李少琪 0860908 AI智慧學程 HW1 ML

1 Bayesian Linear Regression.

$$P(t|x,X,t)=\int_{-\infty}^{\infty} P(t|\overset{③}{x},w)P(w|\overset{④}{X},t)dw$$

$$P(w|X,t) \propto P(t|X,w)P(w|\alpha)$$

① $P(t|X,w)=N(t|y(X,w),\beta^{-1})$ marginal Gaussian distribution.

$=N(t|w^T\sum_{n=1}^{N}\phi(x_n),\beta^{-1}I)$ $\because P(y|x)=N(y|Ax+b,L^{-1})$

$=N(t|w^TA+b,L^{-1})$

$\therefore A=(\sum_{n=1}^{N}\phi(x_n))^T,\ b=0,\ L=\beta I$

② $P(w|\alpha)=N(w|0,\alpha^{-1}I)$ $\because P(x)=N(x|\mu,\Lambda^{-1})$

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

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

$\rightarrow \because P(x|y)=N(x|\Sigma\{A^TL(y-b)+\Lambda\mu\},\Sigma),\ \Sigma=(\Lambda+A^TLA)^{-1}$

$\therefore P(w|X,t)=N(w|\Sigma\{A^TL(t-b)+\Lambda\mu\},\Sigma),\ \Sigma=(\alpha I+\beta\sum_{n=1}^{N}\phi(x_n)\phi(x_n)^T)^{-1}$

$=N(w|\beta S\sum_{n=1}^{N}\phi(x_n)t_n\},S),$

$S=\Sigma=(\alpha I+\beta\sum_{n=1}^{N}\phi(x_n)\phi(x)^T)^{-1}$

③ $P(t|x,w)=N(t|w^T\phi(x),\beta^{-1})$

$=N(t|w^TA+b,L^{-1})$

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

④ $P(w|X,t)=N(w|\beta S\sum_{n=1}^{N}\phi(x_n)t_n,S)$

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

$\therefore \mu=\beta S\sum_{n=1}^{N}\phi(x_n)t_n,\ \Lambda=S^{-1}$

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

$=N(t|\beta\phi^T(x)S\sum_{n=1}^{N}\phi(x_n)t_n,\beta^{-1}+\phi(x)^TS\phi(x)).$

$=N(t|m(x),s^2(x)).$

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

$s^2(x)=\beta^{-1}+\phi(x)^TS\phi(x).$

$\begin{cases} P(t|x,w)\rightarrow P(y|x) \\ P(w|\alpha)\rightarrow P(x) \\ P(w|x,t)\rightarrow P(x|y) \end{cases}$

$\begin{cases} P(t|x,w)\rightarrow P(y|x) \\ P(w|x,t)\rightarrow P(x) \\ P(t|x,X,t)\rightarrow P(y). \end{cases}$

2 Linear Regression

1. Feature selection

(a)

Size of Training set: 75% of dataset

Size of Validation set: 25% of dataset

M = 1

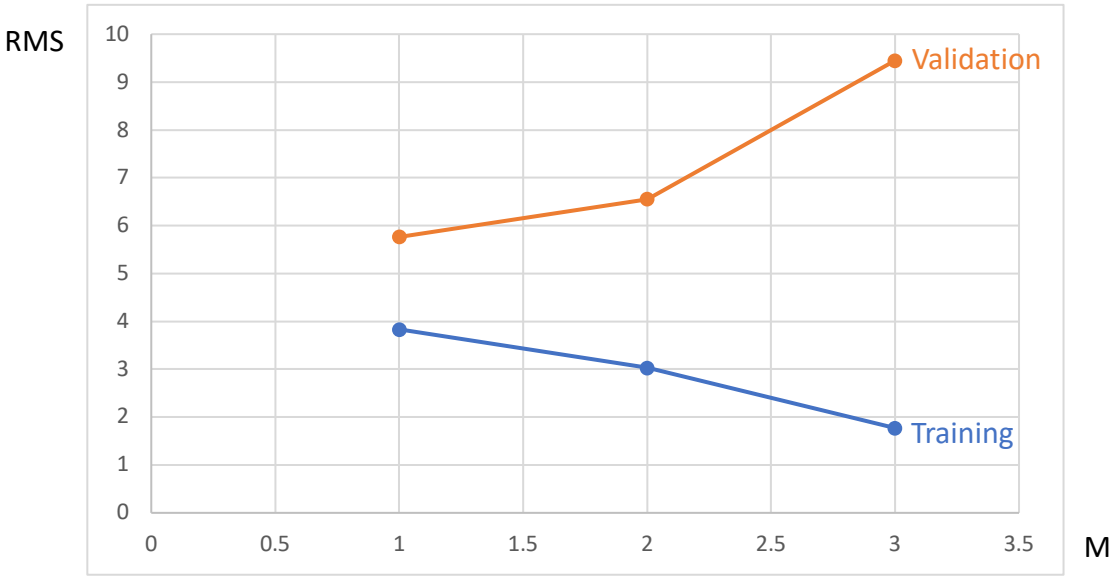
$$\emptyset_j(X) = [1\ x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7\ x_8\ x_9\ x_{10}\ x_{11}\ x_{12}\ x_{13}\ x_{14}\ x_{15}\ x_{16}\ x_{17}]$$

M = 2

$$\emptyset_j(X) = [1\ x_1\ x_2\ \dots\ x_{17}\ x_1x_2\ x_1x_3\ \dots\ x_{16}x_{17}\ x_1^2\ x_2^2\ \dots\ x_{17}^2]$$

➔ Weight 越多, Model 越複雜, 越容易 overfitting

M	1	2	3
Training RMS	3.8366	3.0300	1.7722
Validation RMS	5.7634	6.5501	9.4487



(b) 每次作訓練時, 選一個 feature 不加入訓練, 若 RMS error 變大, 則該 feature 對於資料較為重要。

Feature	AMB_TEMP	CH4	CO	NMHC	NO	NO2	NOx	O3	PM10	RAINFALL
TrainRMS	3.843	3.84	3.94	3.84	3.84	3.84	3.83	3.85	5.97	3.85
ValidRMS	5.70	5.79	6.03	5.80	5.75	5.76	5.76	5.73	6.39	5.82
Feature	RH	SO2	THC	WD_HR		WIND_DIREC		WIND_SPEED		WS_HR
TrainRMS	3.85	3.85	3.84	3.87		3.86		3.84		3.84
ValidRMS	5.75	5.75	5.77	5.77		5.77		5.77		5.80

可以發現去掉 PM10 做訓練的話, RMS error 最大, 可以推測 PM10 對於資料較為有重要性。

2. Maximum likelihood approach

(a) Features: PM10, CO, RAINFALL

Basis function: 先 polynomial 再 sigmoid

M = 1

$$\emptyset_j(X) = [\sigma(1)\ \sigma(x_1)\ \sigma(x_2) \dots \sigma(x_{17})]$$

M = 2

$$\emptyset_j(X) = [\sigma(1)\ \sigma(x_1)\ \sigma(x_2) \dots \sigma(x_4)\ \sigma(x_1x_2)\ \sigma(x_1x_3) \dots \sigma(x_3x_4)\ \sigma(x_1^2) \dots \sigma(x_4^2)]$$

(b) 4 folds

	1 ~ 274	275 ~ 548	549 ~ 822	823 ~ 1096
F_1	valid	train		
F_2				
F_3				
F_4				

	F_1		F_2		F_3		F_4	
	Train	Valid	Train	Valid	Train	Valid	Train	Valid
M = 1	6.743	8.211	7.013	7.44	7.250	6.732	7.271	6.748
M = 2	6.695	8.018	6.875	7.455	7.104	6.705	7.132	6.876
M = 3	6.091	11.575	6.322	7.739	6.695	5.940	6.617	6.308
M = 4	5.558	1851.157	5.453	6.517	5.718	7.999	5.513	5.543
M = 5	5.024	2755.871	4.058	12.135	3.964	54.219	4.101	4.651
M = 6	4.840	444.567	2.871	29394.821	3.830	10.808	2.021	137.833

從上表可以看出模型越複雜參數越多，越容易 overfitting

3. Maximum a posteriori approach

(a) $W_{MAP} = argmax_w\{\frac{\beta}{2}\sum_{n=1}^{N}(W\emptyset(x_n) - t_n) + \frac{\alpha}{2}W^TW\}$

→ $W^* = (\alpha I + \beta\emptyset(X)^T\emptyset(X))^{-1}\emptyset(X)^Tt$

令 $\alpha = 0.5,\ \beta = 1$

	F_1		F_2		F_3		F_4	
	Train	Valid	Train	Valid	Train	Valid	Train	Valid
M = 1	8.618	11.672	9.004	11.179	8.957	11.508	9.309	13.136
M = 2	8.237	11.125	8.548	12.148	8.562	10.107	8.831	15.051
M = 3	7.321	11.645	7.670	21.294	7.849	13.504	8.001	19.412
M = 4	6.368	7.993	6.646	12.841	6.935	8.912	6.861	13.751
M = 5	5.935	7.696	6.141	10.566	6.471	7.383	6.334	11.623
M = 6	5.775	7.825	5.899	7.440	6.276	6.402	6.122	9.499

(b)因為 MAP 的 error function 是將 sum-of-squares error 再做 regularization, 所以有對 weight 加上懲罰, 比較不會 overfitting, 從上面兩張表也可發現 MAP 沒有 overfitting.