum_{k=1}^{N} C^{-1}(\bar{x}_k - \bar{\mu})(\bar{x}_k - \bar{\mu})C^{-1}$$

$$C = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{\mu}_1)(x_k - \hat{\mu}_1)^t$$

Rule of thumb: At least 30 examples are required for the estimation of parameters even if $\bar{x}_i$'s are independently and identically distributed.

- No. of parameters in the estimation of mean$= d$
- No. of parameters in the estimation of covariance matrix$=$ $\frac{d(d+1)}{2} + d$.
- Total no. of examples required $= 30 * [\frac{d(d+1)}{2} + d]$

Note: Estimation of C becomes poor as the dimension of the feature vevector increases. If the dimension is large and the data available for training is less, the matrix may become ill-conditioned.

Revisiting linear regression parameter estimation using MLE I

$$t = y(\bar{x}, \bar{w}) + \epsilon \quad (2)$$

$\epsilon \sim \mathcal{N}(0; \beta^{-1})$ β^{-1} is referred to as precision, β is the variance.
Clearly as variance increases, precision of estimates decreases.
Objective: Estimation of $\bar{w}$ under the assumption that
$p(t|\bar{x}, \bar{w}, \beta) = \mathcal{N}(t|y(\bar{x}, \bar{w}), \beta^{-1}$
What does this mean? It means that the true values are distributed around the estimated value (mean), and are Gaussian distributed.
$E[t|\bar{x}] = \int tp(t|\bar{x})dt = y(\bar{x}, \bar{w})$.
where $y(\bar{x}, \bar{w})$ is the conditional mean of hte target variables.
Let $\mathcal{X} = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_N\}$, a set of data points.
The likelihood functions given by:
$\ln \mathcal{L}(\bar{w}, \beta) = p(t|\mathcal{X}, \beta) = \ln \prod_{n=1}^{N} \mathcal{N}(t_n|y(\bar{x}, \bar{w}), \beta^{-1})$
$\ln \mathcal{L}(\bar{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \beta E_D(\bar{w})$
$E_D(\bar{w}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - \bar{w}^t \bar{\Phi}(\bar{x}))^2$

Revisiting linear regression parameter estimation using MLE II

$\mathcal{L}(\bar{w}, \beta) = \beta \sum_{n=1}^{N}(t_n - \bar{w}^t\bar{\Phi}(\bar{x}))^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi$

$\nabla \ln \mathcal{L}(\bar{w}, \beta) = \sum_{n=1}^{N}(t_n - \bar{w}^t\bar{\Phi}(\bar{x}))\bar{\Phi}(\bar{x})^t = 0$

$\hat{w}_{Ml} = [(\bar{\Phi}^t\bar{\Phi})^{-1}\bar{\Phi}^t]\bar{T}$

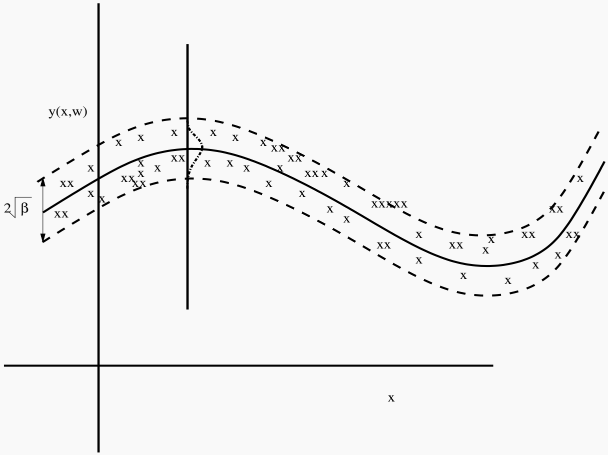
where $\bar{T}$ is a vector $[t_1, t_2, ..., t_N]^t$

Estimation of β yields

$\beta_{ML} = \frac{1}{N}\sum_{n=1}^{N}(t_n - y(\bar{x}_n, \bar{w}))^2$

The values of t_n hover around the expected value with a variance of β_{ML}

Illustration of Parameter Estimation using MLE



Maximum a posteriori (MAP) parameter estimation

MAP estimation of the parameter θ is given as:

$$\bar{\theta}_{MAP} = \mathcal{L}(\bar{\theta})P(\bar{\theta}) \tag{3}$$

Note: $\bar{\theta}_{MLE} = \bar{\theta}_{MAP}$ if $p(\bar{\theta})$ is a uniform prior.

Let $P(D|\mu) = \prod_{k=1}^{N} p(x_k|\mu)$; where $p(x_k|\mu)\backslash N(x_k; \mu, \sigma)$

$$p(\mu|D) = \frac{P(D|\mu)P(\mu)}{P(D)}$$

$$\hat{\theta}_{MAP} = \arg\max_{x} \frac{\ln P(D|\bar{\theta})P(\theta)}{P(D)}$$

$$l(\mu) = ln(\prod_{k=1}^{N} p(x_k|\mu)p(\mu))$$

$$l(\mu) == \frac{1}{2}\sum_{k=1}^{N}\frac{(x_k-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}$$

$$\frac{\partial l(\mu)}{\partial \mu} = 0$$

$$\hat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}x_k + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \quad (4)$$

If $\sigma_0^2 >> \sigma^2$, or if N is more, $\hat{\mu}_{MAP} = \frac{1}{N}\sum_{k=1}^{k} x_k$. That is, $\bar{\theta}_{MLE} = \bar{\theta}_{MAP}$

Multivariate Gaussian distributions

Case 1:

$$P(x|\mu)\backslash N(\bar{x};\bar{\mu},\sigma^2 I)p(\bar{\mu}) = N(\mu_i\mu_0,\sigma^2 I)$$

$$\hat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\mu_0 + \frac{N\hat{\mu}_{MLE}}{\sigma^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} \tag{5}$$

If $\sigma_0^2 >> \sigma^2$, or if N is more, $\hat{\mu}_{MAP} = \frac{1}{N}\sum_{k=1}^{k} x_k$. That is, $\bar{\theta}_{MLE} = \bar{\theta}_{MAP}$

Case 2: $\bar{\mu}$, C is a general matrix

$$\hat{\mu}_{MAP} = (NC^{-1} + C_0^{-1})^{-1}(C_0^{-1}\bar{\mu}_0 + NC^{-1}\bar{\mu}_{ML}) \tag{6}$$

Maximum entropy distribution

- Try to maximize entropy

Cost Function:

$$\sum_{k=1}^{N} p(x) \ln p(x) + \lambda\left(\sum_{k=1}^{N} p(x) - 1\right) \tag{7}$$

Assume $p(x)$ is a maximum entropy distribution.

H.W Work on maximum entropy estimation for discrete distribution.

Kullback–Leibler divergence (KL-divergence)

- It is computed as:

$$KL - divergence = \int p(x) \ln \frac{p(x)}{q(x)} dx \tag{8}$$

- KL-divergence is used for comparing different density functions
- The larger the KL-divergence, the more different are the density functions.
- KL-divergence is not symmetric. In practice, we take KL-divergence in both directions and take the average, which is known as **symmetric KL-divergence**.