/

CS 189 Introduction to Machine Learning
Fall 2017 HW9

This homework is due **Monday, October 30 at 10pm.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW[n] Write-Up"
2. Submit all code needed to reproduce your results, "HW[n] Code".
3. Submit your test set evaluation results, "HW[n] Test Set".

After you've submitted your homework, be sure to watch out for the self-grade form.

(a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None
Comments : none

(b) Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all sols are entirely in my words & that I have not looked at another stud's sols. I have credited all external sources in this write up

Problem # 2 Classification policy

a) The Bayes decision rule minimizes the risk

$$R(f(x)|x) = E[L(f(x), y)] = \sum_{i=1}^{c} L(f(x), y)\, P(Y=i|x)$$

Consider the given rule : 2 scenarios

1. given $P(Y=i|x) \geq P(Y=j|x)\ \forall j$ and $P(Y=i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$

if we choose class i, the risk is : ← policy

$$R_1(i|x) = \sum_{j=1}^{c} L(\underbrace{f(x)}_{i}, j)\, P(Y=j|x)$$

$$= \cancel{L(i,i)}^{\,0}\, P(Y=i|x) + \sum_{\substack{j=1 \\ j\neq i}}^{c} \underbrace{L(j,i)}_{\lambda_s} P(Y=j|x)$$

$$= \lambda_s \sum_{\substack{j=1 \\ j\neq i}}^{c} P(Y=j|x)$$

if we choose class $j' \neq i$ the risk is : ← non-policy

$$R_2(j'|x) = \sum_{j=1}^{c} L(f(x), j)\, P(Y=j|x)$$

$$= \lambda_s P(Y=i|x) + \lambda_s \sum_{\substack{j=1 \\ j\neq i,j'}}^{c} P(Y=j|x)$$

$$+ \cancel{L(j',j')}^{\,0}\, P(Y=j'|x)$$

$$= \lambda_s P(Y=i|x) + \lambda_s \sum_{\substack{j=1 \\ j\neq i,j'}}^{c} P(Y=j|x)$$

if we choose doubt, the risk is : ← non-policy

$$R_3(c+1|x) = \sum_{j=1}^{c} L(\overset{c+1}{f(x)}, j)\, P(Y=j|x)$$

$$= \lambda_r \underbrace{\sum_{j=1}^{c} P(Y=j|x)}_{1} = \lambda_r$$

now we compare R_1, R_2 & R_3:

$$R_1 = \lambda_s \sum_{\substack{j=1 \\ j\neq i}}^{c} P(Y=j|x) = \lambda_s P(Y=j'|x) + \sum_{\substack{j=1 \\ j\neq i,j'}}^{c} P(Y=j|x)$$

$$\left(\begin{array}{l}\text{since } P(Y=j|x) \leq P(Y=i|x) \\ \text{given } \forall j\end{array}\right) \qquad \leq \underbrace{\lambda_s P(Y=i|x) + \sum_{\substack{j=1 \\ j\neq i,j'}}^{c} P(Y=j|x)}_{R_2}$$

$$R_1 = \lambda_s \sum_{\substack{j=1 \\ j\neq i}}^{c} P(Y=j|x) + \lambda_s P(Y=i|x) - \lambda_s P(Y=i|x)$$

$$\text{"} = \lambda_s \underbrace{\sum_{j=1}^{c} P(Y=j|x)}_{1} - \lambda_s \underbrace{P(Y=i|x)}_{\geq 1-\frac{\lambda_r}{\lambda_s}}$$

$$\leq \lambda_s - \lambda_s\left(1-\frac{\lambda_r}{\lambda_s}\right) = \lambda_s - \lambda_s + \lambda_r = \lambda_r = R_3$$

i.e $R_1 \leq R_3$

Thus, R_1 is minimum

2. Given $P(Y=i|x) \geq P(Y=j|x) \ \forall j$ and $P(Y=i|x) < 1-\frac{\lambda_r}{\lambda_s}$

$$\Rightarrow \quad P(Y=j|x) < 1-\frac{\lambda_r}{\lambda_s} \quad \forall j$$

(we can always find the largest $P(Y=i|x)$ thus $P(Y=i|x) \geq P(Y=j|x) \ \forall j$ always holds)

If we choose doubt $\leftarrow$ policy

$$R_1 = \lambda_r$$

If we choose i or j' $\leftarrow$ non-policy

$$R_2 = \lambda_s \sum_{\substack{j=1 \\ j \neq i \text{ or } j'}}^{c} P(Y=j|x)$$

Since
$$P(Y=j'|x) < 1 - \frac{\lambda_r}{\lambda_s} \quad \forall j'$$

$$\lambda_s P(Y=j'|x) < \lambda_s - \lambda_r$$

$$\lambda_r < \underbrace{\lambda_s \underbrace{\sum_{j=1}^{c} P(Y=j|x)}_{1} - \lambda_s P(Y=j'|x)}_{\lambda_s \sum_{\substack{j=1 \\ j \neq i \text{ or } j'}}^{c} P(Y=j|x)}$$

$$\Rightarrow \quad R_1 < R_2$$

Thus R_1 is minimum

In all scenario, the given policy has the minimum risk, thus the policy is the Bayes decision rule b.) If $\lambda_r = 0$, we should alway choose doubt, since the risk for choosing doubt is $R = 0$ minimum

Intuitively, it is correct because we are not charged for doubt

If $\lambda_r > \lambda_s$, we should never choose doubt because

the risk $R = \lambda_r > \lambda_s > \underbrace{\lambda_s \underbrace{\sum_{j=1}^{c} P(Y=j|x)}_{1} - \lambda_s P(Y=j'|x)}_{R' \text{ : cost for choosing arbitrary } j'}$

Thus the risk for choosing doubt is always larger than the risk for choosing arbitrary j'

Intuitively, it is correct because the cost for doubt is too large.

Problem #3 LDA & CCA

a.) The Bayes decison rule minimize the risk

$$R(f(X)|X) = E[L(f(X), X)]$$
$$= \sum_{\ell=1}^{n} P(L=\ell|X)\, L(f(X), \ell)$$

Assume the loss function:

$$L(f(X), \ell) = \begin{cases} 0 & \text{if } f(X) = \ell \\ 1 & \text{if } f(X) \neq \ell \end{cases}$$

Thus to minimize the risk, we have to choose the label ℓ corresponding to maximum $P(L=\ell|X)$

i.e
$$f(X) = \underset{\ell}{\operatorname{argmax}}\, P(L=\ell|X) \quad \text{MAP}$$
$$= \underset{\ell}{\operatorname{argmax}} \frac{P(X|L=\ell)P(L=\ell)}{\int_\ell P(X|L=\ell)P(L=\ell)\,d\ell} \quad \text{const}$$
$$= \underset{\ell}{\operatorname{argmax}} \log P(X|L=\ell) + \underbrace{\log P(L=\ell)}_{\log \pi_\ell}$$

$$P(X|L=\ell) = \frac{1}{(\sqrt{2\pi})^d |\Sigma|^{1/2}} \exp\left(\frac{-1}{2}(X-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell)\right)$$

$$\Rightarrow f(X) = \underset{\ell}{\operatorname{argmin}}\left(\log(\sqrt{2\pi})^d|\Sigma|^{1/2} + \frac{1}{2}(X-\mu_\ell)\Sigma^{-1}(x-\mu_\ell) - \log\pi_\ell\right)$$

since we have only 2 label $\{1, 2\}$ so we will choose label 1 if

$$\log(\sqrt{2\pi})^d|\Sigma|^{1/2} + \frac{1}{2}(X-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \log\pi_1$$
$$\leq \log(\sqrt{2\pi})^d|\Sigma|^{1/2} + \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) - \log\pi_2$$

$$X^T\Sigma^{-1}X - 2\mu_1^T\Sigma^{-1}X + \mu_1^T\Sigma^{-1}\mu_1 - 2\log\pi_1 \leq X^T\Sigma^{-1}X - 2\mu_2^T\Sigma^{-1}X$$
$$+ \mu_2^T\Sigma^{-1}\mu_2 - 2\log\pi_2$$

$$\underbrace{2(\mu_1^T - \mu_2^T)\Sigma^{-1}X + 2\log\frac{\pi_1}{\pi_2} - \mu_1\Sigma^{-1}\mu_1 + \mu_2^T\Sigma^{-1}\mu_2}_{f(x)} \geq 0$$

MAP $\nearrow \rightarrow$ $f(x)$ linear in X

In summary, if $f(x) \geq 0$ choose label 1, else label 2

For MLE, we try to maximize $P(X \mid L=l)$. Similar to above but w/o $P(L=l)$ term. Thus

$$\arg\max_l P(X \mid L=l) = \arg\min_l \left(\log(\sqrt{2\pi})^d |\Sigma|^{1/2} + \frac{1}{2}(X-\mu_l)^T\Sigma^{-1}(X-\mu_l) \right)$$

The final expression is

$$\underbrace{2(\mu_1^T - \mu_2^T)\Sigma^{-1}X - \mu_1^T\Sigma^{-1}\mu_1 + \mu_2^T\Sigma^{-1}\mu_2}_{f(x)} \geq 0$$

MLE $\nearrow \longrightarrow$ $f(x)$

if $f(x) \geq 0$ choose label 1, else label 2.

Two decision rules are the same if $\log\frac{\pi_1}{\pi_2} = 0$

i.e $\frac{\pi_1}{\pi_2} = 1$ or $\pi_1 = \pi_2$

(b) $\Sigma_{XX} = E[(X-\mu_X)(X-\mu_X)^T]$

denote $l_1 = 1$, $l_2 = 2$

$$\mu_X = E[X] = \int_X X P(X)\, dX = \int_X X (\underbrace{P(X|l_1)P(l_1)}_{\pi_1} + P(X|l_2) \underbrace{P(l_2)}_{\pi_2})\, dX$$

$$= \pi_1 \int_X X P(X|l_1)\, dX + \pi_2 \int X P(X|l_2)\, dX$$

$$= \pi_1 E[X|l_1] + \pi_2 E[X|l_2]$$

$$= \pi_1 \mu_1 + \pi_2 \mu_2$$

if $\pi_1 = \pi_2 = \frac{1}{2} \Rightarrow \mu_X = \frac{\mu_1 + \mu_2}{2}$

Similar to $E[XX^T] = \int_X XX^T P(X)\, dX$

$$= \pi_1 E[XX^T|l_1] + \pi_2 E[XX^T|l_2]$$

$$\Sigma_{XX} = E[XX^T - X\mu_X^T - \mu_X X^T + \mu_X \mu_X^T]$$

$$= E[XX^T] - \underbrace{E[X\mu_X^T]}_{E[X]\mu_X^T} - \underbrace{E[\mu_X X^T]}_{=\mu_X E[X^T]} + \underbrace{E[\mu_X \mu_X^T]}_{=\mu_X\mu_X^T}$$

$$= E[XX^T] - \mu_X \mu_X^T$$

$$= \pi_1 E[XX^T|l_1] + \pi_2 E[XX^T|l_2] - \mu_X \mu_X^T$$

$$\Sigma = V[X|l_1] = E[XX^T|l_1] - \mu_1\mu_1^T$$

$$= V[X|l_2] = E[XX^T|l_2] - \mu_2\mu_2^T$$

$$\Rightarrow E[XX^T|l_1] = \Sigma + \mu_1\mu_1^T$$

$$E[XX^T|l_2] = \Sigma + \mu_2\mu_2^T$$

$$\Sigma_{xx} = (\pi_1 + \pi_2)\,\Sigma + \pi_1 \mu_1 \mu_1^T + \pi_2 \mu_2 \mu_2^T - (\pi_1 \mu_1 + \pi_2 \mu_2)(\pi_1 \mu_1 + \pi_2 \mu_2)^T$$

if $\pi_1 = \pi_2 = \frac{1}{2}$

$$\Sigma_{xx} = \Sigma + \frac{1}{2}\mu_1 \mu_1^T + \frac{1}{2}\mu_2 \mu_2^T - \frac{\mu_1 \mu_1^T + \mu_1 \mu_2^T + \mu_2 \mu_1^T + \mu_2 \mu_2^T}{4}$$

$$= \Sigma + \frac{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}{4}$$

$$\Sigma_{YY} = E[(Y - \mu_y)(Y - \mu_y)^T]$$

$$Y = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ given } L = l_1 \Rightarrow E[Y \mid l_1] = \mu_{y_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$V[Y \mid l_1] = E[YY^T \mid l_1] - \mu_{y_1}\mu_{y_1}^T = 0$$

$$\Rightarrow E[YY^T \mid l_1] = \mu_{y_1}\mu_{y_1}^T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$Y = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \text{ given } L = l_2 \Rightarrow E[Y \mid l_2] = \mu_{y_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$E[YY^T \mid l_2] = \mu_{y_2}\mu_{y_2}^T = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$E[YY^T] = \pi_{y_1} E[YY^T \mid l_1] + \pi_{y_2} E[YY^T \mid l_2]$$

$$= \begin{bmatrix} \pi_{y_1} & 0 \\ 0 & \pi_{y_2} \end{bmatrix}$$

assume $\pi_{y_1} = \pi_{y_2} = \frac{1}{2} \Rightarrow E[YY^T] = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$

$$\mu_y = E[Y] = \pi_{y_1} E[Y \mid l_1] + \pi_{y_2} E[Y \mid l_2]$$

$$= \begin{bmatrix} \pi_{y_1} \\ \pi_{y_2} \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

$$\Sigma_{YY} = E[YY^T] - \mu_y \mu_y^T = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} - \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} [1/2 \;\; 1/2]$$

$$\Sigma_{YY} = \begin{bmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{bmatrix}$$

$$\Sigma_{XY} = E\left[(X-\mu_x)(Y-\mu_y)^T\right] = E[XY^T] - \mu_x \mu_y^T$$

(d) From part (b)

$$\Sigma_{XX} = (\pi_1 + \pi_2)\Sigma + \pi_1 \mu_1 \mu_1^T + \pi_2 \mu_2 \mu_2^T$$

$$- (\pi_1 \mu_1 + \pi_2 \mu_2)(\pi_1 \mu_1 + \pi_2 \mu_2)^T$$

$$\Sigma_{YY} = \begin{bmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{bmatrix}$$

(e) Procedure:

1.) Calculate $\mu_X = \frac{1}{n} \sum_{i=1}^{n} x_i$

2.) Calculate $\Sigma_{XX} = E[(X-\mu_X)(X-\mu_X)^T]$

we know $\Sigma_{YY} = \begin{bmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{bmatrix}$

$\mu_Y = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$

Calculate

$$\Sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)^T]$$

3.) From Σ_{XX} Σ_{YY} Σ_{XY}

Apply question (c) to determine u^*

4.) $\rightarrow$ predict x_{test} base on u^*

Problem #4

a.) See code attach

b.) See code attached

Description:

the second-order & third order has the best performance though the initial error is quite large.

The linear order has similar performance with neural network.

The generative model has largest error

Strengths and weaknesses

1. Generative : large error, no need location of sensors

2.) linear / second / third order : pretty simple to implement and quite stable to run but need more # of sample (n not too small)

3.) Neural network : unstable, complicated model and need a lot of care c.) See code attached

The best $l = 150$

d.) $\underbrace{(I+1)l}_{\text{input \#}} + \underbrace{(l+1)l + \dots +}_{k-1 \text{ time}} \underbrace{(l+1)*O}_{\text{output \#}}$

$= (I+1)l + (k-1)l(l+1) + (l+1)*O$

$I = 7 \qquad O = 2$

$\Rightarrow \quad 8l + (k-1)(l^2+1) + 2(l+1) = 10000$

quadratic eqn interms of l

k best. is $k = 2$

See code & plot attached

e.) Choose $k = 2$

$l = 150$

error is ~15 ← small

See code attached

Problem # 5.

See code attached

$\% = 92.18\%$

Problem #6

Solve prob. 36 for QDA using MAP. $\Sigma_1 \neq \Sigma_2$.
and prove that the decision function $f(x)$ is quadratic

Solution

The decision function is

$$f(x) = \arg\max_{\ell} P(L=\ell \mid x) \quad \swarrow \text{MAP}$$

$$= \arg\max_{\ell} \frac{P(X \mid L=\ell)\, P(L=\ell)}{\int_{\ell} P(X \mid L=\ell)\, P(L=\ell)\, d\ell}$$

$$= \arg\max_{\ell} \log P(X \mid L=\ell) + \log P(L=\ell)$$

$$\Rightarrow f(x) = \arg\min_{\ell} \underbrace{\left(\log(\sqrt{2\pi})^{d} \, |\Sigma_\ell|^{1/2} + \frac{1}{2}(X-\mu_\ell)\, \Sigma_\ell^{-1} (x-\mu_\ell) - \log \pi_\ell \right)}_{f^*(x)}$$

We will choose label 1 if

$$f^*(x \mid L = \ell_1) < f^*(X \mid L = \ell_2)$$

$$\log(\sqrt{2\pi})^d|\Sigma_1|^{1/2} + \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \log\pi_1$$

$$\le \log(\sqrt{2\pi})^d|\Sigma_2|^{1/2} + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) - \log\pi_2$$

$$x^T\Sigma_1^{-1}x - 2\mu_1^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1 - 2\log\pi_1 + \log(2\sqrt{\pi})^d|\Sigma_1|^{1/2}$$

$$\le x^T\Sigma_2^{-1}x - 2\mu_2^T\Sigma_2^{-1}x + \mu_2^T\Sigma_2^{-1}\mu_2 - 2\log\pi_2 + \log(\sqrt{2\pi})^d|\Sigma_2|^{1/2}$$

$$\underbrace{x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x - 2(\mu_1^T\Sigma_1 - \mu_2^T\Sigma_2)x - 2\log\frac{\pi_1}{\pi_2} + \mu_1^T\Sigma_1^{-1}\mu_1 + \mu_2^T\Sigma_2^{-1}\mu_2 + \log\frac{|\Sigma_1|^{1/2}}{|\Sigma_2|^{1/2}} \le 0}_{\text{decision rule}}$$

and it is quadratic.