Coursework Exam Solution

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

Date Available: 3 PM, Friday, November 11, 2022
Due Date: 4 PM, Friday, November 11, 2022
Total Mark: 100 marks

Solution to Question 1 [50 marks] (*min_sup*) count = 3
• Find all frequent itemsets in *D* using the **vertical Apriori** algorithm

[5 marks] for correct horizontal to vertical transformation
The Vertical Data Format of the Transaction Data Set *D*

Itemset	TID_set
A	{T300}
C	{T400, T500}
D	{T200}
E	{T100, T200, T300, T500}
I	{T500}
K	{T100, T200, T300, T400, T500}
M	{T100, T300, T400}
N	{T100, T200}
O	{T100, T200, T500}
U	{T400}
Y	{T100, T200, T400}

[6 marks] for correct C_1
• set of candidate 1-itemsets C_1 = {{A}, {C}, {D}, {E}, {I}, {K}, {M}, {N}, {O}, {U}, {Y}}
- Determine supports of itemsets in C_1 using lengths of their *tid* lists

Itemset	TID_set	sup. count
A	{T300}	1
C	{T400, T500}	2
D	{T200}	1
E	{T100, T200, T300, T500}	4
I	{T500}	1
K	{T100, T200, T300, T400, T500}	5
M	{T100, T300, T400}	3
N	{T100, T200}	2
O	{T100, T200, T500}	3
U	{T400}	1
Y	{T100, T200, T400}	3

Thus, we have C_1 = {{A}:1, {C}:2, {D}:1, {E}:4, {I}:1, {K}:5, {M}:3, {N}:2, {O}:3 (not4), {U}:1, {Y}:3}

[6 marks] for correct F_1
The set of frequent 1-itemsets is F_1 = {{E}:4, {K}:5, {M}:3, {O}:3, {Y}:3}

- Construct vertical *tid* lists of each frequent item [by scanning the data set once]. The result is shown below.

frequent 1-itemsets in Vertical Data Format

Itemset	TID_set
E	{T100, T200, T300, T500}
K	{T100, T200, T300, T400, T500}
M	{T100, T300, T400}
O	{T100, T200, T500}
Y	{T100, T200, T400}

• For $k = 1$
• while $F_1 \neq \varnothing$
- Generate C_2 by joining itemset-pairs in F_1 ($k = 1$): $C_2 = F_1 \bowtie F_1$
[6 marks] for correct C_2
C_2 = {{E, K}, {E, M}, {E, O}, {E, Y}, {K, M}, {K, O}, {K, Y}, {M, O}, {M, Y}, {O, Y}}
- Prune itemsets from C_{k+1} that are infrequent due to their infrequent ($k - 1$)-itemsets.
(i.e., Prune itemsets from C_2 that are infrequent due to their infrequent 1-itemsets. Do nothing in this case.)
- Generate *tid* list of each **candidate itemset** in C_2 by intersecting *tid* lists of the itemset-pair in F_1 that was used to create **the candidate itemset**. The result of intersecting is shown below.

candidate 2-itemsets in Vertical Data Format

Itemset	TID_set
{E, K}	{T100, T200, T300, T500}
{E, M}	{T100, T300}
{E, O}	{T100, T200, T500}
{E, Y}	{T100, T200}
{K, M}	{T100, T300, T400}
{K, O}	{T100, T200, T500}
{K, Y}	{T100, T200, T400}
{M, O}	{T100}
{M, Y}	{T100, T400}
{O, Y}	{T100, T200}

- Determine supports of itemsets in C_2 using lengths of their *tid* lists

Itemset	TID_set	sup. count
{E, K}	{T100, T200, T300, T500}	4
{E, M}	{T100, T300}	2
{E, O}	{T100, T200, T500}	3
{E, Y}	{T100, T200}	2

{K, M}	{T100, T300, T400}	3
{K, O}	{T100, T200, T500}	3
{K, Y}	{T100, T200, T400}	3
{M, O}	{T100}	1
{M, Y}	{T100, T400}	2
{O, Y}	{T100, T200}	2

We have C_2 = {{E, K}:4, {E, M}:2, {E, O}:3, {E, Y}:2, {K, M}:3, {K, O}:3, {K, Y}:3, {M, O}:1, {M, Y}:2, {O, Y}:2}

[6 marks] for correct F_2

- F_2 = Frequent itemsets of C_2 together with their *tid* lists. Thus, we have the set of frequent 2-itemsets, F_2, and their *tid* lists are shown below.

frequent 2-itemsets in Vertical Data Format

Itemset	**TID_set**
{E, K}	{T100, T200, T300, T500}
{E, O}	{T100, T200, T500}
{K, M}	{T100, T300, T400}
{K, O}	{T100, T200, T500}
{K, Y}	{T100, T200, T400}

• That is, we have F_2 = {{E, K}:4, {E, O}:3, {K, M}:3, {K, O}:3, {K, Y}:3}
// removed {E, M}, {E, Y}, {M, O}, {M, Y}, {O, Y}

• For $k = 2$
• while $F_2 \neq \varnothing$
- Generate C_3 by joining itemset-pairs in F_2 ($k = 2$): $C_3 = F_2 \bowtie F_2$
[6 marks] for correct C_3
C_3 = {{E, K, O}, {K, M, O}, {K, M, Y}, {K, O, Y}}.
- Prune itemsets from C_3 that are infrequent due to their infrequent 2-itemsets. Thus, we obtain C_3 = {{E, K, O}}.
// removed {K, M, O}, {K, M, Y}, {K, O, Y}
- Generate *tid* list of each **candidate itemset** in C_3 by intersecting *tid* lists of the itemset-pair in F_2 that was used to create **the candidate itemset**. The result of intersecting is shown below.

candidate 3-itemsets in Vertical Data Format

Itemset	**TID_set**
{E, K, O}	{T100, T200, T500}

- Determine supports of 3-itemsets in C_3 using lengths of their *tid* lists

Itemset	**TID_set**	**sup. count**
{E, K, O}	{T100, T200, T500}	3

We have C_3 = {{E, K, O}:3}

[6 marks] for correct F_3

- F_3 = Frequent itemsets of C_3 together with their *tid* lists. Thus, we have the set of frequent 3-itemsets, F_3, and their *tid* lists are shown below.

frequent 3-itemsets in Vertical Data Format

Itemset	**TID_set**
{E, K, O}	{T100, T200, T500}

• That is, we have $F_3 = \{\{E, K, O\}:3\}$

• For $k = 3$
• while $F_3 \neq \varnothing$
- Generate C_4 by joining itemset-pairs in F_3 ($k = 3$): $C_4 = F_3 \bowtie F_3$

[4 marks] for correct C_4 and F_4
$C_4 = \varnothing \rightarrow F_4 = \varnothing$, and the algorithm terminates.

[5 marks]
In conclusion, the set of all frequent itemsets found is
$F = \{\{E\}:4, \{K\}:5, \{M\}:3, \{O\}:3, \{Y\}:3,$
$\{E, K\}:4, \{E, O\}:3, \{K, M\}:3, \{K, O\}:3, \{K, Y\}:3,$
$\{E, K, O\}:3\}$

Solution to Question 2 [50 marks]
a. [6 marks] 3 marks for each correct support vector
Specify support vectors from the given data set D.
The first two instances $\mathbf{x}_1$ and $\mathbf{x}_2$ have Lagrange multipliers $\lambda_i > 0$ (i.e., $\lambda_1 = 2.7027$, $\lambda_2 = 2.7027$).
Thus, the two support vectors (SVs) are $\mathbf{x}_1 = (2, 2.5)$ and $\mathbf{x}_2 = (2.5, 3.2)$.

b. [40 marks] Determine a decision boundary (DB) of a linear SVM (support vector machine).
[20 marks] 10 marks for w_1 and 10 marks for w_2
Compute $\mathbf{w} = (w_1, w_2)$
Let $\mathbf{w} = (w_1, w_2)$ and b denote parameters of the DB.
Adopting the equation $\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i$, where $N = 8$, we can calculate w_1 and w_2 as follows.
For $m = 2$ SVs $\mathbf{x}_1 = (2, 2.5)$ and $\mathbf{x}_2 = (2.5, 3.2)$ and two Lagrange multipliers $\lambda_1 = 2.7027$ and $\lambda_2 = 2.7027$, we have

$w_j = \sum_{i=1}^{N} \lambda_i y_i x_{ij}$, where x_{ij} is the jth component of $\mathbf{x}_i$ (e.g., $\mathbf{x}_i = (x_{i1}, x_{i2})$)

$$w_1 = \sum_{i=1}^{m} \lambda_i y_i x_{i1} = \lambda_1 y_1 x_{11} + \lambda_2 y_2 x_{21}$$

$w_1 = 2.7027 \times 1 \times 2 + 2.7027 \times (-1) \times 2.5 = -1.3514 \approx -1.35$

$$w_2 = \sum_{i=1}^{m} \lambda_i y_i x_{i2} = \lambda_1 y_1 x_{12} + \lambda_2 y_2 x_{22}$$

$w_2 = 2.7027 \times 1 \times 2.5 + 2.7027 \times (-1) \times 3.2 = -1.8919 \approx -1.89$

Thus, we obtain $\mathbf{w} = (w_1, w_2) = (-1.35, -1.89)$.

• [20 marks] 8 marks for $b^{(1)}$, 8 marks for $b^{(2)}$, 2 marks for average b, and 2 marks for DB
We have $b^{(k)} = y_i - \mathbf{w} \cdot \mathbf{x}_i$, where $\mathbf{x}_i$ are support vectors (i.e., $i = 1, 2$), $k = 1, 2, ..., m$)
For $m = 2$ SVs $\mathbf{x}_1 = (2, 2.5)$ and $\mathbf{x}_2 = (2.5, 3.2)$ and $\mathbf{w} = (w_1, w_2) = (-1.35, -1.89)$, applying the formula $b^{(i)} = y_i - w_1 \times x_{i1} - w_2 \times x_{i2}$ for $i = 1, 2, ..., m$, we obtain

[8 marks]

$b^{(1)} = y_1 - \mathbf{w} \cdot \mathbf{x}_1 = y_1 - w_1 \times x_{11} - w_2 \times x_{12}$

$b^{(1)} = 1 - (-1.35) \times 2 - (-1.89) \times 2.5 = 1 - (-7.425) = 8.425 \approx 8.43$

[8 marks]

$b^{(2)} = y_2 - \mathbf{w} \cdot \mathbf{x}_2 = y_2 - w_1 \times x_{21} - w_2 \times x_{22}$

$b^{(2)} = -1 - (-1.35) \times 2.5 - (-1.89) \times (3.2) = -1 - (9.423) = 8.423 \approx 8.42$

[2 marks]

Averaging the values $b^{(1)}$ and $b^{(2)}$, we obtain $b = 8.425 \approx 8.43$

[2 marks]

With $\mathbf{w} = (w_1, w_2) = (-1.35, -1.89)$, $b = 8.43$, the DB of the linear SVM is
$w_1x_1 + w_2x_2 + b = 0 \Leftrightarrow -1.35x_1 - 1.89x_2 + 8.43 = 0$.

c. [4 marks] Describe how to use the trained linear SVM to classify a test instance **z**.
With the found parameters **w** and b of the DB, a test instance **z** is classified as follows.
$f(\mathbf{z}) = sign(\mathbf{w} \cdot \mathbf{z} + b)$.
- If $f(\mathbf{z}) > 0$ (or $\mathbf{w} \cdot \mathbf{z} + b \gtrsim 1$), then **z** is classified as positive class (i.e., class label $y = 1$).
- If $f(\mathbf{z}) < 0$ (or $\mathbf{w} \cdot \mathbf{z} + b \lesssim -1$), then **z** is classified as negative class (i.e., class label $y = -1$).

End of Coursework Exam Solution