EE655000 Machine Learning HW2

107060011 電機資訊學士班 涂皓鈞

Part 2. Implementation

1), 2) MSE screenshot for MLR and BRL:

```
------------------------------
Start BLR training ...
Start BLR testing ...
------------------------------
Start MLR training ...
Start MLR testing ...
------------------------------
MSE of BLR = 0.0072134935657022625, MSE of MLR = 0.008770368791861702.
```

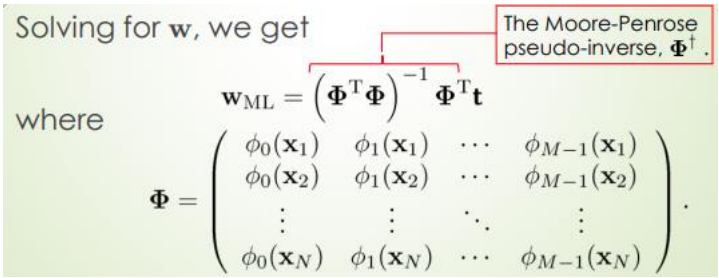
3)

a. The difference between Maximum Likelihood and Bayesian Linear Regression

1. Maximum Likelihood: Assume the regression target as Gaussian random variable, which can be formally represented as follows:

$$P(t|x,W) = N(t|W^T x, \sigma^2)$$

Then to find an optimum W, we use Maximum Likelihood Estimation (MLE), as **the above model is a likelihood**. The optimum W has an analytical solution can be expressed as:



$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

2. Bayesian Linear Regression: Unlike Maximum Likelihood, Bayesian Linear Regression **takes prior into account**. For easy evaluation, **the distribution of prior and posterior are the same form**, so we consider a conjugate prior:

$$P(w) = N(w|m_0, S_0),$$

which is combined with likelihood and it directly gives the form of posterior distribution:

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{3.49}$$

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right) \tag{3.50}$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}. \tag{3.51}$$

we consider a **zero-mean isotropic Gaussian** governed by a single precision parameter α, which can be obtained:

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{3.52}$$

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \tag{3.53}$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}. \tag{3.54}$$

Note that since **the posterior distribution is Gaussian**, its mode coincides with its mean. Thus the **optimum $\mathbf{W}_{MAP}$ is simply equal to $\mathbf{m}_N$**. (α=0.1, β=9 are set here)

b. The impact of different choices of O1 and O2

The following table records the mean square error of Maximum Likelihood and Bayesian Linear Regression in different configurations of O1 and O2. From the table, we can observe that the **mean square error of MLR rapidly rises** in a non-linear way when **O1 and O2 increase**. As for **BLR**, its mean square error also **rises** but **in a smaller amplitude**. I think that's because **there is too much parameters when O1 and O2 increase, resulting in overfitting situation**, and hence the mean square error rises. Further, the **mean square error of BLR is always smaller than that of MLR**. It makes sense that since BLR takes prior distribution into account, it includes much variety in prediction, so it has better performance.

O1	O2	MLR MSE	BLR MSE
2	2	0.007004	0.006958
3	3	0.007228	0.007182
5	5	0.008770	0.007213
7	7	0.027607	0.007617
10	10	1.410484	0.009171
11	11	414.0662	0.008370

Table: MSE of MLR & BLR in different settings of O1 & O2