Week 1 Report

Jiayi Weng
wengjy16@mails.tsinghua.edu.cn

Abstract

Notes for PRML chap. 1-2. MLAPP chap. 1-2 is similar to the chap. in PRML.

1 Likelihood

Likelihood function $p(\mathcal{D}|w)$ is the probability estimation of model's parameter w, given the data $\mathcal{D}$.

Maximum Likelihood Estimation (MLE) means to maximize the likelihood function $p(\mathcal{D}|w)$. It needs to adjust the parameters w to get the model's best probability estimation using data $\mathcal{D}$.

Here is a brief example: we first denote the probability of one coin head up ("H") as p_H. If we flip this coin twice and observe HH, then the likelihood cuntion can be written as

$$p(\mathrm{HH}|p_H = \theta) = \theta^2$$

the likelihood function would get its global maximum point when parameter $\theta = 1$. So we could say $p_H = 1$ is a good estimation according to our observation.

Similarly, the observation "HTH" corresponds to the likelihood function $p(\mathrm{HTH}|p_H = \theta) = \theta^2(1-\theta)$, and will get its maximum when $\theta = \dfrac{2}{3}$.

The negative log of likelihood function is called *error function*. MLE is equivalent to minimize the error function, owing to the fact that the negative logarithm is a monotonically decreasing function.

2 Bayesian Approach

The Bayes' theorem is listed below:

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})} \tag{1}$$

notice that $p(\mathcal{D})$ is a constant when $\mathcal{D}$ is determined, we could rewrite it as

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{2}$$

Different from MLE (only maximize the likelihood), the bayesian approach tries to *maximum a posterior estimation*(MAP), that is the product of likelihood and prior. The uncertainty of w is present as prior probability. Thus MAP is more robust.

MLE gives the point estimation, and MAP gives the probability distribution estimation.

3 Regularization

Assume the parameter w in the model and sampled data from $\mathcal{D}$ are i.i.d..

If we assume w is drawn from uniform distribution, MAP is equal to no regularization. (The prior is uniform distribution, so posterior = likelihood. In other words, MLE $\subset$ MAP.)

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

If we assume w is drawn from laplace distribution, MAP is equal to L1 regularization.

If we assume w is drawn from gaussian distribution, MAP is equal to L2 regularization.

3.1 MLE

Given the data x and label t from a regression problem, if the label obeys gaussian distribution, more formally,

$$p(t|x, \mathrm{w}, \beta) = \mathcal{N}(t|y(\mathrm{x}, \mathrm{w}), \beta^{-1}) \tag{3}$$

Here, $\beta = \sigma^2$ is the precision parameter, and $\mathcal{N}(x|\mu, \sigma^2)$ defines the gaussian distribution by

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{4}$$

If we have the sampled data $\mathcal{D} = \{\mathbf{x}, \mathbf{t}\}$, using MLE to determine w and β, the likelihood function is

$$p(\mathbf{t}|\mathbf{x}, \mathrm{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathrm{w}), \beta^{-1}) \tag{5}$$

and the error function is obtained using Equation 4 and 5 :

$$\begin{aligned} -\log(p(\mathbf{t}|\mathbf{x}, \mathrm{w}, \beta)) &= -\sum_{n=1}^{N} \log(\mathcal{N}(t_n|y(x_n, \mathrm{w}), \beta^{-1})) \\ &= \sum_{n=1}^{N} \frac{1}{2}\log(2\pi\beta^{-1}) + \frac{(t_n - y(x_n, \mathrm{w}))^2}{2\beta^{-1}} \\ &= \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y(x_n, \mathrm{w}))^2 - \frac{N}{2}\log\frac{\beta}{2\pi} \end{aligned} \tag{6}$$

From this result, if hyperparameter β is given, the last term could consider as constant. Also, β does not influence the global minimum location in this function, and the best estimation of β could be obtained from the best w. Consequently, MLE equivalents to minimize the sum of square error under the assumption of gaussian distribution.

3.2 MAP

Once we use Bayesian approach to find out the best w, the equations are

$$p(\mathrm{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) = p(\mathbf{t}|\mathbf{x}, \mathrm{w}, \beta)p(\mathrm{w}|\alpha) \tag{7}$$

where hyperparameter α determines the precision of the distribution w, for simplicity, a gaussian distribution:

$$p(\mathrm{w}|\alpha) = \mathcal{N}(\mathrm{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathrm{w}^T\mathrm{w}\right\} \tag{8}$$

then the negative log of MAP is

$$\begin{aligned} & -\log\left\{p(\mathbf{t}|\mathbf{x}, \mathrm{w}, \beta)p(\mathrm{w}|\alpha)\right\} \\ = \;& \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y(x_n, \mathrm{w}))^2 - \frac{N}{2}\log\frac{\beta}{2\pi} + \frac{\alpha}{2}\mathrm{w}^T\mathrm{w} - \frac{M+1}{2}\log\frac{\alpha}{2\pi} \end{aligned} \tag{9}$$

In this equation, the 2nd and 4th term are not relevant to the global minimum, which can be omitted. Comparing with Equation 6 and 9, the 3rd term $\frac{\alpha}{2}\mathrm{w}^T\mathrm{w}$ is called regularization term. Thus maximize the posterior probability is equivalent to minimizing the sum of square error with a regularization parameter $\lambda = \frac{\alpha}{\beta}$.

4 Others

4.1 Generative Model and Discriminative Model

GM models the joint distribution of $p(\mathrm{x}, \mathcal{C}_k)$ directly to obtain the posterior probability $p(\mathcal{C}_k|\mathrm{x})$. It needs tons of data.

DM models the posterior probability $p(\mathcal{C}_k|\mathrm{x})$ directly. It needs fewer data.

4.2 KL Divergence

Given two distributions $p(x)$ and $q(x)$, the *KL divergence*, also called *relative entropy* or *infomation gain*, is defined as

$$\mathrm{KL}(p||q) = -\int p(\mathrm{x}) \ln \frac{q(\mathrm{x})}{p(\mathrm{x})} \mathrm{dx} \tag{10}$$

It is always non-negative with equality iff. $p(\mathrm{x}) = q(\mathrm{x})$. $\mathrm{KL}(p||q)$ can be interpreted as measuring the expected number of extra bits required to code samples from p using a code optimized for q rather than the code optimized for p.

4.3 Student's t-distribution

Student's t-distribution is the sum of infinity gaussian distribution. It is more robust than gaussian distribution.

References

[1] Nasser M Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007.