106064701 蔡昀霖

1.

1.

(a) 令 $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}$, $\bar{x} = \frac{x_1+x_2}{2}$.

HW1
106064701
蔡昀霖

$$\Rightarrow X^T X^{-1} = \frac{1}{(x_2-x_1)^2} \cdot \begin{bmatrix} x_1^2+x_2^2 & x_1+x_2 \\ x_1+x_2 & 2 \end{bmatrix}$$ 左下右上都少一個負號

$$\Rightarrow SE(\hat{\beta}_j)^2 = \left[(X^T X)^{-1} \cdot \sigma^2 \right]_{jj}$$

$$SE(\hat{\beta}_1)^2 = \frac{2\sigma^2}{(x_2-x_1)^2} = \frac{\sigma^2}{\frac{2(x_1-x_2)^2}{4}} = \frac{\sigma^2}{\frac{(x_1-x_2)^2}{4} + \frac{(x_2-x_1)^2}{4}}$$

$$= \frac{\sigma^2}{(x_1-\bar{x})^2+(x_2-\bar{x})^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{2} (x_i-\bar{x})^2} \quad \#$$

$$SE(\hat{\beta}_0)^2 = \frac{\sigma^2(x_1^2+x_2^2)}{(x_2-x_1)^2}$$

$$= \sigma^2 \left[\frac{\frac{(x_1+x_2)^2}{2} + \frac{(x_1-x_2)^2}{2}}{(x_1-x_2)^2} \right]$$

$$= \sigma^2 \cdot \left[\frac{1}{2} + \frac{\bar{x}}{\sum_{i=1}^{2}(x_i-\bar{x})^2} \right] \quad \#$$

2.

2. (8%+8%+4%) In linear regression, we adopt the linear model

$$Y = X^T\beta + \epsilon,$$

where $X = (1, X_1, \ldots, X_p)^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Let

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \Big(\text{or concisely written as } \{\mathbf{X}, \mathbf{y}\}\Big)$$

be the set of available data points that are generated independently by the above model.

(a) Find β that maximizes the likelihood ratio $p(\mathbf{y}|\mathbf{X}, \beta)$, assuming that σ^2 is known.

(b) Now suppose that the entries of β are i.i.d. $\mathcal{N}(0, \gamma^2)$. Find β that maximizes the posterior probability

$$p(\beta|\mathbf{y}, \mathbf{X}).$$

(c) Comment on the similarities and differences between least squares linear regression and the above schemes.

(a)

$Y = X^T\beta + \epsilon$, 由題目知ϵ ~ N (0, σ^2)，那麼 Y 為 Normal Distribution 的線

性組合，所以 Y~($X^T\beta$, σ^2)， $f_Y(y|X;\beta) = (2\pi\sigma^2)^{-0.5} * \exp(\frac{-1}{2}\left(\frac{y-X^T\beta}{\sigma}\right)^2)$。

→所以 Likelihood Function:

$L(\beta)=\prod_{i=1}^{N} f_Y(y_i|X;\beta)$

$= \prod_{i=1}^{N}(2\pi\sigma^2)^{-0.5} * \exp\left(\frac{-1}{2}\left(\frac{y_i-x_i^T\beta}{\sigma}\right)^2\right)$

$= (2\pi\sigma^2)^{-N/2} * \exp\left(\frac{-1}{2\sigma^2} * \sum_{i=1}^{N}(y_i - x_i^T\beta)^2\right)$ ， i:表示觀察的數目

→log-Likelihood:

$\ln(L(\beta)) = \sum_{i=1}^{N} ln(f_Y(y_i|X,\beta)) = \frac{-N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{-1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2$

→Maximum Likelihood:

對 log-Likelihood 最大化等效於對 Likelihood 做最大化。

求取全微分:$\nabla_\beta(\ln(L(\beta)))$

$= \nabla_\beta\left(-\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2\right)$

$= \frac{1}{\sigma^2}\sum_{i=1}^{N} x_i^T(y_i - x_i\beta)$

$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i^T y_i - \sum_{i=1}^{N} x_i^T x_i\beta\right)$

左邊的圖示抓來的，他的 xi 是我的 xi.transpose!!

→“極值”必發生在全微分=0(梯度為零)

$\sum_{i=1}^{N} x_i^T y_i - \sum_{i=1}^{N} x_i^T x_i\beta = 0 \qquad \beta = \left(\sum_{i=1}^{N} x_i^T x_i\right)^{-1}\sum_{i=1}^{N} x_i^T y_i = (X^T X)^{-1} X^T y$ #得解

左邊的圖示抓來的，他的 xi 是我的 xi.transpose!!

(b)這種算法稱為 最大後驗估計(Maximum-a-Posteriori (MAP) Estimation)

我先將$P(\beta|y;X)$利用貝氏定理化為以下式子

$P(\beta|y,X) = \frac{P(y|\beta,X)P(\beta|X)}{P(y|X)}$ 這裡感覺有點像直接先把 |X 隔開來化簡

$\rightarrow$ 因為 X Given $\rightarrow \frac{P(y|\beta,X)P(\beta|X)}{C(由於X給定)}$

由(a)知式子可以化簡，並且我們要求取 B 使的後驗機率最大

$\rightarrow \underset{\beta}{\operatorname{argmax}}(\frac{\prod_{i=1}^{N} P(y_i|\beta,X)P(\beta|X)}{C})$

同樣的，我們先對$P(\beta|y;X)$取 ln，並由於 p(y,X)是一個常數，所以會得到以下等效的式子:

$$\rightarrow \underset{\beta}{\operatorname{argmax}}\left(\sum_{i=1}^{N}\Big(\ln\big(P(y_i|\beta,X)\big)+\ln\big(p(\beta|X)\big)\Big)\right),$$

$$where\ p(\beta)=(2\pi\gamma^{\,2})^{-0.5}*\exp\left(\frac{-1}{2}\left(\frac{\beta}{\gamma}\right)^2\right)$$

$$\rightarrow \underset{\beta}{\operatorname{argmax}}\left(\sum_{i=1}^{N}\Big(\ln\big(P(y_i|\beta;X)\big)\ \Big)\ -\ln((2\pi\gamma^{\,2})^{0.5})-\frac{1}{2}\left(\frac{\beta}{\gamma}\right)^2\right)-(A)$$

這邊要注意: $P(y_i|\beta;X)$ 意味著在給定 Data X 下 y 在 β 下發生的機率，這說明了其機率分布為$(2\pi\sigma^2)^{-0.5}*\exp\left(\frac{-1}{2}\left(\frac{y_i-x_i^T\beta}{\sigma}\right)^2\right)$

對$\left(\sum_{i=1}^{N}\Big(\ln\big(P(y_i|\beta,X)\big)\ \Big)\ -\ln((2\pi\gamma^{\,2})^{0.5})-\frac{1}{2}\left(\frac{\beta}{\gamma}\right)^2\right)$做$\beta$的偏微分得

$\sum_{i=1}^{N}\left(\frac{y_i-x_i^T\beta}{\sigma^2}\right)*x_i\ \ -\left(\frac{\beta}{\gamma^2}\right)$,則最大值出現在左式為零時。

$\sum_{i=1}^{N}\left(\frac{y_i-x_i^T\beta}{\sigma^2}\right)*x_i\ =\left(\frac{\beta}{\gamma^2}\right)\rightarrow\sum_{i=1}^{N}\left(\frac{y_i}{\sigma^2}\right)=\frac{\beta}{\gamma^2}+\frac{\beta}{\sigma^2}*\sum_{i=1}^{N}(x_i^T)\rightarrow\beta=$

$\frac{\sum_{i=1}^{N}\left(\frac{y_i}{\sigma^2}*x_i\ \right)}{I*\frac{1}{\gamma^2}+\frac{1}{\sigma^2}*\sum_{i=1}^{N}\left(x_i\ *x_i^T\right)}$ ##

先轉化為矩陣比較好表示

$$\frac{\partial L}{\partial\beta}=0\Rightarrow\left(-2X^Ty+2X^TX\beta\right)+\frac{\sigma}{\gamma^2}2\beta=0$$

$$\Rightarrow\beta=\left(X^TX+\frac{\sigma^2}{\gamma^2}I\right)^{-1}X^Ty.$$

(c)

這邊我分點描述:

1.最小平方估計是利用最小化 L2-Norm 求取參數β。

2.最大似然估計則是利用機率的觀點，求取一組參數β使 P(D|β)機率最大。

3.當觀測值來自指數族且滿足輕度條件時，最小平方估計和最大似然估計是相同的。

4.最大似然估計(MLE)是求參數β, 使似然函數 P(D|β)最大。而最大後驗機率估計(MAP)則是想求β使 P(D|β)P(β)最大。求得的β不單單讓似然函數大，β自己出現的先驗概率也盡可能的大。這裡有點神似之後上課應該會上到的正則化概念(懲罰)，只不過一般正則化是使用加法，而這邊使用了乘法，附帶提一點:正則化是避免模型過度擬合(Overfitting)的方法。

5.最大似然估計(MLE)認為參數本身的機率分布是均勻的(Uniform)，即其機率會是一個常數。

2. (c)

$$\text{ML}: \hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i^T \beta)^2$$

$$\text{MAP}: \hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i^T \beta)^2 + \frac{\sigma^2}{\tau^2} \sum_{i=1}^{n} \beta_i^2$$

least square sol.

$\Rightarrow$ The prior distribution acts as a regularier in MAP estimation

3.

3. (10%+10%+8%) Suppose that there were 200 coupons for each of the discount percentages 5%, 10%, 15%, 20%, and 30% (i.e., the input values x_i, for $i = 1, \ldots, 1000$), and that the number of coupons redeemed for the above cases are 31, 52, 68, 101, and 144, respectively (i.e., the number of responses in each case that yields $y_i = 1$).

(a) Fit a simple linear regression to the observed proportions 31/200, 52/200, 68/200, 101/200, and 144/200. List the fitted values for the above discount values. According to this regression, at what price reduction will you get a 25% redemption rate?

(b) Repeat (a) using logistic regression.

(c) Repeat (a) using linear discriminant analysis. (Hint: You will need to use the estimated normal distribution to compute $\Pr(Y = 1|X = x)$.)

(a) ① 依題目要求，分別索取的佔比為 31/200, 52/200, 68/200, 101/200, 144/200 數量的樣本，並把樣本分為一了 $X = 396\times 1$，$y = 396\times 1$ 的矩陣。

$$\begin{bmatrix} 1 & 0.05 \\ \vdots & \vdots \\ 1 & \vdots \end{bmatrix}_{396\times 2} \qquad \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

② $\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 2.281 \\ 0.032 \end{bmatrix} \begin{matrix} \leftarrow \beta_0 \\ \leftarrow \beta_1 \end{matrix}$ #

③ 25% ⇒ $0.09 + 1.71 \times 0.25 = 0.52$

④ 10% ⇒ 0.26，15% ⇒ 0.37，20% ⇒ 0.48，30% ⇒ 0.71

→ 各折扣所對應的兌換率 #

(b) ① 第一步如 (a) -① 所述。

② 迭代解的Beta: $B(k+1) = B(k) - \eta \cdot \nabla J(B(k))$, where. $\nabla J(B(k)) = -\sum_{i=1}^{n} \frac{(2y_i - 1)x_i}{1 + e^{(2y_i-1)B(k)^T x_i}}$

η = 定 = 0.01

③ 當 $B(k+1)$ 不再改變，即可停止。⇒ 即可估測 25% 的兌換率 #

(c)

令 $k=1 \Rightarrow y_i = 1$
$k=2 \Rightarrow y_i = 0$

$$\hat{\mu}_1 = \frac{1}{396} \times (0.05\times 31 + 0.1\times 52 + 0.15\times 68 + 0.2\times 101 + 0.3\times 144) = 0.202$$

$$\hat{\mu}_2 = \frac{1}{604} \times (0.05\times 169 + 0.1\times 148 + 0.15\times 132 + 0.2\times 99 + 0.3\times 56) = 0.132$$

$$\hat{\sigma} = \frac{1}{1000-2} \times \left(\sum_{i: y_i = 0} (x_i - 0.202)^2 + \sum_{i: y_i = 1} (x_i - 0.132)^2 \right) = 0.006$$

$$\hat{\pi}_1 = \frac{396}{1000}$$

$$\hat{\pi}_2 = \frac{604}{1000}$$

$$\delta_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

10% = $\delta_1(0.1)$ = 0.235
15% = $\delta_1(0.15)$ = 0.35
20% = $\delta_1(0.2)$ = 0.48
25% = $\delta_1(0.25)$ = 0.62
30% = $\delta_1(0.3)$ = 0.74

→ 各折扣所對應的兌換率 #

4.

7. It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

date: / /

4.

$R^2 = (TSS - RSS)/TSS$

" in the case of simple linear Regression"

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i)^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - (B_0 + B_1 x_i))^2$$

$B_0 = 0$

$$= \sum_{i=1}^{N} \Big(y_i - \underbrace{\Big(\frac{\sum_{j=1}^{n} x_j y_j}{\sum_{k=1}^{n} x_k^2}\Big)}_{B_1} x_i\Big)^2$$

This not zero

$$TSS - RSS = \sum_{i=1}^{n} (y_i)^2 - \sum_{i=1}^{N} \big(y_i^2 - 2 y_i B_1 x_i + \cancel{(B_1 x_i)^2}\big)$$

$$= \sum_{i=1}^{N} 2 x_i y_i B_1 - \frac{(\sum_{j=1}^{n} x_j y_j)^2}{\sum_{k=1}^{n} x_k^2} = \frac{(\sum_{j=1}^{n} x_j y_j)^2}{\sum_{k=1}^{n} x_k^2}$$

$$\frac{TSS - RSS}{TSS} = \frac{(\sum_{j=1}^{n} x_j y_j)^2}{\sum_{i=1}^{n} (y_i)^2 \times \sum_{k=1}^{n} (x_k)^2} = (Cor(X, Y))^2$$

#

5.

4. When the number of features p is large, there tends to be a deterioration in the performance of KNN and other *local* approaches that perform prediction using only observations that are *near* the test observation for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this curse.

curse of dimensionality

(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10 % of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$,

we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction?

(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10 % of the range of X_1 *and* within 10 % of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make the prediction?

(c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10 % of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

(d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10 % of the training observations. For $p = 1, 2,$ and 100, what is the length of each side of the hypercube? Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.

5.(a)

If $x \in [0.05, 0.95]$, the interval will be $[x-0.05, x+0.05] \Rightarrow 100\%$

If $x < 0.05$, the interval we use is $[0, x+0.05]$. $\Rightarrow (100x+5)\%$

If $x > 0.95$, the interval we use is $[x-0.05, 1]$. $\Rightarrow (105-100x)\%$

$\therefore$ The average fraction is

$$\int_0^{0.05}(100x+5)\,dx + \int_{0.05}^{0.95} 10\,dx + \int_{0.95}^{1}(105-100x)\,dx = 9.75\%$$

(b) $(9.75\%) \times (9.75\%) = 0.950625\%$

(c) $(9.75\%)^{100}$

(d) $p=1 \Rightarrow l=0.1$

$p=2 \Rightarrow l=(0.1)^{\frac{1}{2}}$

$p=100 \Rightarrow l=(0.1)^{\frac{1}{100}}$

(e) $p=1 \Rightarrow 0.1$

$p=2 \Rightarrow (0.1)^{\frac{1}{2}} \approx 0.316$

$p=100 \Rightarrow (0.1)^{\frac{1}{100}} \approx 0.977$

也就是說當我特徵數越高時，在使用固定(10%)的觀察量時，

我們更越需要包含每個特徵的所有範圍 #