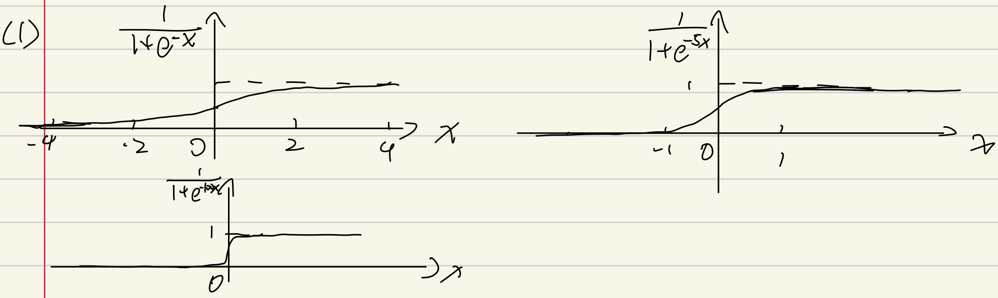
Homework 2

仲泽为 12232814

Due: Mar. 27th Wed 23:59

Q1 (1)



Large weights can lead to overfitting in logistic regression because data around 0 would be clarified very clearly whether 0 or 1, and the probability of one data to be in which class is not represented well in overfitting model.

(2) $\max\limits_{w_0,\dots,w_d} \prod\limits_{i=1}^{n} P(Y_i \mid X_i, w_0, \dots, w_d)\, P(w_0, \dots, w_d)$, with prior of $N(0, I)$

$$\mathcal{L}(\underline{w}) = -\log \prod_{i=1}^{n} P(y^i \mid x^i; \underline{w}) - \log P(\underline{w})$$

Since we can assume all w_j's are independent

$$P(\underline{w}) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w_j^2}{2}\right)$$

$$\mathcal{L}(\underline{w}) = -\sum_{i=1}^{n} \log P(y^i \mid x^i; \underline{w}) - \sum_{j=0}^{d} \log \frac{1}{\sqrt{2\pi}} + \sum_{j=0}^{d} \frac{w_j^2}{2}$$

Then we can have the gradient:

$$\frac{\partial \mathcal{L}(w)}{\partial w_j} = w_j - \sum_{i=1}^{n} x^i [P(y^i / x^i; \underline{w}) - y^i]$$

The gradient descent update rule thus is

$$w_j \leftarrow w_j + \alpha \frac{\partial \mathcal{L}(w)}{\partial w_j}$$

$$w_j \leftarrow (1+\alpha) w_j - \sum_{i=1}^{n} x^i [P(y^i \mid x^i; \underline{w}) - y^i]$$

The prior constrict the weight $\underline{w}$ to not being to large so that dataset (d features n entries) 像识别数字

$\{(x^i, y^i) \mid i=1,\dots,n\}$ $x^i \in \mathbb{R}^d$ y^i is one hot

Q2 (1) Given $p(y_c^i=1 \mid x^i; w) = \frac{\exp(w_c^T x^i)}{\Sigma_{c'} \exp(w_{c'}^T x^i)}$

Add the same value for every weight won't affect the value of the softmax equation. Thus, we can add $(-w_k)$ for every weight

$$p(y_c^i=1 \mid x^i; w) = \frac{\exp[(w_c - w_k)^T x^i]}{\Sigma_{c'} \exp[(w_{c'} - w_k)^T x^i]}$$

Note that $\exp[(w_k - w_k)^T x^i] = \exp(0^T x^i) = e^0 = 1$

$$\Sigma_{c'} \exp[(w_{c'} - w_k)^T x^i] = \sum_{c'=1}^{k-1} \exp[(w_{c'} - w_k)^T x^i] + 1$$

Let $w_{c'}$ denotes $w_{c'} - w_k$

$$p(y_c^i=1 \mid x^i; w) = \begin{cases} \dfrac{\exp(w_c^T x^i)}{1+\sum_{c'=1}^{k-1} \exp(w_{c'}^T x^i)}, & c<k \\ \dfrac{1}{1+\sum_{c'=1}^{k-1} \exp(w_{c'}^T x^i)}, & c=k \end{cases}$$

---

(2) $\mathcal{L}(w) = -\ln \prod_{j=1}^{n} p(y_c^j \mid x^j, w)$ [c is the true class of x^j]

We can note $y_c = 1$ when true class and 0 otherwise

$\sum_{c=1}^{k} y_c^i = 1$

$$\begin{aligned} \mathcal{L}(w) &= -\sum_{j=1}^{n} \sum_{c=1}^{k} y_c^j \ln[p(y_c^j \mid x^j, w)] \\ &= -\sum_{j=1}^{n} \sum_{c=1}^{k} \left[y_c^j \ln \frac{\exp(w_c^T x^j)}{\Sigma_{c'} \exp(w_{c'}^T x^j)} \right] \\ &= -\sum_{j=1}^{n} \sum_{c=1}^{k} \left[y_c^j (w_c^T x^j) - y_c^j \ln\left(\Sigma_{c'} \exp(w_{c'}^T x^j)\right) \right] \end{aligned}$$

---

next page

Q2 (3) $\nabla_{w_c} \sum_{c=1}^{k} y_c^j \ln\left(\sum_{c'} \exp(w_{c'}^T x^j)\right)$

$= \left(\sum_{c=1}^{k} y_c^j\right) \frac{1}{\sum_{c'} \exp(w_{c'}^T x^j)} \cdot \exp(w_c^T x^j) \cdot x^j]$ ($w_{c'} \neq w_c$ is ignored)

$= \frac{\exp(w_c^T x^j)}{\sum_{c'} \exp(w_{c'}^T x^j)} x^j$ for a ~~fixed~~ fixed j

---

(4) $\nabla_{w_c} \mathcal{L}(w) = \sum_{j=1}^{n} y_c^j x^j - \sum_{j=1}^{n} \frac{\exp(w_c^T x^j)}{\sum_{c'} \exp(w_{c'}^T x^j)} x^j$

$= \sum_{j=1}^{n} y_c^j x^j - \sum_{j=1}^{n} x^j \cdot p(y_c^j = 1 \mid x^j; w)$

$= \sum_{j=1}^{n} x^j \left[y_c^j - p(y_c^j = 1 \mid x^j; w) \right]$

---

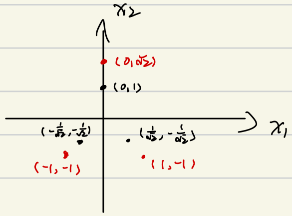
(5) The gradient ascent rule is

$w_c^{t+1} \leftarrow w_c^t + \eta \nabla_{w_c} \mathcal{L}(w)$

where $\nabla_{w_c} \mathcal{L}(w) = \sum_{j=1}^{n} x^j (y_c^j - p(y_c^j = 1 \mid x^j; w^t))$

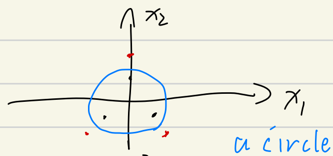
$w_c^{t+1} \leftarrow w_c^t + \eta \cdot \sum_{j=1}^{n} x^j \left(y_c^j - p(y_c^j = 1 \mid x^j; w^t) \right)$

Q3 (1) I couldn't, since these points are not linear separable in 2-dimensional space



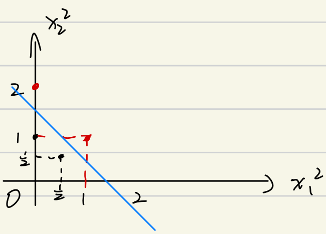
If use kernel, we can have the SVM $x_1^2 + x_2^2 = \left(\frac{\sqrt{2}+1}{2}\right)^2$

which is roughly



(2) By expanding $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ to $\underline{x} = \begin{bmatrix} x_1^2 \\ x_2^2 \end{bmatrix}$, rewrite the point value

Class -1: $\{(0,1), (\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2})\}$, Class $+1$: $\{(0.2), (1,1), (1,1)\}$



The SVM: $x_1^2 + x_2^2 = \frac{3}{2}$

For the new point $(-\frac{1}{2}, \sqrt{2}) \Rightarrow \frac{1}{4} + 2 = \frac{9}{4} > \frac{3}{2}$

The point $(-\frac{1}{2}, \sqrt{2})$ belongs to Class $+1$

Q4 Loss function: $L_\epsilon(\underline{x}, y, f) = |y - f(\underline{x})|_\epsilon = \max\{0, |y - f(\underline{x})| - \epsilon\}$

The SVR cost function: $\frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{n} L_\epsilon(\underline{x}_i, y_i, f)$

$f(\underline{x}) = \underline{w}^T\underline{x}$, $C, \epsilon > 0$

---

(1) Introduce $\xi, \xi^* \geq 0$

SVR cost function: $\frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{n}[L_\epsilon(\underline{x}_i, y_i, f) + \xi_i + \xi_i^*]$

By writing the Loss function part into constraints, we obtain the quadratic problem:

$$\min_{\underline{w}, \xi, \xi^*} \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\begin{aligned} \text{s.t.} \quad & y_i - \underline{x}_i^T\underline{w} \leq \epsilon + \xi_i^* , && \forall i \\ & \underline{x}_i^T\underline{w} - y_i \leq \epsilon + \xi_i , && \forall i \\ & \xi_i , \xi_i^* \geq 0 , && \forall i \end{aligned}$$

---

(2) The Langrangian function:

$$\begin{aligned} L(\underline{w}, \xi, \xi^*, \underline{\mu}, \underline{\mu}^*, \underline{\alpha}, \underline{\alpha}^*) = & \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ & + \sum_{i=1}^{n}\alpha_i^*(y_i - \underline{x}_i^T\underline{w} - \epsilon - \xi_i^*) + \sum_{i=1}^{n}\alpha_i(\underline{x}_i^T\underline{w} - y_i - \epsilon - \xi_i) \\ & - \sum_{i=1}^{n}(\mu_i\xi_i + \mu_i^*\xi_i^*) \end{aligned}$$

---

(3) Stationarity: $\frac{\partial L}{\partial \underline{w}} = \underline{w} - \sum_{i=1}^{n}\alpha_i^*\underline{x}_i + \sum_{i=1}^{n}\alpha_i\underline{x}_i = 0$

$$\Rightarrow \underline{w} = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\underline{x}_i$$

$$\frac{\partial L}{\partial \underline{\xi}} = \sum_{i=1}^{n}(C - \alpha_i - \mu_i) = 0$$

$$\Rightarrow \alpha_i + \mu_i = C , \forall i$$

$$\frac{\partial L}{\partial \underline{\xi}^*} = \sum_{i=1}^{n}(C - \alpha_i^* - \mu_i^*) = 0$$

$$\Rightarrow \alpha_i^* + \mu_i^* = C , \forall i$$

Q4 (3) Complementarity slackness:

$$\alpha_i(\underline{x}_i^T \underline{w} - y_i - \epsilon - \xi_i) = 0 \qquad \forall i$$

$$\alpha_i^*(y_i - \underline{x}_i^T \underline{w} - \epsilon - \xi_i^*) = 0 \qquad \forall i$$

$$\mu_i \xi_i = 0, \quad \mu_i^* \xi_i^* = 0 \qquad \forall i$$

Feasibility:

$$\alpha_i, \alpha_i^*, \mu_i, \mu_i^* \geq 0 \qquad \forall i$$

$$\underline{x}_i^T \underline{w} - y_i - \epsilon - \xi_i \leq 0, \qquad y_i - \underline{x}_i^T \underline{w} - \epsilon - \xi_i^* \leq 0, \quad \forall i$$

Replace all stationarity conditions into the Lagrangian functions to eliminate primal variables

$$L(\underline{\mu}, \underline{\mu}^*, \underline{\alpha}, \underline{\alpha}^*) = \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\underline{x}_i^T \underline{x}_j$$

$$+ \sum_{i=1}^n \sum_{j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j^* - \alpha_j)\underline{x}_i^T \underline{x}_j$$

$$+ \sum_{i=1}^n (C - \alpha_i^* - \mu_i^*)\xi_i^* + \sum_{i=1}^n (C - \alpha_i - \mu_i)\xi_i$$

$$+ \sum_{i=1}^n [\alpha_i^*(y_i - \epsilon) + \alpha_i(-y_i - \epsilon)]$$

$$= \sum_{i=1}^n y_i(\alpha_i^* - \alpha_i) - \epsilon \sum_{i=1}^n (\alpha_i^* + \alpha_i)$$

$$- \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\underline{x}_i^T \underline{x}_j$$

next page

Q4 (3) Thus, the dual form is

$$\max_{\underline{\alpha}^*, \underline{\alpha}} \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) - \epsilon \sum_{i=1}^{n} (\alpha_i^* + \alpha_i) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\, \underline{x}_i^T \underline{x}_j$$

$$\text{s.t.} \quad \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0$$

$$0 \le \alpha_i \le C, \quad 0 \le \alpha_i^* \le C \quad \forall i$$

---

(4) Yes, since part $-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\, \underline{x}_i^T \underline{x}_j$ contains variables in quadratic form and there is no higher times in the objective function. Besides, the constraints are linear. In all, the dual problem satisfies the definition of quadratic problem. Thus, we can use quadratic optimization solvers to solve the problem.

---

(5) Define $M = \{i \mid 0 < \alpha_i^* < C\}$ $N = \{j \mid 0 < \alpha_j < C\}$

Support vectors are $\underline{x}_i$, $i \in M$ and $\underline{x}_j$, $j \in N$

---

(6) $\underline{w} = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)\, \underline{x}_i$

For sample X

$$\hat{y}_j = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)\, \underline{x}_i^T \underline{x}_j \qquad \forall j = 1, \dots, n$$

---

(7) Yes, we can replace $\underline{x}_i^T \underline{x}_j$ by $\phi(\underline{x}_i, \underline{x}_j)$

next page

Q4 (8) $\epsilon \downarrow \Rightarrow$ width of the boundary $\downarrow \Rightarrow$ loss function value $\uparrow$ (total loss & cost)

$\Rightarrow$ may lead to overfitting

$\epsilon \uparrow \Rightarrow$ width of the boundary $\uparrow \Rightarrow$ Total loss $\downarrow$

$\Rightarrow$ may lead to underfitting

---

(9) $C \downarrow \Rightarrow$ penalty on loss $\downarrow \Rightarrow$ cost $\downarrow$

$\Rightarrow$ may lead to underfitting

$C \uparrow \Rightarrow$ penalty on loss $\uparrow \Rightarrow$ cost $\uparrow$

$\Rightarrow$ may lead to overfitting

Q5 Given D with $|D| = 15$,

There are 2 classes: C_1: Yes, C_2: No, $|C_1| = 9$, $|C_2| = 6$

D

The first decision

$$H^{1}(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{3} \approx 0.9710$$

(a) For attribute age: $\overset{n=3}{}$ a_1: young: $|D_1| = 5$; a_2: middle: $|D_2| = 5$; a_3: old: $|D_3| = 5$

$|D_{ik}|$: $|D_{11}| = 2$, $|D_{12}| = 3$; $|D_{21}| = 3$, $|D_{22}| = 2$; $|D_{31}| = 4$, $|D_{32}| = 1$

$$H^{(1)}(D|age) = \sum_{i=1}^{n}\frac{|D_i|}{|D|}H^{(2)}(D_i) = -\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$$

$$= \sum_{i=1}^{3}\frac{|D_i|}{|D|}\left[-\frac{|D_{i1}|}{|D_i|}\log_2\frac{|D_{i1}|}{|D_i|} - \frac{|D_{i2}|}{|D_i|}\log_2\frac{|D_{i2}|}{|D_i|}\right]$$

$$= \frac{|D_1|}{|D|}\left(-\frac{|D_{11}|}{|D_1|}\log_2\frac{|D_{11}|}{|D_1|} - \frac{|D_{12}|}{|D_1|}\log_2\frac{|D_{12}|}{|D_1|}\right) + \frac{|D_2|}{|D|}\left(-\frac{|D_{21}|}{|D_2|}\log_2\frac{|D_{21}|}{|D_2|} - \frac{|D_{22}|}{|D_2|}\log_2\frac{|D_{22}|}{|D_2|}\right)$$

$$+ \frac{|D_3|}{|D|}\left(-\frac{|D_{31}|}{|D_3|}\log_2\frac{|D_{31}|}{|D_3|} - \frac{|D_{32}|}{|D_3|}\log_2\frac{|D_{32}|}{|D_3|}\right)$$

$$\approx \frac{1}{3}(0.9710 + 0.9710 + 0.7219) \approx 0.8880$$

We also note that under attribute age:

$H^{(2)}(D_1) \approx 0.9710$ $\quad H^{(2)}(D_2) \approx 0.9710$ $\quad H^{(2)}(D_3) \approx 0.7219$

$$g(D, age) = H(D) - H(D|age) \approx 0.0830$$

$$H_{age}(D) = -\sum_{j=1}^{3}\frac{|D_j|}{|D|}\log_2\frac{|D_j|}{|D|} = -\left(\frac{1}{3}\log_2\frac{1}{3}\right)\times 3 \approx 1.5850$$

$$g_R(D, age) = \frac{g(D, age)}{H_{age}(D)} \approx \frac{0.0379}{1.5850} \approx 0.0524$$

(b) For attribute work: $n=2$. a_1: Yes, $|D_1| = 5$; a_2: No, $|D_2| = 10$

$|D_{ik}|$: $|D_{11}| = 5$, $|D_{12}| = 0$; $|D_{21}| = 4$, $|D_{22}| = 6$

$$H^{(1)}(D, work) = \frac{1}{3}H^{(2)}(D_1) + \frac{2}{3}H^{(2)}(D_2) = \frac{1}{3}\left(-\frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5}\right) + \frac{2}{3}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)$$

$$\approx \frac{1}{3}\times 0 + \frac{2}{3}\times 0.9710 \approx 0.6473$$

Under attribute work $H^{(2)}(D_1) = 0$, $H^{(2)}(D_2) \approx 0.9710$

Q5

$$g(D, work) = H(D) - H(D|work) \approx 0.3237$$
$$H_{work}(D) = -\sum_{j=1}^{2} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx 0.9183$$
$$g_R(D, work) = \frac{g(D, work)}{H_{work}(D)} \approx 0.3525$$

(c) For attribute house: $n=2$ a_1: Yes, $|D_1| = 6$; a_2: No, $|D_2| = 9$

$|D_{ik}|$: $|D_{11}| = 6$, $|D_{12}| = 0$; $|D_{21}| = 3$, $|D_{22}| = 6$

$$H^{(2)}(D|house) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H^{(2)}(D_i) = \frac{2}{5}\left(-\frac{6}{6}\log_2\frac{6}{6} - \frac{0}{6}\log_2\frac{0}{6}\right) + \frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)$$
$$\approx \frac{2}{5}\times 0 + \frac{3}{5}\times 0.9183 \approx 0.5510$$

Under attribute house, $H^{(2)}(D_1) = 0$, $H^{(2)}(D_2) \approx 0.5510$

$$g(D, house) = H(D) - H(D|house) \approx 0.4200$$
$$H_{house}(D) = -\sum_{j=1}^{2} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} \approx 0.9710$$
$$g_R(D, house) = \frac{g(D, house)}{H_{house}(D)} \approx 0.4325$$

(d) For attribute credit: $n=3$ a_1: normal, $|D_1| = 5$; a_2: good, $|D_2| = 6$; a_3: excellent, $|D_3| = 4$

$|D_{ik}|$: $|D_{11}| = 1$, $|D_{12}| = 4$; $|D_{21}| = 4$, $|D_{22}| = 2$; $|D_{31}| = 4$, $|D_{32}| = 0$

$$H^{(2)}(D|credit) = \sum_{i=1}^{3} \frac{|D_i|}{|D|} H^{(2)}(D_i) = \frac{1}{3}\left(-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}\right) + \frac{2}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{4}{15}\times 0$$
$$\approx \frac{1}{3}\times 0.7219 + \frac{2}{5}\times 0.9183 + \frac{4}{15}\times 0 \approx 0.6080$$

Under credit, $H^{(2)}(D_1) \approx 0.7219$, $H^{(2)}(D_2) \approx 0.9183$, $H^{(2)}(D_3) = 0$

$$g(D, credit) = H(D) - H(D|credit) \approx 0.3630$$
$$H_{credit}(D) = -\sum_{j=1}^{3} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{5}\log\frac{2}{5} - \frac{4}{15}\log\frac{4}{15} \approx 1.5656$$
$$g_R(D, credit) = \frac{g(D, credit)}{H_{credit}(D)} \approx 0.2319$$

Q5

$0.0524 < 0.2319 < 0.3525 < 0.4325$

$g_R(D, age) < g_R(D, credit) < g_R(D, work) < g_R(D, house)$

Chose house for the first decision attribute

$D_1 : \{4, 8, 9, 10, 11, 12\}$, $D_2 : \{1, 2, 3, 5, 6, 7, 13, 14, 15\}$

Note that data in D_1 purely belong to class 1 : Yes, we don't need to create branch for this node

For D_2, we have $H^{(2)}(D_2) \approx 0.9183$

---

② The second decision

Let D denotes the former D_2 :

$|D| = 9$, $|C_1| = 3$, $|C_2| = 6$

$H(D) \approx 0.9183$

ID	Age	Work	Credit	Class
1	y	n	Nor	N
2	y	n	G	N
3	y	y	G	Y
5	y	n	Nor	N
6	m	n	Nor	N
7	m	n	G	N
13	o	y	G	Y
14	o	y	Ex	Y
15	o	n	Nor	N

(a) For attribute age : $n=3$, a_1: young, $|D_1| = 4$; a_2 : middle, $|D_2| = 2$; a_3 : old, $|D_3| = 3$

$|D_{ik}|$: $|D_{11}| = 1$, $|D_{12}| = 3$; $|D_{21}| = 0$, $|D_{22}| = 2$; $|D_{31}| = 2$, $|D_{32}| = 1$

$H(D|age) = \sum_{i=1}^{3} \frac{|D_i|}{|D|} H(D_i) = \frac{4}{9}\left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}\right) + \frac{2}{9} \times 0 + \frac{1}{3}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$

$\approx \frac{4}{9} \times 0.8113 + \frac{2}{9} \times 0 + \frac{1}{3} \times 0.9183 \approx 0.6667$

Under attribute age : $H(D_1) \approx 0.8113$, $H(D_2) = 0$, $H(D_3) \approx 0.9183$

$g(D, age) = H(D) - H(D|age) \approx 0.2516$

Q5

$$H_{age}(D) = -\sum_{j=1}^{3} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} = -\frac{4}{9}\log_2\frac{4}{9} - \frac{2}{9}\log_2\frac{2}{9} - \frac{1}{3}\log_2\frac{1}{3} \approx 1.5305$$

$$g_R(D, age) = \frac{g(D, age)}{H_{age}(D)} \approx 0.1644$$

(b) For attribute work: $n=2$. a_1: Yes, $|D_1| = 3$; a_2: No, $|D_2| = 6$

$|D_{ik}|$: $|D_{11}| = 3$, $|D_{12}| = 0$; $|D_{21}| = 0$, $|D_{22}| = 6$

Actually at here we could tell the training set is purely separated and the decision tree can be explicitly constructed, but keep on as an algorithm.

$$H(D|work) = \sum_{i=1}^{2} \frac{|D_i|}{|D|} H(D_i) = \frac{1}{3}\left(-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right) + \frac{2}{3}\left(-\frac{0}{6}\log_2\frac{0}{6} - \frac{6}{6}\log_2\frac{6}{6}\right) = 0$$

$$g(D, work) = H(D) - H(D|work) = H(D) \approx 0.9183$$

$$H_{work}(D) = -\sum_{j=1}^{2} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx 0.9183$$

$$g_R(D, work) = \frac{g(D, work)}{H_{work}(D)} = 1$$

(c) For attribute credit: $n=3$. a_1: normal: $|D_1| = 4$; a_2: good, $|D_2| = 4$; a_3: excellent: $|D_3| = 1$

$|D_{ik}|$: $|D_{11}| = 0$, $|D_{12}| = 4$; $|D_{21}| = 2$, $|D_{22}| = 2$; $|D_{31}| = 1$, $|D_{32}| = 0$.

$$H(D|credit) = \sum_{i=1}^{3} \frac{|D_i|}{|D|} H(D_i) = \frac{4}{9}\left(-\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4}\right) + \frac{4}{9}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + 0$$

$$= \frac{4}{9} \approx 0.4444$$

$$g(D, credit) = H(D) - H(D|credit) \approx 0.4739$$

$$H_{credit}(D) = -\sum_{j=1}^{k} \frac{|D_j|}{|D|} \log \frac{|D_j|}{|D|} = -\frac{4}{9}\log_2\frac{4}{9} - \frac{4}{9}\log_2\frac{4}{9} - \frac{1}{9}\log_2\frac{1}{9} \approx 1.3921$$

$$g_R(D, credit) = \frac{g(D, credit)}{H_{credit}(D)} \approx 0.3404$$

Q5 $0.2516 < 0.3404 < 1$

$g_R(D, age) < g_R(D, credit) < g_R(D, work)$

Choose work as the second decision attribute

$D_1 = \{3, 13, 14\}$, $D_2 = \{1, 2, 5, 6, 7, 15\}$

Note that data in D_1 purely belong to class 1: Yes while data in D_2 purely belong to class 2: No

Thus, the decision tree can be explicitly constructed with "House" as the attribute for the first distribution and "Work" as the attribute for the second distribution which is: