Machine Learning: Homework 2

Sen Wang

February 16, 2019

1 Linear Regression

The objective function for optimization is,

$$E_D\mathbf{w} = \frac{1}{2}\sum_{n=1}^{N} r_n\{y_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 \tag{1}$$

So by taking partial derivatives with respect to each w_m, we can obtain,

$$\sum_{j=0}^{M}[w_j \sum_{n=1}^{N} r_n\phi_m(\mathbf{x}_n)\phi_j(\mathbf{x}_n)] = \sum_{n=1}^{N} r_n y_n \phi_m(\mathbf{x}_n) \tag{2}$$

if we rearrange we can see that this is a system of linear equations of the form $\mathbf{Aw} = \mathbf{b}$, where,

$$A_{ij} = \sum_{n=1}^{N} r_n\phi_i(\mathbf{x}_n)\phi_j(\mathbf{x}_n), \quad w_j = w_j, \quad b_i = \sum_{n=1}^{N} r_n y_n \phi_i(\mathbf{x}_n) \tag{3}$$

If we think about the SSE as data dependent noise variance, then we treat $\mathbf{w}^T\phi(\mathbf{x}_n)$ as the mean for the given input. However, if we think about it in terms of replicated data points, then, the SSE can be interpreted as the average error across different datasets.

2 Evidence

2.1 Marginal Likelihood with known variance

The marginal likelihood can be obtained by simply using the normalisation constants from lecture 3, which gives,

$$p(\mathbf{t}|\alpha, \beta) = \frac{Z_N}{Z_0 Z_l} = \frac{(2\pi)^{M/2}|S_N|^{1/2}}{(\frac{2\pi}{\alpha})^{M/2}(\frac{2\pi}{\beta})^{N/2}} = \frac{\alpha^{M/2}\beta^{N/2}|S_N|^{1/2}}{(2\pi)^N/2} \tag{4}$$

2.2 Marginal Likelihood of Unknown Variance

(a). Show that the corresponding posterior distribution takes the same functional form.

By expanding the product of the likelihood and the priors, and ignore the normalization constants, we can obtain,

$$p(\mathbf{w}, \beta|\mathbf{t}) \propto \beta^{a_0+\frac{M+N}{2}-1} \exp\{-\frac{\beta}{2}[(\mathbf{t} - \Phi\mathbf{w})^T(\mathbf{t} - \Phi\mathbf{w}) + (\mathbf{w} - \mathbf{m}_0)^T\mathbf{S}_0^{-1}\mathbf{w} - \mathbf{m}_0) - 2b_0]\} \tag{5}$$

If we define,

$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \Phi^T\Phi)^{-1} \tag{6}$$

$$\mathbf{w}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \Phi^T\mathbf{t}) \tag{7}$$

$$a_N = a_0 + \frac{N}{2} \tag{8}$$

$$b_N = b_0 + \frac{1}{2}(\mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{t}^T\mathbf{t} - \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N) \tag{9}$$

Then we can obtain the following results by completing the squares,

$$p(\mathbf{w}, \beta|\mathbf{t}) \propto \beta^{a_N+\frac{M}{2}-1} \exp\{\frac{\beta}{2}[(\mathbf{w} - \mathbf{m}_N)^T\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) + 2b_N]\} \tag{10}$$

$$= \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\mathrm{Gam}(\beta|a_N, b_N) \tag{11}$$

(b) Show the marginal probability of the data.

Similar to part A, we can apply the same relationship through the normalisation constants as,

$$p(\mathbf{t}) = \frac{Z_N}{Z_0 Z_l} \tag{12}$$

$$= (2\pi)^{M/2}|\mathbf{S}_N|^{1/2}\frac{\Gamma(a_N)}{b_N^{a_N}}b_0^{a_0}\Gamma(a_0)(\frac{1}{2\pi})^{M/2}\frac{1}{|\mathbf{S}_0|^{1/2}} \tag{13}$$

$$= \frac{1}{2\pi^{N/2}}\frac{b_0^{a_0}}{b_N^{a_N}}\frac{|\mathbf{S}_N|^{1/2}}{|\mathbf{S}_0|^{1/2}}\frac{\Gamma(a_N)}{\Gamma(a_0)} \tag{14}$$

3 Mixture of Conjugate Priors

3.1 Show that a mixture of conjugate priors is indeed a conjugate prior

If we express the mixture of conjugate priors as,

$$\pi(\theta) = \sum_{i=1}^{K} w_i \pi_i(\theta) \tag{15}$$

where w_i are the corresponding weights we assign to each prior, then we can write the posterior as,

$$\pi(\theta|\mathcal{D}) = \frac{\sum_{i=1}^{K} w_i \pi_i(\theta) f(\mathcal{D}|\theta}{A} = \sum_{i=1}^{K} \frac{w_i}{A} \pi_i(\theta) f(\mathcal{D}|\theta) \tag{16}$$

where the mixture of the priors gives the posterior with the same form as the prior.

3.2 Show a plot of comparison between the prior and the posterior

For the problem at hand, we can easily define the likelihood, prior and posterior as,

$$p(\mathcal{D}|\theta) = \binom{N_1 + N_0}{N_1} \theta^{N_1} (1-\theta)^{N_0} \tag{17}$$

$$p(\theta) = 0.5\text{Beta}(\theta|a_1, b_1) + 0.5\text{Beta}(\theta|a_2, b_2) \tag{18}$$

$$p(\theta|\mathcal{D}) = 0.5\text{Beta}(\theta|N_1 + a_1, N_0 + b_1) + 0.5\text{Beta}(\theta|N_1 + a_2, N_0 + b_2) \tag{19}$$

Plotting the PDF of the distributions above, we obtain the plot below,

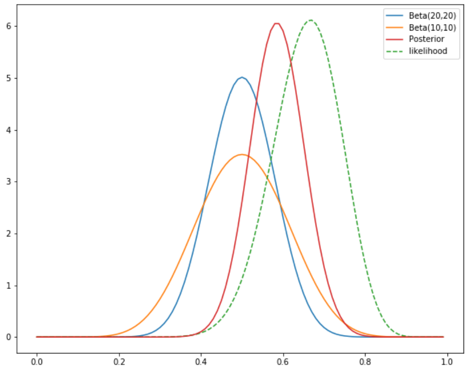

Figure 1: Comparison between prior and posterior

4 Optimal Threshold on Classification Probability

4.1 Show equivalence of decision criteria and derive θ as a functon of λ_{01} and λ_{10}

The equivalence can be shown through the calculation of the expected conditional loss function,

$$\begin{aligned}
\rho(\hat{y}|\mathbf{x}) &= \sum_{y} L(y, \hat{y}) p(y|\mathbf{x}) && (20)\\
&= L(\hat{y}=0, y=1)p(y=1|\mathbf{x}) + L(\hat{y}=1, y=0)p(y=0|\mathbf{x}) && (21)\\
&= \lambda_{01}p(y=1|\mathbf{x}) + \lambda_{10}p(y=0|\mathbf{x}) && (22)\\
&= \begin{cases} \lambda_{01}p(y=1|\mathbf{x}) & \text{if } \hat{y}=0 \\ \lambda_{10}p(y=0|\mathbf{x}) & \text{if } \hat{y}=1 \end{cases} && (23)
\end{aligned}$$

So, we choose $y = 0$ if,

$$\lambda_{01}p_1 < lambda_{10}p_0 \qquad (24)$$

$$p_1 < \frac{\lambda_{10}}{\lambda_{01} + \lambda_{10}} \qquad (25)$$

So we obtain the relation,

$$\theta < \frac{\lambda_{10}}{\lambda_{01} + \lambda_{10}} \qquad (26)$$

4.2 Derive the loss function where the threshold is 0.1

For $\theta = 0.1$, we can use the above relation to find that,

$$\frac{\lambda_{01}}{\lambda_{10}} = 9 \qquad (27)$$

So if we let $\lambda_{10} = a$, then the loss function is defined as,

$$L(\hat{y}, y) = \begin{cases} 9a \text{ if fales negative} \\ a \text{ if fales positive} \end{cases} \qquad (28)$$

5 Bayes Factor

5.1 Show the marginal likelihood

Similar to previous questions, the mariginal likelihood can be simply obtained through the normalisation constants,

$$p(N_1|N) = \frac{Z_N}{Z_0 Z_l} = \frac{\mathrm{B}(N_1+1, N-N_1+1)}{\mathrm{B}(1,1)} \binom{N}{N_1}^{-1} = \frac{1}{N+1} \tag{29}$$

5.2 Derive the Bayes Factor

The null hypothesis assumes fair coin, so the likelihood is,

$$p(\mathcal{D}|M_0) = (\frac{1}{2})^N \tag{30}$$

The alternative hypothesis gives the likelihood of the model as shown in previous part,

$$p(\mathcal{D}|M_1) = \frac{1}{N+1} \tag{31}$$

So calculating the Bayes Factor, we obtain,

$$\mathrm{BF}1,0 = \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{2^N}{N+1} \tag{32}$$

So if we have $N_1 = 0$ and $N = 10$, the Bayes Factor is,

$$\mathrm{BF}1,0 = \frac{2^{10}}{11} >> 1 \tag{33}$$

If we have $N_1 = 90$ and $N = 100$, then,

$$\mathrm{BF}1,0 = \frac{2^{100}}{101} >> 1 \tag{34}$$

6 Behaviour of training set error with increasing sample size, Multi-output regression and Ridge regression

6.1 Explain why with sufficiently complex model, the error on the training set can increase as we get more training data, until we reach some plateau

With complex model, initially the model is fine-tuned to the noise so the training error is very small. However, as the model start generalising, there is decreasing bias but the training error starts to increase. When the current model is optimal, the error reaches plateau.

6.2 Compute MLE for W from the above data

After applying the feature function on the input data, we have,

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{35}$$

We already know $\mathbf{Y}$ from the question, so the MLE of $\mathbf{W}$ can be obtained by,

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{pmatrix} \tag{36}$$

6.3 Show the optimizer of ridge regression

Consider w_0 first, we take derivative of w_0 of the error function and set it to zero, we obtain

$$\sum_{n=1}^{N} y_n - \mathbf{X}_n\mathbf{w} - w_0 = 0 \tag{37}$$

$$\sum_{n=1}^{N} y_n - Nw_0 = 0 \tag{38}$$

$$w_0 = \bar{y} \tag{39}$$

Now for the non-bias parameters, we take derivative of the error function by each parameter, we obtain,

$$\sum_{n=1}^{N} (y_n + \mathbf{X}_n\mathbf{w} - w_0)\mathbf{X}_n^j + 2\lambda w_j = 0 \tag{40}$$

After rearranging, we can obtain,

$$\sum_{n=1}^{N} \mathbf{X}_n^j \mathbf{X}_n \mathbf{w} + \lambda w_j = \sum_{n=1}^{N} y_n \mathbf{X}_n^j \tag{41}$$

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})\mathbf{w} = \mathbf{X}^T \mathbf{y} \tag{42}$$

And we obtain the final result,

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \tag{43}$$

6.4 Performing a linear regression with MLE

(a) Compute the unknown coefficients based on MLE with $M = 2, 4, 10, 14$. Compute and plot the mean square error for the training and the test set.

Using the above results from Part B, we can compute the MLE for **w** and the exact values can be accessed from the Jupyter Notebook attachment. The MSE for different model and the plots for both training and test sets are also plotted below.

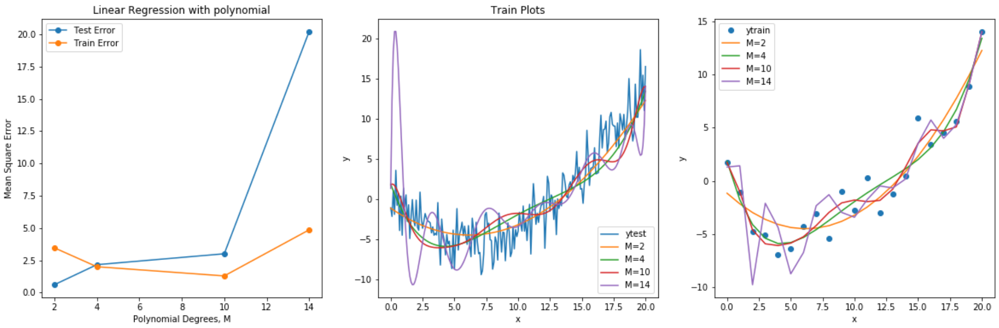

Figure 2: Error for train and set sets.

(b) Repeat with ridge regression

Using the above results from Part C, we can compute the MLE for **w** and the exact values can be accessed from the Jupyter Notebook attachment. The MSE for different model and the plots for both training and test sets are also plotted below.

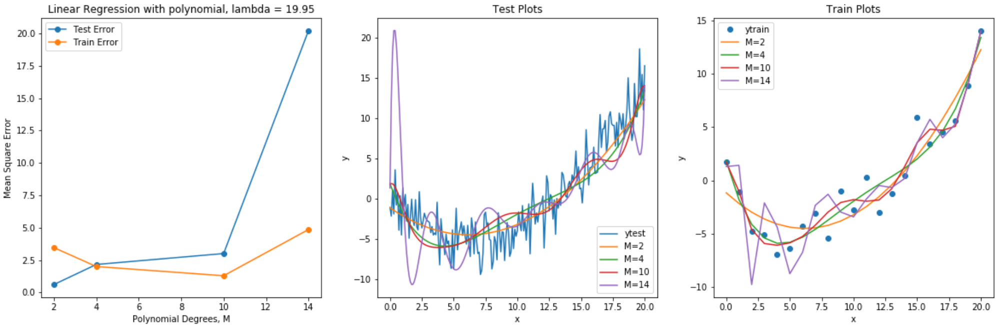

Figure 3: Error for train and set sets.

7 Bayesian linear regression

7.1 Derive the expression for the posterior, marginal posterior, predictive distribution and the model evidence

By expanding the product of the likelihood and the priors, and ignore the normalization constants, we can obtain,

$$p(\mathbf{w}, \sigma^2|\mathbf{y}) \propto (\frac{1}{\sigma^2})^{a+\frac{M+N}{2}+1} \exp\{-\frac{1}{2\sigma^2}[(\mathbf{y}-\Phi\mathbf{w})^T(\mathbf{t}-\Phi\mathbf{w}) + \gamma^{-1}\mathbf{w}^T\mathbf{w} - 2b]\} \tag{44}$$

If we define,

$$\mathbf{V}_N = (\gamma^{-1}\mathbf{I} + \Phi^T\Phi)^{-1} \tag{45}$$

$$\mathbf{w}_N = \mathbf{V}_N\Phi^T\mathbf{y} \tag{46}$$

$$a_N = a_0 + \frac{N}{2} \tag{47}$$

$$b_N = b_0 + \frac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{m}_N^T\mathbf{V}_N^{-1}\mathbf{m}_N) \tag{48}$$

Then we can obtain the following results by completing the squares,

$$p(\mathbf{w}, \sigma^2|\mathbf{t}) \propto (\frac{1}{\sigma^2})^{a_N+\frac{M}{2}+1} \exp\{\frac{1}{2\sigma^2}[(\mathbf{w}-\mathbf{w}_N)^T\mathbf{V}_N^{-1}(\mathbf{w}-\mathbf{w}_N) + 2b_N]\} \tag{49}$$

$$= \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \beta^{-1}\mathbf{V}_N)\mathrm{IG}(\sigma^2|a_N, b_N) \tag{50}$$

For the marginal posterior, we apply the sum rule of probability,

$$p(\mathbf{w}|\mathcal{D}) = \int_0^\infty p(\mathbf{w}|\sigma^2|\mathcal{D})d\sigma^2 \tag{51}$$

$$\propto [1 + \frac{(\mathbf{w}-\mathbf{w}_N)^T\mathbf{V}_N^{-1}(\mathbf{w}-\mathbf{w}_N}{2b_N}]^{-\frac{2a_N+D}{2}} \tag{52}$$

$$= \mathcal{T}_D(\mathbf{w}_N, \frac{b_N}{a_N}\mathbf{V}_N, 2a_B) \tag{53}$$

For the predictive distribution, we will apply a concept,

$$p(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \int\int p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2)p(\mathbf{w}, \sigma^2)d\mathbf{w}d\sigma^2 \tag{54}$$

$$\propto \int\int (\frac{1}{\sigma^2})^{a_N+\frac{m+M}{2}+1} \exp\{-\frac{1}{2\sigma^2}[(\mathbf{y}-\Phi\mathbf{w})^T(\mathbf{y}-\Phi\mathbf{w}) \tag{55}$$

$$+ (\mathbf{w}-\mathbf{w}_N)^T\mathbf{V}_N^{-1}(\mathbf{w}-\mathbf{w}_N) + 2b_N]\}d\mathbf{w}d\sigma^2 \tag{56}$$

Now by rearranging the part in exponential, we can obtain,

$$(\mathbf{y}-\Phi\mathbf{w})^T(\mathbf{y}-\Phi\mathbf{w}) + (\mathbf{w}-\mathbf{w}_N)^T\mathbf{V}_N^{-1}(\mathbf{w}-\mathbf{w}_N) + 2b_N] \tag{57}$$

$$=(\mathbf{w} - (\Phi^T\Phi + \mathbf{V}_N^{-1})^{-1}(\Phi^T\mathbf{y} + \mathbf{V}_N^{-1}\mathbf{w}_N))^T(\Phi^T\Phi + \mathbf{V}_N^{-1})^T \tag{58}$$

$$(\mathbf{w} - (\Phi^T\Phi + \mathbf{V}_N^{-1})^{-1}(\Phi^T\mathbf{y} + \mathbf{V}_N^{-1}\mathbf{w}_N)) + 2\beta \tag{59}$$

$$2\beta = -(\Phi^T\mathbf{y} + \mathbf{V}_N^{-1}\mathbf{w}_N)^T(\Phi^T\Phi + \mathbf{V}_N^{-1})^T(\Phi^T\mathbf{y} + \mathbf{V}_N^{-1}\mathbf{w}_N) + \mathbf{w}_N^T\mathbf{V}_N^{-1}\mathbf{w}_N + \mathbf{y}^T\mathbf{y} + 2b_N \tag{60}$$

We then obtain the predictive distribution, which is

$$p(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \mathcal{T}_m(\mathbf{y}|\Phi\mathbf{w}_N, \frac{b_N}{a_N}(\mathbf{I}_m - \Phi\mathbf{V}_N\Phi^T), 2a_N) \tag{61}$$

For the model evidence, we again use the normalisation constants,

$$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_l} = (2\pi)^{-N/2}\gamma^{-M/2}|\mathbf{V}_N|^{1/2}\frac{\Gamma(a_N)}{\Gamma(a)}\frac{b^a}{b_N^{a_N}} \tag{62}$$

7.2 Plot the predictive mean and the predictive error bars.

The model output and the plots are as shown below.

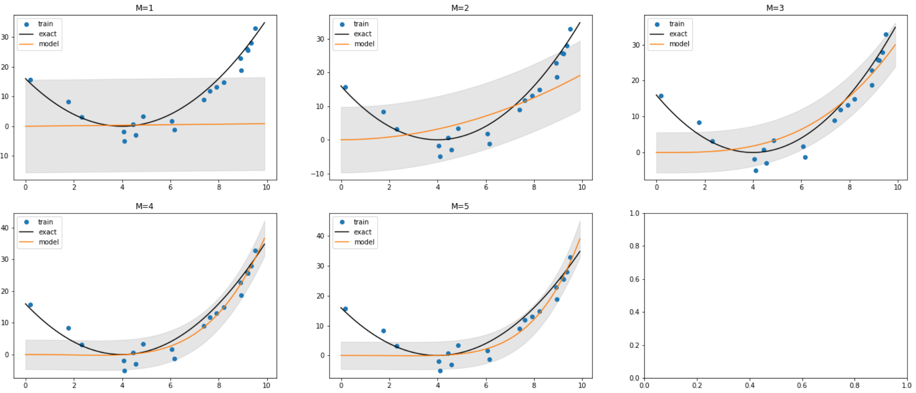

Figure 4: Non-centered model output

7.3 Sample plots

The sample plots are as shown below.

It can be observed that as the model complexity increases, there are increasing variances of the model.

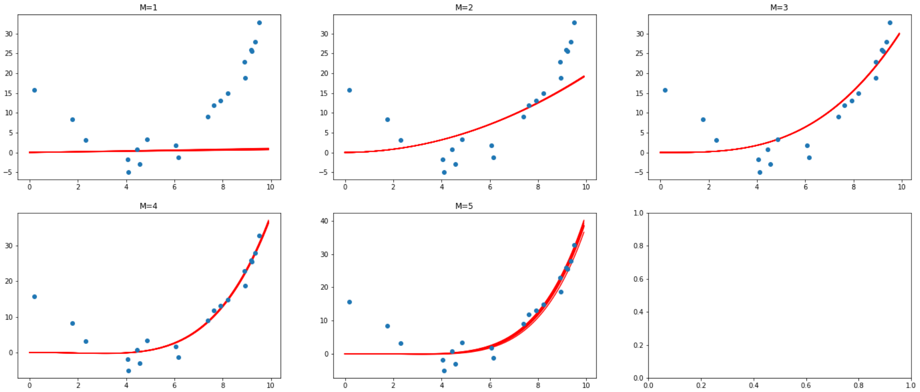

Figure 5: Sample Plots

7.4 Model Evidence

Using expression from part A, the model evidence is plotted below. The model output and the plots are as shown below.

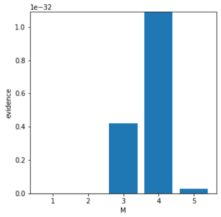

Figure 6: Model Evidence

7.5 Data Centering

The equivalent plot as part B is performed on centered data and the results are as shown below, The model output and the plots are as shown below. It can be observed that while the errors are similar, there is a modification on the bias term.

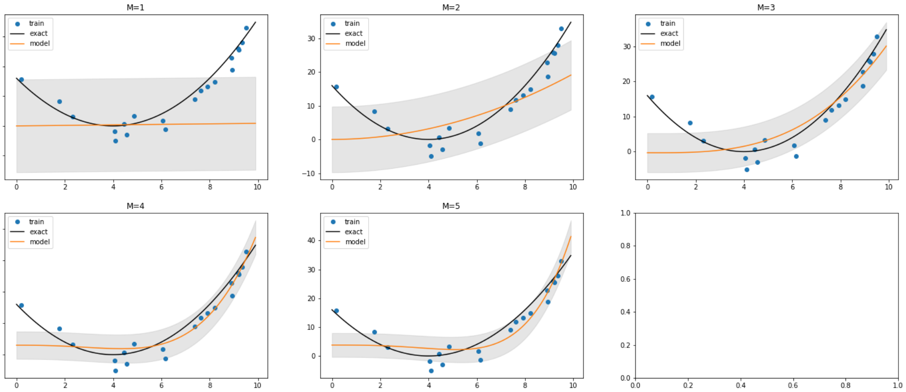

Figure 7: Centered model output