Machine Learning (Homework 1)

Due date : 10/28

1 Bayesian Linear Regression (15%)

For a given input value x, the corresponding target value t is assumed as a Gaussian distribution $p(t|x,\mathbf{w},\beta) = \mathrm{N}(t|y(x,\mathbf{w}),\beta^{-1})$ and the prior distribution of **w** is also assumed as a Gaussian distribution $p(\mathbf{w}|\alpha) = \mathrm{N}(\mathbf{w}|0,\alpha^{-1}\mathbf{I})$.

A linear regression function is expressed by $y(x,\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\varphi(x)$ where $\varphi(x)$ is a basis function. We are not only interested in the value **w** but also in making prediction of t for new test data x. We multiply the likelihood function of new data $p(t|x,\mathbf{w},\beta)$ and the posterior distribution of the training data $p(\mathbf{w}|\mathbf{x},\mathbf{t})$ and take the integral over **w** to find the predictive distribution

$$\int_{-\infty}^{\infty} p(t|x,\mathbf{w},\beta)p(\mathbf{w}|\mathbf{x},\mathbf{t})d\mathbf{w}.$$

Please derive this predictive distribution which is a Gaussian distribution of the form $p(t|x,\mathbf{x},\mathbf{t}) = \mathrm{N}(t|m(x),s^2(x))$ where

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n$$

$$s^2(x) = \beta^{-1} + \varphi(x)^{\mathrm{T}}\mathbf{S}\varphi(x).$$

Here, the matrix $\mathbf{S}^{-1}$ is given by $\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}}$. You may use the formulas shown in page 93.

2 Jensens Inequality (10%)

Convexity implies

$$f(\lambda a + (1 - \lambda)b) \le \lambda f(a) + (1 - \lambda)f(b) \tag{1}$$

A convex function satisfies

$$f\Big(\sum_{i=1}^{M} \lambda_i x_i\Big) \le \sum_{i=1}^{M} \lambda_i f(x_i) \tag{2}$$

where $\lambda_i \ge 0$ and $\sum_i \lambda_i = 1$. Please use the technique of proof by induction to derive equation (2) from equation (1).

MLHW1 - 0753407
劉宭守

1 Bayesian Linear Regression

$$P(t|x,X,t) = \int_{-\infty}^{\infty} p(t|x,W,\beta)\,p(W|X,t)\,dw$$

1° $P(W|X,t) \propto p(t|X,W)\,p(W|\alpha)$

by equations in page 93,

$P(t|X,W) = N(t|W^T\Phi(X), \beta^{-1}I) = N(t|W^TA+b, L^{-1})$

$\rightarrow A = \Phi(X)^T, b=0, L=\beta I$

$p(W|\alpha) = N(W|0,\alpha^{-1}I) = N(W|\mu,\Lambda^{-1}) \rightarrow \mu=0, \Lambda=\alpha I$

$P(W|X,t) = N(W|\Sigma\{A^TL(W-b)+\Lambda\mu\}, \Sigma)$, where $\Sigma = (\alpha I + A^TLA)^{-1}$

Substitute $A = \Phi(X)^T, b=0, L=\beta I, \mu=0, \Lambda=\alpha I$

$\Rightarrow N(W|S(\Phi^T(X)\beta t), S)$ where $S = (\alpha I + \Phi(X)\beta\Phi(X)^T)^{-1}$

2° By equations in page 93.

$p(t|W,X) = N(t|W^T\Phi(X), \beta^{-1}) = N(t|W^TA+b, L^{-1}) \rightarrow A = \Phi(X), b=0, L=\beta I$

$p(W|X,t) = N(W|S(\beta\Phi(X)t), S) = p(W|\mu,\Lambda^{-1}) \rightarrow \mu = S(\beta\Phi(X)t, \Lambda^{-1}) = S$

Substitute $A = \Phi(X)^T, b=0, L=\beta I, \mu=0, \Lambda=\alpha I$

$P(t|x,X,t) = N(t|A\mu+b, L^{-1}+A\Lambda^{-1}A^T)$

$$= N(t|\beta\Phi(x)^TS\Phi(X)t, \beta^{-1}+\Phi(x)^TS\Phi(x))$$

2 Proof Jensen's Inequality by induction

假設 $n=m\geq 2$ 時 $\sum_{i=1}^{m} f(\lambda_i x_i) \leq \sum_{i=1}^{m} \lambda_i f(x_i)$ 成立，欲證明 $\sum_{i=1}^{m+1} f(\lambda_i x_i) \leq \sum_{i=1}^{m+1} \lambda_i f(x_i)$

$\because$ all x_i's lie in interval $[a,b]$, so does linear combination $\sum_{i=1}^{m}\lambda_i x_i$, The combination can be represented differently: $\sum_{i=1}^{m+1}\lambda_i x_i = \lambda_{m+1}x_{m+1} + \sum_{i=1}^{m}\lambda_i x_i = \lambda_{m+1}x_{m+1} + (1-\lambda_{m+1})\sum_{i=1}^{m}\frac{\lambda_i}{1-\lambda_{m+1}}x_i$

且已知 $\boxed{\sum_{i=1}^{m}\frac{\lambda_i}{1-\lambda_{m+1}} = 1}$, $\sum_{i=1}^{m}\frac{\lambda_i}{1-\lambda_{m+1}}x_i \in [a,b]$,

$$f\left(\sum_{i=1}^{m+1}\lambda_i x_i\right) = f\left(\lambda_{m+1}x_{m+1} + (1-\lambda_{m+1})\sum_{i=1}^{m}\frac{\lambda_i}{1-\lambda_{m+1}}x_i\right)$$

$$\leq \lambda_{m+1}f(x_{m+1}) + (1-\lambda_{m+1})\,f\left(\sum_{i=1}^{m}\frac{\lambda_i}{1-\lambda_{m+1}}x_i\right)$$

$$\leq \lambda_{m+1}f(x_{m+1}) + (1-\lambda_{m+1})\boxed{\sum_{i=1}^{m}\frac{\lambda_i}{1-\lambda_{m+1}}}\,f(x_i)$$

$$= \lambda_{m+1}f(x_{m+1}) + \sum_{i=1}^{m}\lambda_i f(x_i)$$

$$= \sum_{i=1}^{m+1}\lambda_i f(x_i)$$

3 Polynomial Regression (75%)

In real-world applications, the dimension of data is usually more than one. Here, the California Housing Prices data set is given in (housing.csv). The data set pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. Please build a regression model for estimation of the values given in item median house value (i.e. house price) by applying a polynomial function

$$y(x,\mathbf{w}) = w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M = \sum_{j=0}^{M} wjxj$$

and minimizing the error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 .$$

A general polynomial with coefficients for data dimension up to two $\mathbf{x} = [x_1\ x_2]^{>}$ is formed by

$$y(\mathbf{x},\mathbf{w}) = w_0 + \sum_{i=1}^{D} w_ix_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij}x_ix_j.$$

The data set contains 3 columns of different features that a certain house has. In this exercise, the first 90% samples are used as the Training Set and the last 10% samples are used as the Testing Set.

Data Description

Number of Instances: 20640

Number of Attributes: 4 (3-dimensional input + 1-dimensional target)

Attribute Information:

- Total Rooms: Total number of rooms within a block
- Population: Total number of people residing within a block

- Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- Median House Value: Median house value for households within a block (measured in US Dollars)

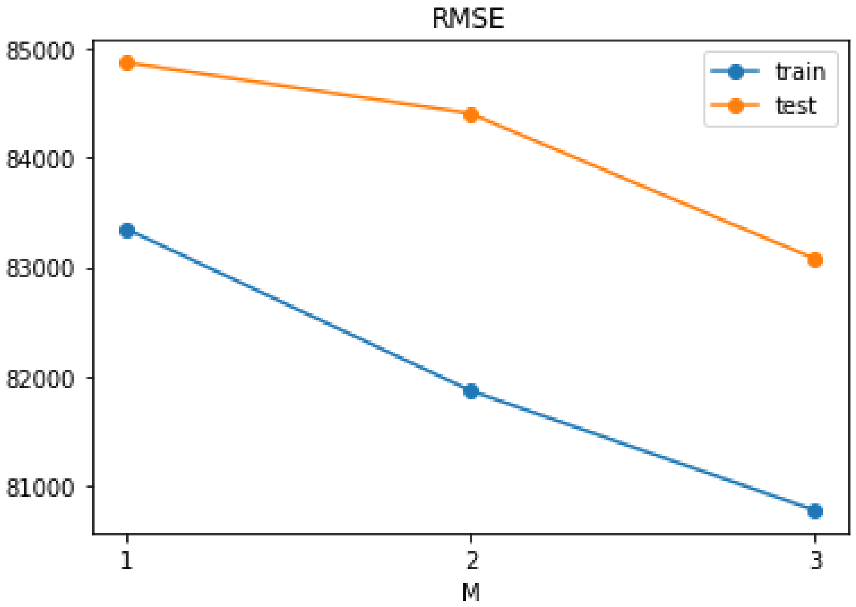
1. In the training stage, please apply the polynomials of **order *M* = 1 to *M* = 3** over the 3-dimensional input data, **evaluate the corresponding Root-Mean-Square error** ($E_{RMS} = \sqrt{2E(\mathbf{w})/N}$) on the Training Set and Test Set and plot their RMS error versus order *M*. Describe in details about what you see in the plot.

   橫軸為 M，而縱軸為 RMSE 的值，可以觀察到兩件事情：

   a. 隨著 M 增加，RMSE 皆下降：因為當函式越複雜時越 Robust，錯誤率越低。

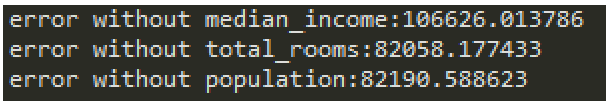
   b. Training set 的 RMSE 低於 Testing set 的 RMSE：因為是用 training data 去 train model，所以餵入 Testing data 的時候 Error 會比較高是正常的。

   另外，因為只用三個 attribute 就要預測房價有點太少，因此錯誤會到達八萬多是正常的現象。



2. Please apply the polynomials of order *M* = 3 and **select the most contributive attribute** or dimension which has the lowest RMS error on the Training Set.

將三個 attribute 分別抽出一個以後去預估並計算錯誤率，可以看到在沒有 median_income 這項 attribute 的情況下，錯誤率最高，因此可以此推斷出 median_income 是 **most contributive attribute**。



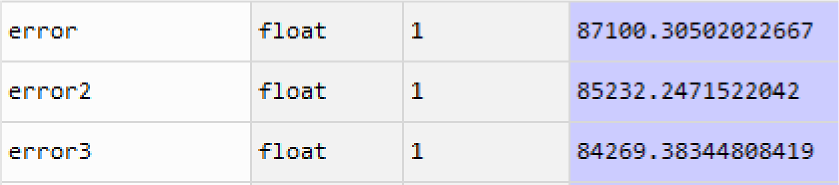
```
error without median_income:106626.013786
error without total_rooms:82058.177433
error without population:82190.588623
```

3. **Considering the regularized error function**

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

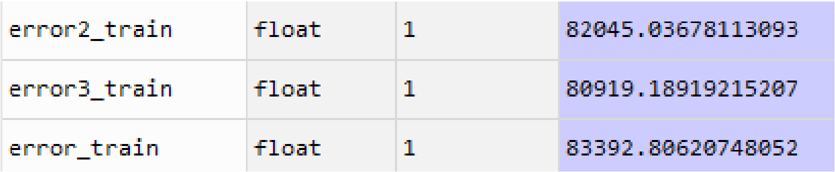
where $\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + w_2^2 + \ldots + w_M^2$. Please set two values for regularization parameter as ***λ* = 0.1 and *λ* = 0.001** and repeat **part 1**. (note: $E_{RMS} = \sqrt{2E(\mathbf{w})/N}$ is calculated using *E*(w) not E_e(w)) Also, plot the regularized regression result on Training Set and Testing Set for various order *M* from 1 to 3. Compare the result with different *λ* and describe the difference between **part 1** and **part 3**.

圖一是尚未加入 λ 時 M 分別為 1-3 時的 RMSE



error	float	1	87100.30502022667
error2	float	1	85232.2471522042
error3	float	1	84269.38344808419

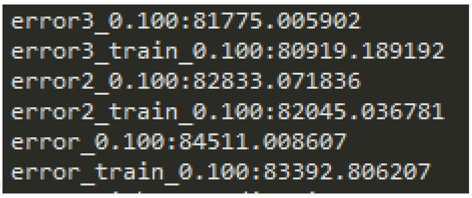
圖一(test)



error2_train	float	1	82045.03678113093
error3_train	float	1	80919.18919215207
error_train	float	1	83392.80620748052

圖二(train)

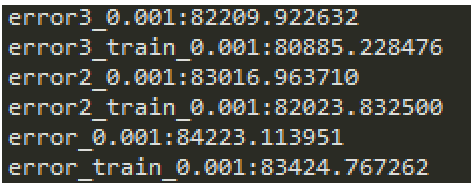
圖三是加入 λ=0.1 時 M 分別為 1-3 時的 RMSE



```
error3_0.100:81775.005902
error3_train_0.100:80919.189192
error2_0.100:82833.071836
error2_train_0.100:82045.036781
error_0.100:84511.008607
error_train_0.100:83392.806207
```

圖三(train&testλ=0.1)

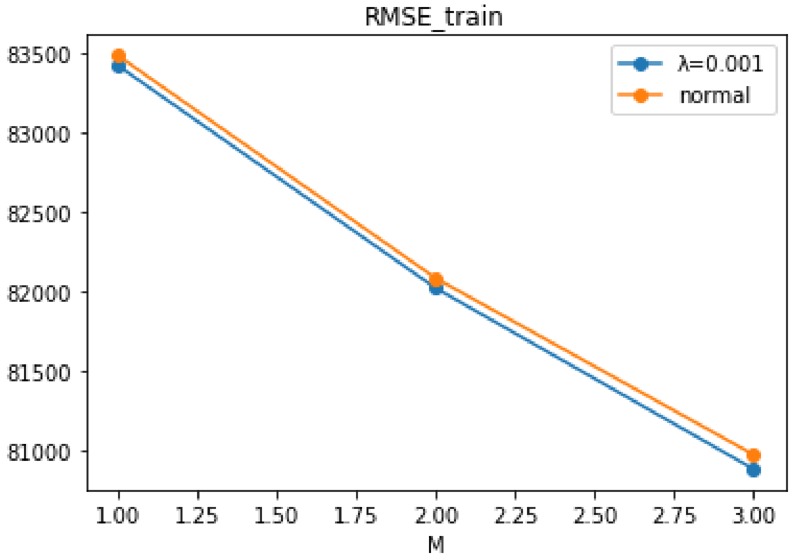
圖四是加入 λ=0.001 時 M 分別為 1-3 時的 RMSE



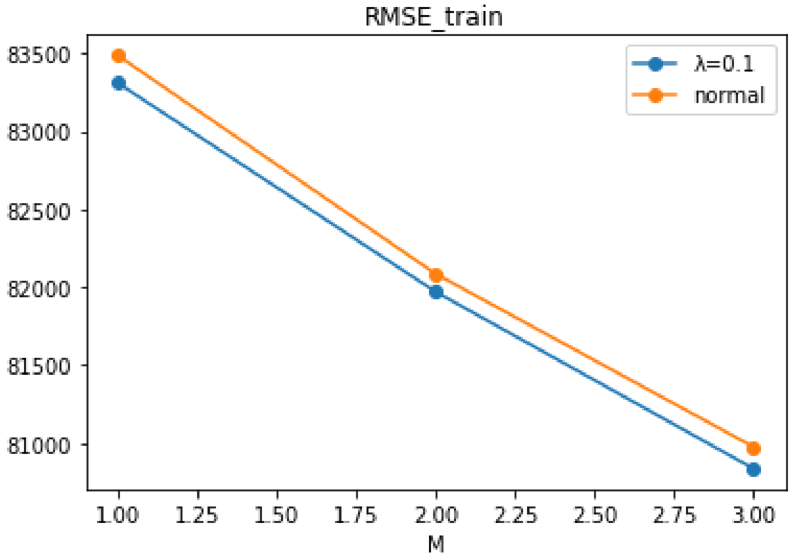
```
error3_0.001:82209.922632
error3_train_0.001:80885.228476
error2_0.001:83016.963710
error2_train_0.001:82023.832500
error_0.001:84223.113951
error_train_0.001:83424.767262
```

圖四(train&test_λ=0.001)

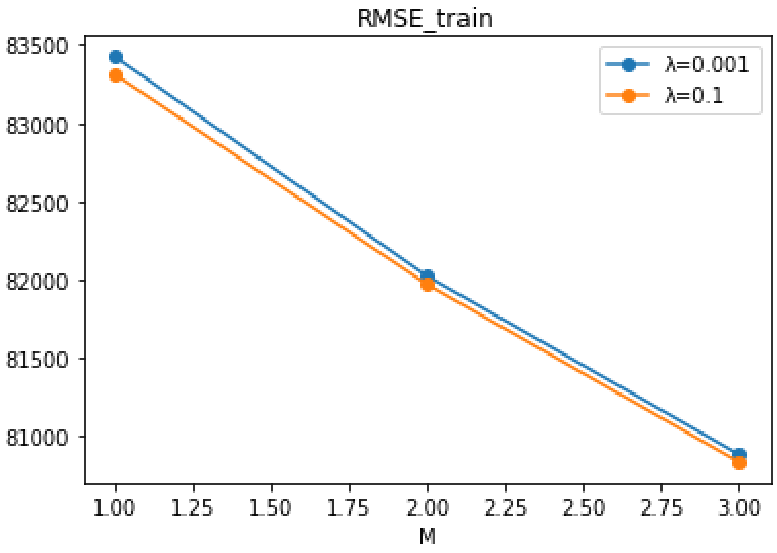
比較上圖三種狀況，可以看到加入 λ 這項 regularized term 以後，真的有讓 RMSE 下降，不過因為沒有發生 overfitting 的問題，所以加上這項的效果並不大。另外隨著 λ 的改變，圖五為 λ=0.001，圖六為 λ=0.1，可以看到兩者讓 RMSE 下降的效果不太一樣，當 λ 比較大的時候對整體 RMSE 的影響也較大，如圖七所示。



圖五



圖六


RMSE_train
83500
83000
82500
82000
81500
81000
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
M
λ=0.001
λ=0.1

圖七