Assignment 1 Solution

(COMP3605 - Introduction to Data Analytics, 2022-2023)

Date Available: Tuesday, September 27, 2022
Due Date: 11.50 PM, Wednesday, October 12, 2022
Total Mark: 100 marks

Students are NOT asked to use Laplacian correction
Some students did not read the Question 1 carefully and they presented their answer wrongly. Consequently, they get zero mark for this.
Some students carelessly computed $P(x_k \mid C_1)$ and $P(x_k \mid C_2)$. Thus, their results are wrong.

Solution to Question 1 [40 marks]
Given test instance $\boldsymbol{X}$ = (MP = Yes, WP = Yes, CCI = No, Gender = Female).
Let the class labels be C_1 = Yes, C_2 = No.
The used formulas are $P(C_i \mid \boldsymbol{X}) = \dfrac{P(\boldsymbol{X} \mid C_i)P(C_i)}{P(\boldsymbol{X})}$,

$P(\boldsymbol{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$, $P(x_k \mid C_i) = |C_{i,x_k}| / |C_{i,D}|$

Check $P(\boldsymbol{X} \mid C_i)P(C_i) > P(\boldsymbol{X} \mid C_j)P(C_j)$? for $1 \le j \le m, j \neq i$.
$P(C_i) = |C_{i,D}|/|D|$, where $|D| = 10$, $|C_{1,D}| = 5$, $|C_{2,D}| = 5$
[6 marks, 3 marks for $P(C_1)$ and 3 marks for $P(C_2)$]
$P(C_1) = |C_{1,D}|/|D| = P(\text{Yes}) = 5/10 = 0.5$,
$P(C_2) = |C_{2,D}|/|D| = P(\text{No}) = 5/10 = 0.5$
$P(x_k \mid C_i) = |C_{i,x_k}| / |C_{i,D}|$, i.e., $P(x_k \mid C_1) = |C_{1,x_k}| / |C_{1,D}|$, $P(x_k \mid C_2) = |C_{2,x_k}| / |C_{2,D}|$
• The calculations for $P(C_1 \mid \boldsymbol{X}) = P(\text{LIP = Yes} \mid \boldsymbol{X})$ are
[8 marks, 2 marks for each correct computation $P(x_k \mid C_1)$]
$P(\text{MP = Yes} \mid C_1) = 5/5 = 1$, $P(\text{WP = Yes} \mid C_1) = 4/5 = 0.8$,
$P(\text{CCI = No} \mid C_1) = 2/5 = 0.4$, $P(\text{Gender = Female} \mid C_1) = 3/5 = 0.6$
[8 marks = 4 + 4]

$P(\boldsymbol{X} \mid C_1) = P(\boldsymbol{X} \mid C_1) = \prod_{k=1}^{n} P(x_k \mid C_1) = 5/5 \times 4/5 \times 2/5 \times 3/5 = 24/125 = 0.192$

$P(\boldsymbol{X} \mid C_1)P(C_1) = \left(\prod_{k=1}^{n} P(x_k \mid C_1)\right) \times P(C_1) = 24/125 \times 5/10 = 12/125 = 0.096$

• The calculations for $P(C_2 \mid \boldsymbol{X}) = P(\text{LIP = No} \mid \boldsymbol{X})$ are
[8 marks, 2 marks for each correct computation $P(x_k \mid C_2)$]
$P(\text{MP = Yes} \mid C_2) = 2/5 = 0.4$, $P(\text{WP = Yes} \mid C_2) = 0/5 = 0$,
$P(\text{CCI = No} \mid C_2) = 5/5 = 1$, $P(\text{Gender = Female} \mid C_2) = 1/5 = 0.2$
Notes:
1. without using Laplacian correction, students still can get full mark for this section, i.e., 8 marks.
2. some students use correction for $P(x_k \mid C_2)$ only, they can get full mark for this section.
3. some students use correction for both $P(x_k \mid C_2)$ and $P(x_k \mid C_1)$, they can get full mark for this section.

4. some students use Laplacian correction for $P(\text{WP} = \text{Yes} \mid C_2) = 0/5 = 0$ only, they can get full mark for this section. The attribute WP has two possible values, namely {Yes, No}. That is, we have $q = 2$ counts. Thus, $P(\text{WP} = \text{Yes} \mid C_2) = 0/5 = 0$ becomes $P(\text{WP} = \text{Yes} \mid C_2) = (0 + 1)/(5 + 2) = 1/7 = 0.1428 = 0.143$.
5. some students use m-estimate approach: $P(x_k \mid C_i) = (n_c + mp) / (n + m)$, where $n = 5$, n_c is the number of training examples from class C_i that take on the value x_k, [$m = 2, p = 1/2$] or [$m = 4, p = 1/4$], they can get full mark for this section.

/* with correction $P(x_k \mid C_2)$ only: use m-estimate approach: $P(x_k \mid C_i) = (n_c + mp) / (n + m)$, where $n = 5$, $m = 4, p = 1/4$, and n_c is the number of training examples from class C_i that take on the value x_k, below calculations are accepted
$P(\text{MP} = \text{Yes} \mid C_2) = (2 + 1)/(5 + 4) = 3/9 = 1/3 = 0.333$,
$P(\text{WP} = \text{Yes} \mid C_2) = (0 + 1)/(5 + 4) = 1/9 = 0.111$,
$P(\text{CCI} = \text{No} \mid C_2) = (5 + 1)/(5 + 4) = 6/9 = 2/3 = 0.666$,
$P(\text{Gender} = \text{Female} \mid C_2) = (1 + 1)/(5 + 4) = 2/9 = 0.222$ */
[8 marks = 4 + 4]

$$P(\boldsymbol{X} \mid C_2) = P(\boldsymbol{X} \mid C_2) = \prod_{k=1}^{n} P(x_k \mid C_2) = 2/5 \times 0/5 \times 5/5 \times 1/5 = 0$$

$$P(\boldsymbol{X} \mid C_2)P(C_2) = \left(\prod_{k=1}^{n} P(x_k \mid C_2)\right) \times P(C_2) = 0 \times 5/10 = 0$$

/* with correction $P(x_k \mid C_2)$ only:

$$P(\boldsymbol{X} \mid C_2) = P(\boldsymbol{X} \mid C_2) = \prod_{k=1}^{n} P(x_k \mid C_2) = 3/9 \times 1/9 \times 6/9 \times 2/9 = 0.333 \times 0.111 \times 0.666 * 0.222$$
$$= 0.00548 = 0.0055$$

$$P(\boldsymbol{X} \mid C_2)P(C_2) = \left(\prod_{k=1}^{n} P(x_k \mid C_2)\right) \times P(C_2) = 0.00548 \times 5/10 = 0.00274 = 0.003 \text{ */}$$

• [2 marks] We have $P(\boldsymbol{X} \mid C_1)P(C_1) = 0.096 > P(\boldsymbol{X} \mid C_2)P(C_2) = 0$. Thus, the naïve Bayesian classifier will predict that the class label LIP = Yes for the given test instance $\boldsymbol{X}$.

/* with correction $P(x_k \mid C_2)$ only
• [2 marks] We have $P(\boldsymbol{X} \mid C_1)P(C_1) = 0.096 > P(\boldsymbol{X} \mid C_2)P(C_2) = 0.003$. Thus, the naïve Bayesian classifier will predict that the class label LIP = Yes for the given test instance $\boldsymbol{X}$. */

Some students correctly apply Laplacian correction as follows.
$P(\text{MP} = \text{Yes} \mid C_2) = 3/7 = 0.429$, $P(\text{WP} = \text{Yes} \mid C_2) = 1/7 = 0.143$,
$P(\text{CCI} = \text{No} \mid C_2) = 6/7 = 0.857$, $P(\text{Gender} = \text{Female} \mid C_2) = 2/7 = 0.286$

$$P(\boldsymbol{X} \mid C_2) = P(\boldsymbol{X} \mid C_2) = \prod_{k=1}^{n} P(x_k \mid C_2) = 3/7 \times 1/7 \times 6/7 \times 2/7 = 0.015$$

$$P(\boldsymbol{X} \mid C_2)P(C_2) = \left(\prod_{k=1}^{n} P(x_k \mid C_2)\right) \times P(C_2) = 0.015 \times 5/10 = 0.0075$$

We have $P(\boldsymbol{X} \mid C_1)P(C_1) = 0.096 > P(\boldsymbol{X} \mid C_2)P(C_2) = 0.0075$. Thus, the naïve Bayesian classifier will predict that the class label LIP = Yes for the given test instance $\boldsymbol{X}$.

Solution to Question 2 [30 marks]

a. [10 marks] information gain for **Temperature**.

Let the class labels be C_1 = Yes, C_2 = No. $|D| = 14$, $|C_{1,D}| = 9$, $|C_{2,D}| = 5$, $p_i = |C_{i,D}| / |D|$,
$p_1 = |C_{1,D}| / |D| = 9/14 = 0.643$, $p_2 = |C_{2,D}| / |D| = 5/14 = 0.357$.

$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$, (bits), where $\log_2 0 = 0$.

[3 marks]

$Info(D) = -[p_1\log_2(p_1) + p_2\log_2(p_2)] = -[0.643\times\log_2(0.643) + 0.357\times\log_2(0.357)]$
$Info(D) = -[0.643\times(-0.637) + 0.357\times(-1.485)] = -[-0.410 - 0.531] = 0.9403 \approx 0.940$

[6 marks = 2 + 2 + 2]

$Info_A(D) = \sum_{j=1}^{v} |D_j| / |D| \times Info(D_j)$, $Info(D_j) = -\sum_{i=1}^{m} p_{i,j} \log_2(p_{i,j})$, $p_{i,j} = |C_{i,D_j}| / |D_j|$

$Info_A(D) = |D_1|/|D| \times Info(D_1) + |D_2|/|D| \times Info(D_2) + |D_3|/|D| \times Info(D_3)$
$p_{1,1} = 3/4 = 0.75$, $p_{2,1} = 1/4 = 0.25$; $p_{1,2} = 4/6 = 0.667$, $p_{2,2} = 2/6 = 0.333$;
$p_{1,3} = 2/4 = 0.5$, $p_{2,3} = 2/4 = 0.5$

$Info_{\mathbf{Temperature}}(D) = 4/14\times[-(p_{1,1}\log_2(p_{1,1}) + p_{2,1}\log_2(p_{2,1}))] +$
$6/14\times[-(p_{1,2}\log_2(p_{1,2}) + p_{2,2}\log_2(p_{2,2}))] +$
$4/14\times[-(p_{1,3}\log_2(p_{1,3}) + p_{2,3}\log_2(p_{2,3}))]$

$Info_{\mathbf{Temperature}}(D) = 4/14\times[-(3/4\times\log_2(3/4) + 1/4\times\log_2(1/4))] +$
$6/14\times[-(4/6\times\log_2(4/6) + 2/6\times\log_2(2/6))] +$
$4/14\times[-(2/4\times\log_2(2/4) + 2/4\times\log_2(2/4))]$

$Info_{\mathbf{Temperature}}(D) = 0.286\times[-(0.75\times(-0.415) + 0.25\times(-2.0))] +$
$0.429\times[-(0.667\times(-0.585) + 0.333\times(-1.585))] +$
$0.286\times[-(0.5\times(-1.0) + 0.5\times(-1.0))]$

$Info_{\mathbf{Temperature}}(D) = 0.286\times[-(-0.311 - 0.5)] + 0.429\times[-(-0.39 - 0.528)] + 0.286\times[-(-0.5 - 0.5)]$
$Info_{\mathbf{Temperature}}(D) = 0.286\times0.811 + 0.429\times0.918 + 0.286\times1.0 = 0.232 + 0.394 + 0.286$
$Info_{\mathbf{Temperature}}(D) = 0.9111 \approx 0.911$ // we accept some rounding errors

[1 mark]

$Gain$(**Temperature**) $= Info(D) - Info_{\mathbf{Temperature}}(D) = 0.940 - 0.911 = 0.029$ (or 0.0292)

b. [10 marks] gain ratio for **Humidity** using $Gain$(**Humidity**) = 0.1518.

$GainRatio(A) = Gain(A) / SplitInfo_A(D)$, $SplitInfo_A(D) = -\sum_{j=1}^{v}[(|D_j| / |D|)\times\log_2(|D_j| / |D|)]$

[6 marks]

$SplitInfo_A(D) = -[(|D_1|/|D|)\times\log_2(|D_1|/|D|) + (|D_2|/|D|)\times\log_2(|D_2|/|D|)]$
$SplitInfo_{\mathbf{Humidity}}(D) = -[7/14\times\log_2(7/14) + 7/14\times\log_2(7/14)]$
$SplitInfo_{\mathbf{Humidity}}(D) = -[0.5\times\log_2(0.5) + 0.5\times\log_2(0.5)]$
$SplitInfo_{\mathbf{Humidity}}(D) = -[0.5\times(-1.0) + 0.5\times(-1.0)]$
$SplitInfo_{\mathbf{Humidity}}(D) = -[-0.5 - 0.5] = 1.0$ // we accept some rounding errors

[4 marks]

- Using $Gain$(**Humidity**) = 0.1518, we obtain
  $GainRatio$(**Humidity**) = $Gain$(**Humidity**) / $SplitInfo_{\mathbf{Humidity}}(D)$
  = 0.1518 / 1.0 = 0.152 (or 0.1518) // we accept some rounding errors

c. [10 marks] Gini index(**Outlook**), splitting subset {Sunny, Rainy}. $|D| = 14$, $|D_1| = 9$, $|D_2| = 5$

$Gini_A(D) = (|D_1| / |D|)\times Gini(D_1) + (|D_2| / |D|)\times Gini(D_2)$

$Gini(D_j) = 1 - \sum_{i=1}^{m} p_{i,j}^{\ 2} = 1 - [(p_{1,j})^2 + (p_{2,j})^2]$, $p_{i,j} = |C_{i,D_j}| / |D_j|$

[4 marks]

$Gini(D_1) = 1 - [(p_{1,1})^2 + (p_{2,1})^2]$, $p_{1,1} = |C_{1,D_1}| / |D_1|$, $p_{2,1} = |C_{2,D_1}| / |D_1|$

$Gini(D_1) = 1 - [(5/10)^2 + (5/10)^2] = 1 - (0.5^2 + 0.5^2) = 1 - (0.25 + 0.25)$

$Gini(D_1) = 1 - 0.5 = 0.5$

[4 marks]

$Gini(D_2) = 1 - [(p_{1,2})^2 + (p_{2,2})^2]$, $p_{1,2} = |C_{1,D_2}| / |D_2|$, $p_{2,2} = |C_{2,D_2}| / |D_2|$

$Gini(D_2) = 1 - [(4/4)^2 + (0/4)^2] = 1 - (1.0^2 + 0.0^2)$

$Gini(D_2) = 1 - 1.0 = 0.0$

[2 marks]

$Gini_{\textbf{Outlook} \in \{\text{Sunny, Rainy}\}}(D) = |D_1|/|D| \times Gini(D_1) + |D_2|/|D| \times Gini(D_2)$

$Gini_{\textbf{Outlook} \in \{\text{Sunny, Rainy}\}}(D) = 10/14 \times Gini(D_1) + 4/14 \times Gini(D_2)$

$Gini_{\textbf{Outlook} \in \{\text{Sunny, Rainy}\}}(D) \approx 0.714\times0.5 + 0.286\times0.0 \approx 0.357 + 0.0$

$Gini_{\textbf{Outlook} \in \{\text{Sunny, Rainy}\}}(D) \approx 0.357$ (or 0.3571) // we accept some rounding errors

Solution to Question 3 [30 marks]

$|D| = p + n = 150$, $p = 50$, $n = 100$; $p_1 = 45$, $n_1 = 15$, $p_1 + n_1 = 60$; $p_2 = 4$, $n_2 = 1$, $p_2 + n_2 = 5$.

a. [5 marks] Compute the rule accuracy of R_1 and R_2.

$accuracy(R_1) = n_{correct} / n_{covers} = n_{correct} / (p_1 + n_1) = 45 / 60 = 3/4 = 75\%$.

$accuracy(R_2) = n_{correct} / n_{covers} = n_{correct} / (p_2 + n_2) = 4 / 5 = 80\%$.

We have $accuracy(R_2) > accuracy(R_1)$. Thus, according to the rule accuracy metric, R_2 is a better rule than R_1.

b. [5 marks] Compute the rule coverage of R_1 and R_2.

$coverage(R_1) = n_{covers} / |D| = (p_1 + n_1) / |D| = 60 / 150 = 2/5 = 0.4$.

$coverage(R_2) = n_{covers} / |D| = (p_2 + n_2) / |D| = 5 / 150 = 0.0333$.

We have $coverage(R_1) > coverage(R_2)$. Thus, according to the rule coverage metric, R_1 is a better rule than R_2.

c. [10 marks] Compute *FOIL_GAIN* for R_1 and R_2 with respect to R_0

Assume that the initial rule R_0: {} → + covers $p_0 = 50$ positive examples and $n_0 = 100$ negative examples.

• Compute the FOIL's information gain for the rule R_1 with respect to R_0.

The rule R_1 covers $p_1 = 45$ positive examples and $n_1 = 15$ negative examples.

The FOIL's information gain for the rule R_1 with respect to R_0 is computed as

$$\begin{aligned}
FOIL_Gain(R_0, R_1) &= p_1 \times \{\log_2[p_1 / (p_1 + n_1)] - \log_2[p_0 / (p_0 + n_0)]\} \\
&= 45 \times \{\log_2[45/(45 + 15)] - \log_2[50/(50 + 100)]\} \\
&= 45 \times [\log_2(45/60) - \log_2(50/150)] \\
&= 45 \times [\log_2(3/4) - \log_2(1/3)] \\
&= 45 \times [-0.41504 - (-1.58496)] \\
&= 45 \times (-0.41504 + 1.58496)
\end{aligned}$$

$= 45 \times 1.1699 \approx 52.6466 \approx 52.65$

• Compute the FOIL's information gain for the rule R_2 with respect to R_0.
The rule R_2 covers $p_2 = 4$ positive examples and $n_2 = 1$ negative examples.
The FOIL's information gain for the rule R_2 with respect to R_0 is computed as

$$
\begin{aligned}
FOIL\_Gain(R_0, R_2) &= p_2 \times \{\log_2[p_2 / (p_2 + n_2)] - \log_2[p_0 / (p_0 + n_0)]\} \\
&= 4 \times \{\log_2[4/(4 + 1)] - \log_2[50/(50 + 100)]\} \\
&= 4 \times [\log_2(4/5) - \log_2(1/3)] \\
&= 4 \times [-0.32193 - (-1.58496)] \\
&= 4 \times (-0.32193 + 1.58496) \\
&= 4 \times 1.2630 \approx 5.0521 \approx 5.05
\end{aligned}
$$

We have $FOIL\_Gain(R_0, R_1) > FOIL\_Gain(R_0, R_2)$. Therefore, according to the FOIL's information gain metric, R_1 is a better rule than R_2.

d. [10 marks] Laplace

• The Laplace measure takes into account the rule coverage and is computed by

$$\text{Laplace} = (f_+ + 1) / (n + k),$$

where n is the number of examples covered by the rule, f_+ is the number of positive examples covered by the rule, $k = m$ is the number of classes.

• If the rule coverage is large, then its Laplace measure asymptotically approaches the rule accuracy f_+ / n. That is, the rule that has the Laplace measure close to its accuracy is a better rule.

• The Laplace measure for R_1 is Laplace(R_1) = (45 + 1) / (60 + 2) = 46/62 = 23/31 ≈ 74.19%, which is very close to its accuracy of 75% (i.e., 75 – 74.19 = 0.81).

• The Laplace measure for R_2 is Laplace(R_2) = (4 + 1) / (5 + 2) = 5/7 ≈ 71.43%, which is quite far from its accuracy of 80% (i.e., 80 – 71.43 = 8.57) because R_2 has a much lower coverage.

• Laplace(R_1) is close to its accuracy. Thus, R_1 is a better rule than R_2.

End of Assignment 1 Solution