Notes for PRML

Hui Wei

November 11, 2019

1 Chapter 1 Introduction

Note 1 *Why we need to maximize the logarithm of the likelihood function, rather than maximizing it directly? For example, we sample N iid variables which is Gaussian. So the likelihood function is*

$$p(\mathbf{x}|\mu,\sigma^2)=\prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2)$$

It is because not only taking the logarithm simplifies the mathematical analysis, but it will also help numerically since the product of a large number of small probabilities can easily underflow the numerical precision of the computer, and it will be resolved by the sum of the log probability.

$$\ln p(\mathbf{x}|\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi)$$

Note 2 *In the curve fitting example, after we know about the data set* $\mathbf{x}=(x_1,\ldots,x_N)^T$*, we can use* $y(x,\mathbf{w})=\sum_{j=0}^{M}w_jx^j$ *to compute the mean of the target* $\mathbf{t}=(t_1,\ldots,t_N)^T$*. However, we need to consider the actual noise, so we need to express our uncertainty over the value of the target variable using a probability distribution.*

Note 3 *How to understand every item in the Bayesian Theorem?*

$$p(\mathcal{C}_k|\mathbf{x})=\frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

$p(\mathcal{C}_k|\mathbf{x})$ *is posterior probability, which means after the data is observed, we can get the probability of every class for each data point.* $p(\mathbf{x}|\mathcal{C}_k)$ *is class-condition density, which means in a specific class, what the feature* $\mathbf{x}$ *is like.* $p(\mathcal{C}_k)$ *is prior probability, which means before we can observe data, the probability of every class, we can get it from the statistics. For* $p(\mathbf{x})$*, we can use* $\sum_k p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$*.*

Note 4 *The understanding of the regression model. See Pattern Recognition and Machine Learning Page 29 and 47.*

2 Chapter 2 Probability Distributions

Note 5 *Why maximizing the likelihood function can lead to the over-fitting?*

Let's take Bernoulli distribution as an example.

$$Bern(x|\mu)=\mu^x(1-\mu)^{1-x}$$

where $x\in\{0,1\}$*.Note we write the* ***hyperparameter*** μ *at the position of condition.*

For a data set $\mathcal{D} = \{x_1, \ldots, x_n\}$, we assume they are extracted independently and identically from $p(x|\mu)$, so that the logarithm of the likelihood is

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n \ln(1-\mu))\}.$$

We can get the derivative of it w.r.t μ and get

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Assume m is the number of $x_n = 1$, then

$$\mu = \frac{m}{N}$$

Suppose after observing the event three times, it happen to have $x_n = 1$ three times. So for the maximum likelihood, $N = m = 3$ and $\mu = 1$, and it is an unreasonable result, which is over-fitting to the small dataset.

Note 6 *How to tackle this kind of problem?*

*In PRML, we can first assume a prior distribution $p(\mu)$, such like Beta distribution. Then we use **sequential** approach to learning when we adopt a Bayesian viewpoint.Note that this approach is independent of the choice of prior and of the likelihood function and depends only on the assumption of i.i.d data.(PRML) Note that usually, we use the conjugate function of the likelihood function as the prior distribution, since in this way, the posterior distribution has the same form as the likelihood function and prior distribution.*

Note 7 *Note that*

$$E[\mathbf{x}|\mu] = \sum_{\mathbf{x}} \mathbf{p}(\mathbf{x}|\mu)\mathbf{x}$$

for some vector **x.**

Note 8 *The **limitations** of Gaussian distribution:*

*1.There are totally $\frac{D(D+3)}{2}$ independent parameters in Σ and μ. **Solution**: use restricted forms of the covariance matrix.*

*2.It is intrinsically unimodal(i.e. has a single maximum) and so is unable to provide a good approximation to multimodal distributions. **Solution**:use the latent variables to solve this two problems.*

Note 9 *Assume we have the dataset $\mathbf{x} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$. The unbiased estimator for the expectation is the sample average $\mathbf{E}[\mathbf{x}] = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{x_n}$. The unbiased estimator for the variance is $cov[\mathbf{x}] = \frac{\mathbf{1}}{\mathbf{N-1}} \sum_{\mathbf{n=1}}^{\mathbf{N}} (\mathbf{x_n} - \bar{\mathbf{x}})^{\mathbf{2}}$, which is not the sample variance. For why we use $N - 1$, please refer to wikipedia about Bessel correction.*

Note 10 *When we use Bayesian inference, we need to choose the prior distribution for some parameters. A convenient way is to choose the* ***conjugate*** *distribution of the maximum likelihood function, so that the posterior distribution has the same formula as the prior distribution and maximum likelihood function.*

Note 11 *Gaussian distribution is sensitive for some outliers of data (PRML), while t-distribution is more robust than Gaussian distribution. This is the reason why the linear regression is not robust, since we assume that the maximum likelihood function is Gaussian. We can replace Gaussian distribution with t-distribution in linear regression to solve this problem.*

Note 12 *For the periodic variable, we need to map the distribution in the real-axis into the polar-coordination system. Note that for the periodic variable, the distribution of it should be periodic. There is a very important method for indicating the periodic variable. Since the statistic information of it heavily dependent on the choice of the origin, we use the two-dimensional vector to represent every observation data point.* $\boldsymbol{x}_n = (cos\theta_n + sin\theta_n)$*, here we only care about* θ*, and assume the radius* $r = 1$*. The reason is that now we are concerned about the periodic variables, and we use radian to represent them instead of their length. However, for computing the mean and the variance of the periodic variable, we need to consider the length, that is* $\bar{\boldsymbol{x}}_n = (\bar{r}cos\bar{\theta}, \bar{r}sin\bar{\theta})$*.*

Note 13 *There is another shortcoming of Gaussian distribution. When there is more than one peak, more than one mountain, the original Gaussian distribution cannot represent such a distribution. Solution: use the Mixture of Gaussian model to describe such a distribution. The reason we can do this is the linear combination of Gaussian models can almost depict any continuous distribution. Of course, we can use the mixture of different models besides Gaussian model.*

Note 14 *Mixtures of Gaussians:*

$$p(\mathbf{x}) = \sum_{\mathbf{k=1}}^{\mathbf{K}} \pi_{\mathbf{k}} \mathcal{N}(\mathbf{x}|\mu_{\mathbf{k}}, \mathbf{\Sigma_k})$$

A good interpretation of the Mixture of Gaussian is: we can consider pi_k *as the prior distribution, telling us how likely we pick k-th component of the entire mixture model.* $\mathcal{N}(\mathbf{x}|\mu_{\mathbf{k}})$ *is the density of k-th component. The linear combination of all components is the entire mixture model.*

Note 15 *The difference between the parametric and the nonparametric methods for estimating parameters in the model is: for parametric method, we need to assume a specific functional forms governed by a small number of parameters whose value are to be determined from a data set, while for nonparametric one, we make few assumptions about the form of the distribution and we let the data "speak for itself". However, there still exists a shortcoming for nonparametric method such like K-Nearest-Neighbor and Kernel density estimator, that is it needs to store the whole dataset. So if the dataset is very large, it is computationally expensive.*

3 Chapter 3 Linear Models for Regression

Note 16 *The linear model is linear for the parameter* $\boldsymbol{w}$*. It is also the linear combination of the basis functions. One of the limitation of polynomial basis function is that they are global functions of the input variable, so that changes in one region of input space affect all other regions.*

Note 17 *For the maximum likelihood for the linear regression, there are two important assumptions: 1. all the data are extracted identically and independently from the same distribution. 2. the target variable is given by a deterministic function with additive Gaussian noise.*

Note 18 *Introduce a very important concept of linear algebra: Moor-Penrose pseudo-inverse.*

$$\mathbf{\Phi}^{\dagger} \equiv (\mathbf{\Phi^T \Phi})^{-1}\mathbf{\Phi^T}$$

Why we need that? Since it is the generalization form for the matrix inverse for non-square matrices. If we solve the linear regression problem to get the parameter $\mathbf{w}$*, we can get:*

$$\mathbf{w_{ML}} = (\mathbf{\Phi^T \Phi})^{-1}\mathbf{\Phi^T t}$$

which is known as the normal equations. $\mathbf{\Phi}$ *is the design matrix and the elements in it is value of the basis function for every data point.* $\mathbf{t}$ *is the target value vector.*

Note 19 *Two important conclusion for maximum likelihood for linear regression: 1. The bias* w_0 *compensates for for the difference between the average (over the training set) of the target values and the weighted sum of the averages of the basis function values. The reason we say "compensates" since in the expression of the linear regression* w_0 *is additive. 2. The inverse of the noise precision* β *is given by the residual variance of the target values around the regression function.*

Note 20 *For the number of output (target value)* $K > 1$*, the common way is to use the same set of basis functions* $\phi(\mathbf{x})$*, but different coefficient set* $\mathbf{w}$*.*

Note 21 *How to use the Bayesian method to get the linear regression function?(general method)*

1. like before, we need to get the likelihood function from the i.i.d. data.

2. according to the form of the likelihood function, we can assume the prior function has the same form of distribution, and get the corresponding parameters from normalizing the posterior distribution.

3. to maximize the posterior distribution(MAP) and get the coefficient vector $\mathbf{w}$ *of the linear regression function* $y(\mathbf{x}, \mathbf{w})$*.*

Note 22 *The features of the Bayesian method:*

1. it can naturally contain the regularization item in the log of posterior (only for the the Gaussian distribution, don't know others), which will overcome overfitting, which is one of the most biggest shortcomings of maximized likelihood function (MLP).

2. it is very suitable to the sequential data, we can treat any posterior distribution as the prior distribution for the newly observed data. At the same time, since it is suitable for the sequential data, we don't need to store the whole dataset, which is very efficient for the memory.