機器學習理論 HW2

108064535 陳文遠

Step 1. 使用 Maximum Likelihood & Least Squares 法來訓練 Training_set.csv 檔中的資料

- ☐ 假設有一線性方程式，而題目給定此自變數為一個 feature vector 如下：

$$\emptyset(\mathrm{x}) = [\emptyset_1(\mathrm{x}), \emptyset_2(\mathrm{x}), \dots, \emptyset_\mathrm{P}(\mathrm{x}), \emptyset_{\mathrm{P}+1}, \emptyset_{\mathrm{P}+2}]^\mathrm{T} \quad for \;\; P = O_1 \times O_2$$

其中，

$$\emptyset_\mathrm{k}(x) = e^{-\frac{(x_1-\mu_i)^2}{2s_1} - \frac{(x_2-\mu_j)^2}{2s_2}} \quad for \;\; 1 \le i \le O_1, 1 \le j \le O_2$$

$$\mathrm{k} = \mathrm{O}_2 \times (i-1) + j$$

$$\mu_\mathrm{i} = s_1 \times (i-1) + x_{1\_min} \;, \qquad \mu_j = s_2 \times (j-1) + x_{2\_min}$$

$$\mathrm{s}_1 = \frac{x_{1\_max} - x_{1\_min}}{O_1 - 1} \;, \qquad \mathrm{s}_2 = \frac{x_{2\_max} - x_{2\_min}}{O_2 - 1}$$

最後，

$$\emptyset_{\mathrm{P}+1}(x) = x_3(Research \;\; Experience) \;\; and \;\; \emptyset_{\mathrm{P}+2}(x) = 1(\mathrm{bias})$$

Step 2. 將 Testing_set.csv 中的測試資料送進 Maximum Likelihood & Least Squares 的 Model 中來預測 chance of admit 以及計算其 squared error $\left(\mathbf{y}(\mathbf{x}) - \mathbf{t}(\mathbf{x})\right)^2$

- ☐ 假設線性函數為(ε 為 Added Gaussian Noise)：

$$\mathrm{t} = \mathrm{y}(\mathrm{x}, \mathrm{w}) + \varepsilon = \sum_{j=1}^{P+2} w_j \emptyset_j(x) + \varepsilon \;\; where \;\; p(\varepsilon|\beta) = N(\varepsilon|0, \beta^{-1})$$

其機率分布可以寫成：

$$\mathrm{p}(\mathrm{t}|\mathrm{x}, \mathrm{w}, \beta) = \mathrm{N}(\mathrm{t}|\mathrm{y}(\mathrm{x}, \mathrm{w}), \beta^{-1})$$

其中，$\mathrm{X} = \{\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}\}$ 是我們輸入的資料(分數)，$\mathrm{t} = [\mathrm{t}_1, \mathrm{t}_2, \dots, \mathrm{t}_\mathrm{n}]^\mathrm{T}$ 為 Targets (chance of admit)，則我們可以獲得下方的 likelihood function：

$$\mathrm{p}(\mathrm{t}|\mathrm{X}, \mathrm{w}, \beta) = \prod_{n=1}^{N} N(t_n|w^T\emptyset(x_n), \beta^{-1})$$

再將 likelihood function 取對數：

$$\ln[\mathrm{p}(\mathrm{t}|\mathrm{X}, \mathrm{w}, \beta)] = \sum_{n=1}^{N} N(t_n|w^T\emptyset(x_n), \beta^{-1})$$

$$= \frac{N}{2}\ln(\beta) - \frac{N}{2}\ln(2\pi) - \beta \times \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\emptyset(x_n)\}^2$$

最後將上述式子對 w 微分令為 0 來求極值：

$$\frac{d}{dw}\ln[\mathrm{p}(\mathrm{t}|\mathrm{X},\mathrm{w},\beta)] = \beta\sum_{n=1}^{N}\{t_n - w^T\emptyset(x_n)\}\emptyset(x_n)^T = 0$$

$$\Rightarrow \mathrm{w_{ML}} = (\Phi^T\Phi)^{-1}\Phi^T t \;\text{；}\; proved$$

- 接下來，將 Testing_csv 檔中的測試資料代入 $\emptyset_{\mathrm{k}}(x)$ 來構成 testing data 的 feature vector ($\emptyset$(x))：

$$\emptyset_{\mathrm{k}}(x) = e^{-\frac{(x_1-\mu_i)^2}{2s_1}-\frac{(x_2-\mu_j)^2}{2s_2}} \quad for\ \ 1 \le i \le O_1, 1 \le j \le O_2$$

同樣地，

$$\emptyset_{\mathrm{P+1}}(x) = x_3(Research\ \ Experience)\ \ and\ \ \emptyset_{\mathrm{P+2}}(x) = 1(\mathrm{bias})$$

- 最後求出 predict 的值：

$$\hat{y} = \Phi\mathrm{w_{ML}} = \emptyset(\mathrm{x})\mathrm{w_{ML}}$$

Squared Error 為：

$$\mathrm{Squared\ \ Error} = \big(\mathrm{y(x)} - \mathrm{t(x)}\big)^2 = (\hat{y} - \mathrm{y})^2 = (\hat{y} - \mathrm{x}_4)^2$$

<u>Step 3.</u> 使用 Bayesian Linear Regression 法來訓練 Training_csv 檔中的資料

- 根據貝式定理我們得知

$$\mathrm{p(w|t) \propto p(t|w)p(w)}$$

假設事前機率 (priori probability) 為

$$\mathrm{p(w) \sim N(w|m_0, S_0) = N(w|0, \alpha^{-1}I)}$$

$$\mathrm{m_0} = 0, \qquad S_0 = \alpha^{-1}I$$

假設後驗機率 (posteriori probability) 為

$$\mathrm{p(w|t) \sim N(w|m_N, S_N)}$$

經推導後可得 $\mathrm{m_N} = \beta S_N\Phi^T t$ 以及 $\mathrm{S_N^{-1}} = \alpha I + \beta\Phi^T\Phi$，其中 Φ 與 **<u>Step 1.</u>** 的求法一樣

- 在此實驗中，我僅先將 α 以及 β 設為 1

<u>Step 4.</u> 將 Testing_set.csv 中的測試資料送進 Bayesian Linear Regression 的 Model 中來預測 w, chance of admit 以及計算其 squared error $\big(\mathbf{y(x) - t(x)}\big)^2$

- 權重值為 $\mathrm{w = w_{MAP} = m_N}$，此題計算出的 w 分別為：

-0.18731298 0.07518969 0.03911943 0.49466673 0.35914929

☐ 接下來，將 Testing_csv 檔中的測試資料代入 $\emptyset_{\mathrm{k}}(x)$ 來構成 testing data 的 feature vector ($\emptyset$(x))：

$$\emptyset_{\mathrm{k}}(x) = e^{-\frac{(x_1-\mu_i)^2}{2s_1}-\frac{(x_2-\mu_j)^2}{2s_2}} \quad for \ \ 1 \le i \le O_1, 1 \le j \le O_2$$

同樣地，

$$\emptyset_{\mathrm{P+1}}(x) = x_3(Research \ \ Experience) \ \ and \ \ \emptyset_{\mathrm{P+2}}(x) = 1(\text{bias})$$

☐ 最後求出 predict 的值：

$$\hat{y} = \Phi \mathrm{w_{MAP}} = \emptyset(\mathrm{x})\mathrm{w_{MAP}}$$

Squared Error 為：

$$\text{Squared Error} = \left(\mathrm{y(x)} - \mathrm{t(x)}\right)^2 = (\hat{y} - \mathrm{y})^2 = (\hat{y} - \mathrm{x_4})^2$$

<u>**Step 5.**</u> 討論

1. Maximum likelihood and least squares 以及 Bayesian linear regression 的差異：

已知貝式定理如下，其中 θ 為欲估計的數，D 為輸入的資料集

$$\mathrm{p(\theta|D)} = \frac{p(D|\theta) \times p(\theta)}{p(D)} \Rightarrow posterior = \frac{likelihood \times prior}{evidence}$$

Maximum likelihood and least squares 的目的是為了找出一個 θ 來使得 likelihood p(D|θ)。換句話說在此方法中，他將 $\frac{p(\theta)}{p(D)}$ 視為一個常數而非隨機變數，並且不需要借助 prior probability 來估計 θ

而在 Bayesian linear regression 中會完全計算出 posterior probability p(θ|D)，並且在此方法中會將 θ 視為隨機變數，倘若 p(θ|D) 的變異數足夠小，就將他的期望值視為估計值。Bayesian 法與 Maximum likelihood 法的關鍵區別就是他允許採用 prior information

2. $\mathbf{O_1}, \boldsymbol{O_2}$ 對結果的影響

為了比較其影響，我分別將 Maximum likelihood 法以及 Bayesian 法對 100 的樣本估計的 squared error 進行加總，接著修改 $\mathrm{O_1}, O_2$ 的值來觀察何時會得到最小 squared error

從觀察的結果中會發現 $\mathrm{O_1}, O_2$ 對 Bayesian 法的結果影響不大，對 Maximum likelihood 法的影響較大，而當 $\mathrm{O_1} = 2, O_2 = 4$ 時我可以得到最小的總 squared error

當 $\mathrm{O_1} = 2,\ O_2 = 4$ 時，Maximum likelihood 法之 100 個樣本的 squared error 加總為 0.4363359203046397，而 Bayesian 法之 100 個樣本的 squared error 加總為 0.44922167547056285