Machine Learning (Homework 1)

0753420 郭家瑄

1. Bayesian Linear Regression

$$\int_{-\infty}^{\infty} p(t|x,\underline{w},\beta)\,p(\underline{w}|\underline{x},\underline{t})\,dw = p(t|x,\underline{x},\underline{t})$$

1. $p(\underline{w}|\underline{x},\underline{t}) \propto p(\underline{t}|\underline{x},\underline{w})\,p(\underline{w}|\alpha)$

by equations in p.93

$$p(\underline{t}|\underline{x},\underline{w}) = N(\underline{t}|\underline{w}^T\Phi(x), \beta^{-1}I) = N(\underline{t}|\underline{w}^T A+b, L^{-1})$$

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

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

$$\rightarrow \mu=0,\ \Lambda=\alpha I$$

$$p(\underline{w}|\underline{x},\underline{t}) = N(\underline{w}|\Sigma\{A^TL(\underline{w}-b)+\Lambda\mu\},\Sigma), \text{ where } \Sigma=(\alpha I+A^TLA)^{-1}$$

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

$$\rightarrow N(\underline{w}|S(\Phi^T(x)\beta\underline{t}), S), \text{ where } S=(\alpha I+\Phi(x)\beta\Phi(x)^T)^{-1}$$

2. by equations in p.93

$$p(t|\underline{w},x)=N(t|\underline{w}^T\Phi(x), \beta^{-1}) = N(t|\underline{w}^TA+b, L^{-1})$$

$$\rightarrow A=\Phi(x),\ b=0,\ L=\beta I$$

$$p(\underline{w}|\underline{x},\underline{t}) = N(\underline{w}|S(\beta\Phi(x)\underline{t}, S) = p(\underline{w}|\mu,\Lambda^{-1})$$

$$\rightarrow \mu = S(\beta\Phi(x)\underline{t}),\ \Lambda^{-1}=S$$

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

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

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

2. Jensens Inequality

when X takes on two values the inequality is

$$f(\lambda x_1+(1-\lambda)x_2)\le \lambda f(x_1)+(1-\lambda)f(x_2)$$

This is true by the defintion of convex functions.

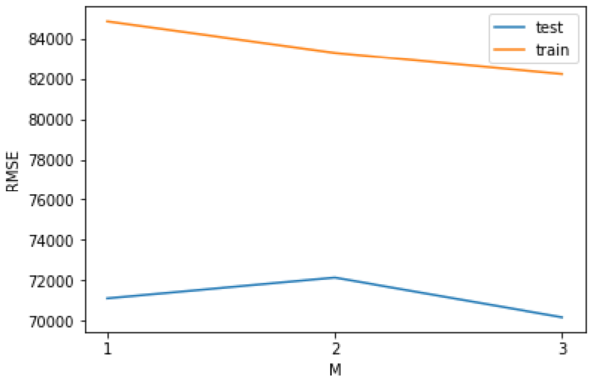
Inductive hypothesis: suppose the theorem is true for distribution with $k-1$ values.

$$\begin{aligned}
\sum_{i=1}^{M}\lambda_i f(x_i) &= \lambda_M f(x_M)+(1-\lambda_M)\sum_{i=1}^{M-1}\frac{\lambda_i}{1-\lambda_M}f(x_i)\\
&\ge \lambda_M f(x_M)+(1-\lambda_M)f\left(\sum_{i=1}^{M-1}\frac{\lambda_i}{1-\lambda_M}\cdot x_i\right)\\
&\ge f\left(\lambda_M x_M+(1-\cancel{\lambda_M})\sum_{i=1}^{M-1}\frac{\lambda_i}{\cancel{1-\lambda_M}}\cdot x_i\right)\\
&= f\left(\sum_{i=1}^{M}\lambda_i x_i\right)
\end{aligned}$$

$$\Rightarrow \sum_{i=1}^{M}\lambda_i f(x_i)\ge f\left(\sum_{i=1}^{M}\lambda_i x_i\right)_{\#}$$

3. Polynomial Regression

(1)



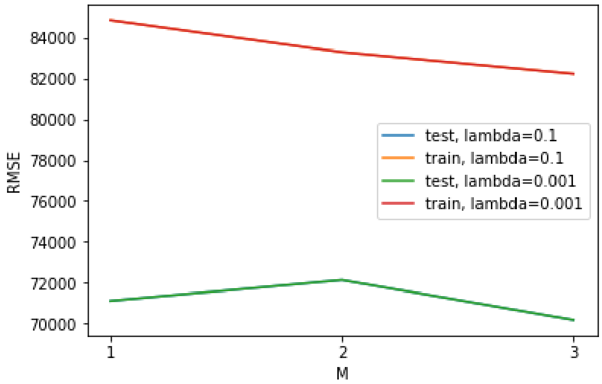
不管是 Train 或是 Test，計算出來的 RMSE 都非常大，由此可知，要用一個簡單線性模型、跟 3 個特徵值就要預測房價，是非常不實際的。另外，應該是因為在切 Train data 跟 Test data 的時候，資料沒有特別 shuffle，就剛好分到了 Test 的 prediction error 小於 train error。

(2)

```
RMSE after remove feature [ total room ] = 83730.34554561331
RMSE after remove feature [ population ] = 83893.56376074895
RMSE after remove feature [ median income ] = 107630.60095137352
```

上圖是實驗結果，當 M=3 的時候，拿掉 total room 這個 feature 的 RMSE 是 83730.34554561331，拿掉 population 的時候 RMSE 為 83893.56376074895，而當我們把 median income 這個 feature 拿掉時，RMSE 飆升到 107630.60095137352。所以我們可以得知，**median income** 是最模型學習時非常重要的參考指標。

(3)



從圖中可能看不出來，但是其實有 4 條線。藍線、綠線幾乎重疊在一起；橘線、紅線也幾乎重疊在一起。是因為不管是 M=1,2,3，這個 model 都還沒 overfitting 整個資料集，Error 都還很大，所以在這個情況下加入 regularization term 並不會有甚麼很好的效果，所以看到他們兩兩近乎交疊，是正常的狀況。