Machine learning HW1

0510894 電機 4D 翁紹恩

一、Bayesian Linear Regression

0510894 翁紹恩

$P(t|x, \mathbf{X}, \mathbf{t}) = \int_{-\infty}^{\infty} P(t|x, \overline{w}) P(\overline{w}|\mathbf{X}, \mathbf{t})\, d\overline{w}$ evidence

$P(t|x, \overline{w}) = N(t|y(x, \overline{w}), \beta^{-1}) = N(t|\overline{w}^T\phi(x), \beta^{-1})$ likelihood

$P(\overline{w}) = N(\overline{w}|0, \alpha^{-1}I)$ prior.

by Marginal and Conditional Gaussians equations in book P93.

$P(x) = N(x|\mu, \Lambda^{-1})$ prior

$P(y|x) = N(y|Ax + b, L^{-1})$ likelihood

$\Rightarrow P(y) = N(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$ evidence

$P(x|y) = N(x|\Sigma\{A^TL(y-b) + \Lambda\mu\}, \Sigma)$ posterior

$\because P(\overline{w}|\mathbf{X}, \mathbf{t}) \propto P(\mathbf{t}|\mathbf{X}, \overline{w}) P(\overline{w})$

$P(\mathbf{t}|\mathbf{X}, \overline{w}) = N(\mathbf{t}|\overline{w}^T\phi(x), \beta^{-1}I)$ likelihood

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

$P(\overline{w}) = N(\overline{w}|0, \alpha^{-1}I)$ prior

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

$\Sigma = (\Lambda + A^TLA)^{-1} = (\alpha I + \phi(x)\beta I\phi(x)^T)^{-1} = (\alpha I + \phi(x)\beta\phi(x)^T)^{-1}$

$P(\overline{w}|\mathbf{X}, \mathbf{t}) = N(\overline{w}|\Sigma\{\phi(x)\beta\mathbf{t}\}, \Sigma)$ posterior

$\Rightarrow P(t|x, \overline{w}) = N(t|\overline{w}^T\phi(x), \beta^{-1})$ likelihood

$A' = \phi(x)^T,\ b' = 0,\ L' = \beta$

$P(\overline{w}|\mathbf{X}, \mathbf{t}) = N(\overline{w}|\Sigma\{\phi(x)\beta\mathbf{t}\}, \Sigma)$ prior $\quad S^{-1} = \alpha I + \beta\sum_{n=1}^{N}\phi(x_n)\phi(x)^T$

$= N(\overline{w}|S(\phi(x)\beta\mathbf{t}), S) \quad S = \Sigma.$

$= N(\overline{w}|\mu', \Lambda^{-1})$

$\mu' = S(\beta\phi(x)\mathbf{t}),\ \Lambda = S^{-1},\ \Sigma' = (\Lambda + A'^TL'A')^{-1} = (S^{-1} + \phi(x)\beta\phi(x)^T)^{-1}$

$P(t|x, \mathbf{X}, \mathbf{t}) = N(t|\phi(x)^TS(\beta\phi(x)\mathbf{t}), \beta^{-1} + \phi(x)^TS\phi(x))$ evidence

$m(x) = \phi(x)^TS\beta\phi(x)\mathbf{t} \quad S^2(x) = \beta^{-1} + \phi(x)^TS\phi(x)$

$= \beta\phi(x)^TS\sum_{n=1}^{N}\phi(x_n)t_n$

#

1. Feature selection

在這部分，training set 和 validation set 的比例為 7：3(768：328)

a. RMS error：

	M=1	M=2
Training	3.6079	2.8550
Validation	6.1140	6.9692

➔雖然在 M=1 training set 時 RMS error 比較大，但是 validation set 出來的結果卻比較小，可以看出，在 training set 時，M=2 是比較 fit 整個數據的，但反而因此 overfitting，導致在 validation 上的誤差較大。

b. Analyze the weights of polynomial models(M=1)

The remove data	RMS error
AMB_TEMP	3.6149
CH4	3.6119
CO	3.7076
NMHC	3.6083
NO	3.6079
NO2	3.6083
NOx	3.6081
O3	3.6214
PM10	5.6394
RAINFALL	3.6103
RH	3.6235
SO2	3.6452
THC	3.6108
WD_HR	3.6560
WIND_DIREC	3.6435
WIND_SPEED	3.6091
WS_HR	3.6112

➔從本表中可以看出，當移除第九筆資料，PM10 時，所得出的 RMS error 最大且有最明顯變化，沒有 PM10 時會導致極大誤差，因此 PM10 是影響 Training set RMS error 最大的因素。

2. Maximum likelihood approach

在本題當中使用了三種訓練模型(Polynomial, Gaussian, Sigmoidal)做探討

a. 未使用 N-fold cross validation

使用全部參數(x M=1)套入

Polynomial ➔ $\Phi=[1, x_1, x_2 \ldots, x_D]$

Gaussian➔$\Phi=[1, \exp(\frac{-(x_1-\mu_1)^2}{2\sigma_1^2}), \exp(\frac{-(x_2-\mu_2)^2}{2\sigma_2^2}) \ldots, \exp(\frac{-(x_D-\mu_D)^2}{2\sigma_D^2})]$

Sigmoidal➔$\Phi=[1, \frac{1}{1+\exp(-\frac{(x_1-\mu_1)}{\sigma_1})}, \frac{1}{1+\exp(-\frac{(x_2-\mu_2)}{\sigma_2})} \ldots, \frac{1}{1+\exp(-\frac{(x_D-\mu_D)}{\sigma_D})}]$

每一個 model 前面加上 1 做為 intercept term

x M=1

	Training	Testing
Gaussian	8.3566	8.5781
Polynomial	3.6078	6.1140
Sigmoidal	3.9221	5.2801

x M=2(沒有刪資料)

	Training	Testing
Gaussian	4.6459	9.5798
Polynomial	2.8549	6.969
Sigmoidal	2.8169	7.0366

從兩者資料可以看出，只要使用 M=2 的模型，因為輸入的參數太多，因此都會 overfitting，所以要減少使用 data 的量。

根據第 1.b 題得出的數據，我取前五個對數據影響較大的參數，分別是 CO、O3、PM10、RH、WD_HR 只用這五個參數做二階相乘。

Polynomial ➔ $\Phi=[1, x_1, x_2 \ldots, x_D, x_3x_8, x_3x_9 \ldots, x_{11}x_{14}]$

Gaussian➔$\Phi=[1, \exp(\frac{-(x_1-\mu_1)^2}{2\sigma_1^2}), \exp(\frac{-(x_2-\mu_2)^2}{2\sigma_2^2}) \ldots, \exp(\frac{-(x_D-\mu_D)^2}{2\sigma_D^2}),$

$\exp(\frac{-(x_3x_8-\mu_{3*8})^2}{2\sigma_{3*8}^2}), \exp(\frac{-(x_3x_9-\mu_{3*9})^2}{2\sigma_{3*9}^2}) \ldots, \exp(\frac{-(x_{11}x_{14}-\mu_{11*14})^2}{2\sigma_{11*14}^2})]$

Sigmoidal➔$\Phi=[1, \frac{1}{1+\exp(-\frac{(x_1-\mu_1)}{\sigma_1})}, \frac{1}{1+\exp(-\frac{(x_2-\mu_2)}{\sigma_2})} \ldots, \frac{1}{1+\exp(-\frac{(x_D-\mu_D)}{\sigma_D})},$

$\frac{1}{1+\exp(-\frac{(x_3x_8-\mu_{3*8})}{\sigma_{3*8}})}, \frac{1}{1+\exp(-\frac{(x_3x_9-\mu_{3*9})}{\sigma_{3*9}})}, \ldots, \frac{1}{1+\exp(-\frac{(x_{11}x_{14}-\mu_{11*14})}{\sigma_{11*14}})}]$

x M=2(二階取五筆資料)

	Training	Testing
Gaussian	6.5369	7.1877
Polynomial	3.3279	6.0078
Sigmoidal	3.5148	5.3715

由上表和 M=1 時比較，可以發現 overfitting 的問題就消失了。

b. 使用 4-fold cross validation

x M=1

	Training	Testing

Gaussian	8.2385	8.6905
Polynomial	3.9964	4.8668
Sigmoidal	4.1535	4.2553

x M=2(沒有刪資料)

	Training	Testing
Gaussian	4.8783	8.1891
Polynomial	3.2170	5.7974
Sigmoidal	3.1329	5.5475

x M=2(二階取五筆資料)

	Training	Testing
Gaussian	6.4846	7.1467
Polynomial	3.7406	4.7350
Sigmoidal	3.8363	4.4117

使用 M=2 的時候，很明顯的沒刪參數時因為參數太多，所以有 overfitting 的現象，在如上題適當的選取幾個影響大的參數後，overfitting 的結果明顯改善。

而用 cross validation 的方法後，可以觀察到 error 整體都有下降的趨勢，讓我們對於參數的調整可以更加客觀。

3. Maximum a posteriori approach

和第 2 題的模型差異為 w 的改變，w=$(\lambda I + \emptyset^T\emptyset)^{-1}\emptyset^T y$

a. 取 λ = 10

x M=1

	Training	Testing
Gaussian	8.5289	8.5433
Polynomial	3.7287	6.4873
Sigmoidal	4.4205	5.3803

x M=2(沒有刪資料)

	Training	Testing
Gaussian	6.6072	7.6802
Polynomial	2.8896	6.6814
Sigmoidal	3.6998	5.1839

x M=2(二階取五筆資料)

	Training	Testing
Gaussian	7.5493	7.6582
Polynomial	3.3916	6.2601
Sigmoidal	3.8463	5.1569

可以看到加入 regularization term 之後，本來 overfitting 的結果，現在和有先挑選過參數的結果是差不多的。

b. 使用 4-fold cross validation

x M=1

	Training	Testing
Gaussian	8.3913	8.6591
Polynomial	4.2162	5.1141
Sigmoidal	4.6589	4.5394

x M=2(沒有刪資料)

	Training	Testing
Gaussian	6.5611	7.3550
Polynomial	3.2779	5.4216
Sigmoidal	3.9390	4.1457

x M=2(二階取五筆資料)

	Training	Testing
Gaussian	7.4215	7.7062
Polynomial	3.8923	4.8512
Sigmoidal	4.1454	4.1233

使用 cross-validation 在 Gaussian model 上並沒有差很多，但是在 Polynomial 和 Sigmoidal model 上，testing set error 都有明顯下降。

c. Compare the result between maximum likelihood approach and maximum a posteriori approach.

綜合上述資料，可以發現兩者最主要的差異，在參數很多、模型十分複雜時，maximum a posteriori approach 可以有效的減少 RMS error，如此一來對於我們在訓練模型上也較為方便不用做參數篩選，但是可以看出，還是有做適當篩選的模型可以得到平均起來較好的結果。