intro math

Shumin An

March 2019

1 Introduction

We use the X below represents data samples.

$$X_{N\times p} = (x_1, x_2, \cdots, x_N)^T, x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$$

As we know, there are two point of views: **Frequentists** and **Bayesians**

For Frequentists, they assumed θ in $p(x|\theta)$ is a constant value. The probability of the data samples is $p(X|\theta) \underset{iid}{=} \prod_{i=1}^{N} p(x_i|\theta)$. Thus, the θ should give the maximum probability. In order to get that θ, we can compute **MLE**:

$$\theta_{MLE} = \underset{\theta}{argmax} \log p(X|\theta) \underset{iid}{=} \underset{\theta}{argmax} \sum_{i=1}^{N} \log p(x_i|\theta) \tag{1}$$

For bayesians, they assumed θ subject to a prior distribution. According to the Bayesian Theorem, the posterior probability could be written as:

$$p(\theta|X) = \frac{p(X|\theta)\cdot p(\theta)}{p(X)} = \frac{p(X|\theta)\cdot p(\theta)}{\int\limits_{\theta} p(X|\theta)\cdot p(\theta) d\theta}$$

In order to find θ, we should maximize the posterior. However, there's no need to actually get the value of posterior probability, since p(X) is unrelated with θ. So here come to the **MAP**:

$$\theta_{MAP} = \underset{\theta}{argmax}\, p(\theta|X) = \underset{\theta}{argmax}\, p(X|\theta)\cdot p(\theta) \tag{2}$$

2 Basic Math

2.1 Gaussian Distribution

2.1.1 MLE in One Dimension

Basically, the probability density function of Gaussian Distribution could be written as:

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{3}$$

Next, we compute its MLE only consider in one dimension situation:

$$\log p(X|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x_i-\mu)^2/2\sigma^2)$$

We could see that the parameter has two parts: mean and variance, so we need to compute them respectively. For μ we have:

$$\mu_{MLE} = \underset{\mu}{argmax} \log p(X|\theta) = \underset{\mu}{argmax} \sum_{i=1}^{N} (x_i-\mu)^2$$

Let its partial derivatives equals 0, we could get the value of μ:

$$\frac{\partial}{\partial\mu} \sum_{i=1}^{N} (x_i-\mu)^2 = 0 \longrightarrow \mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{4}$$

And same process for σ:

$$\sigma_{MLE} = \underset{\sigma}{argmax} \log p(X|\theta) = \underset{\sigma}{argmax} \sum_{i=1}^{N} [-\log\sigma - \frac{1}{2\sigma^2}(x_i-\mu)^2] = \underset{\sigma}{argmin} \sum_{i=1}^{N} [\log\sigma + \frac{1}{2\sigma^2}(x_i-\mu)^2]$$

$$\frac{\partial}{\partial\sigma} \sum_{i=1}^{N} [\log\sigma + \frac{1}{2\sigma^2}(x_i-\mu)^2] = 0 \longrightarrow \sigma_{MLE}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i-\mu)^2 \tag{5}$$

The μ and σ above are estimated value based on input data. We could compute whether there is a bias on them. For μ:

$$\mathbb{E}_{\mathcal{D}}[\mu_{MLE}] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N} \sum_{i=1}^{N} x_i] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathcal{D}}[x_i] = \mu$$

For σ:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\sigma^2_{MLE}] &= \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}(x_i^2-2x_i\mu_{MLE}+\mu^2_{MLE})\\
&= \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2_{MLE}] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2+\mu^2-\mu^2_{MLE}]\\
&= \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2]-\mathbb{E}_{\mathcal{D}}[\mu^2_{MLE}-\mu^2] = \sigma^2-(\mathbb{E}_{\mathcal{D}}[\mu^2_{MLE}]-\mu^2)\\
&= \sigma^2-(\mathbb{E}_{\mathcal{D}}[\mu^2_{MLE}]-\mathbb{E}^2_{\mathcal{D}}[\mu_{MLE}]) = \sigma^2-Var[\mu_{MLE}]\\
&= \sigma^2-Var[\frac{1}{N}\sum_{i=1}^{N}x_i] = \sigma^2-\frac{1}{N^2}\sum_{i=1}^{N}Var[x_i] = \frac{N-1}{N}\sigma^2
\end{aligned}
$$

We could see that σ has a bias comes from using μ_{MLE}

2.1.2 MLE in Multi-Dimension

Multi-Dimension Gaussian Distribution could be written as:

$$
p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{6}
$$