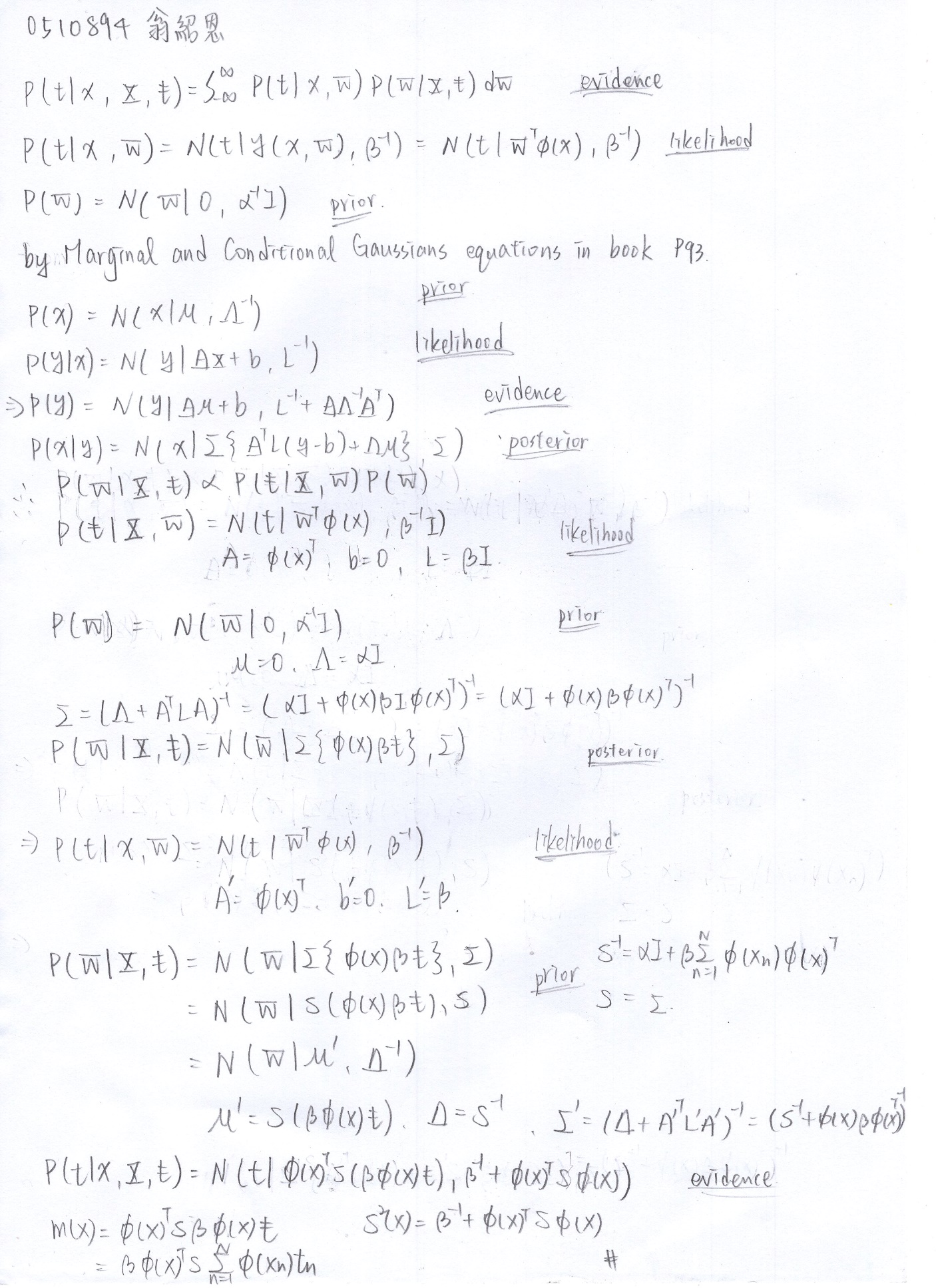
Machine learning HW1

0510894 電機4D 翁紹恩

1. Bayesian Linear Regression



1. Feature selection

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

1. 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。

1. 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。

1. Maximum likelihood approach

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

1. 未使用N-fold cross validation

使用全部參數(x M=1)套入

Polynomial 🡺 Φ=[1,,…,]

Gaussian🡺Φ=[1,,…, ]

Sigmoidal🡺Φ=[1, , …, ]

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 🡺 Φ=[1,,…,,,…,]

Gaussian🡺Φ=[1,,…, , , …, ]

Sigmoidal🡺Φ=[1, , …, , , , … , ]

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。

1. 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。

1. Maximum a posteriori approach

和第2題的模型差異為w的改變，w=

1. 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。

1. 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。

1. 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。