Homework 4 - OSDA

Jexembayev Ruslan

December 16, 2023

The entire code for these tasks is available at the following GitHub link: `https://github.com/Jexembayev/OSDA`

1 Task 1.1

1.1 Data Table

No	Type	Mount	Price	Con	Snow	Ice	Dur	Accegrade
1	SK	F	206	1.9	1.4	1.8	2.7	F
2	SRK	F/R	520	2.1	0.8	3.8	2.3	F
3	SK	F	160	1.7	1.9	1.6	3.7	F
4	SK	F	213	1.7	2.0	2.4	3.4	F
5	SMS	F/R	598	1.6	2.4	7.0	2.8	F
6	SK	F	109	2.0	1.9	2.4	3.7	F
7	SRK	F/R	325	2.0	2.1	3.2	2.8	F
8	SMS	F/R	498	1.5	3.3	3.5	2.0	T
9	SRK	F/R	396	2.8	2.1	3.1	2.5	T
10	SRK	F/R	325	2.2	2.2	4.6	3.2	T
11	SRK	F/R	389	2.0	2.2	3.3	4.3	T
12	SRK	F	298	2.5	2.3	3.3	2.8	T
13	SK	F	149	1.9	2.5	4.0	3.8	T
14	SMS	F/R	684	1.7	3.3	4.4	2.2	T
15	SK	F	99	2.8	2.2	2.5	4.0	T
16	SK	F	140	2.6	2.3	3.3	3.4	T
17	SK	F	215	2.3	3.8	4.8	2.3	T

1.2 Analysis for Rows 15, 16, and 17

Row	Intersect	Extension	Classification
15	{'System': 'SK', 'Mount': 'F' 'Price': (99, 206), 'Con': (1.9, 2.8), 'Snow': (1.4, 2.2), 'Ice': (1.8, 2.5), 'Dur': (2.7, 4.0)}	Ext - [1, 5]	F
16	{'System': 'SK', 'Mount': 'F' 'Price': (140, 206), 'Con': (1.9, 2.6), 'Snow': (1.4, 2.3), 'Ice': (1.8, 3.3), 'Dur': (2.7, 3.4)}	Ext - [1]	F
17	{'System': 'SK', 'Mount': 'F' 'Price': (206, 215), 'Con': (1.9, 2.3), 'Snow': (1.4, 3.8), 'Ice': (1.8, 4.8), 'Dur': (2.3, 2.7)}	Ext - [1]	F

1.3 Conclusions

In the original dataset, rows 15, 16, and 17 were True, but in our analysis, they are False, resulting in 0% accuracy. This discrepancy arises from using only the first matching row without considering additional factors. Let's move on to the second task to see what happens when we shuffle the training rows.

2 Task 1.2

2.1 Data Table

No	System	Mount	Price	Con	Snow	Ice	Dur	Accegrade
1	SK	F	149	1.9	2.5	4.0	3.8	T
2	SRK	F / R	520	2.1	0.8	3.8	2.3	F
3	SRK	F / R	389	2.0	2.2	3.3	4.3	T
4	SK	F	213	1.7	2.0	2.4	3.4	F
5	SMS	F / R	598	1.6	2.4	7.0	2.8	F
6	SK	F	109	2.0	1.9	2.4	3.7	F
7	SRK	F / R	325	2.0	2.1	3.2	2.8	F
8	SMS	F / R	498	1.5	3.3	3.5	2.0	T
9	SRK	F / R	396	2.8	2.1	3.1	2.5	T
10	SK	F	160	1.7	1.9	1.6	3.7	F
11	SRK	F / R	389	2.0	2.2	3.3	4.3	T
12	SRK	F	298	2.5	2.3	3.3	2.8	T
13	SK	F	206	1.9	1.4	1.8	2.7	F
14	SMS	F / R	684	1.7	3.3	4.4	2.2	T
15	SK	F	99	2.8	2.2	2.5	4.0	T
16	SK	F	140	2.6	2.3	3.3	3.4	T
17	SK	F	215	2.3	3.8	4.8	2.3	T

2.2 Analysis for Rows 15, 16, and 17

Row	Intersect	Extension	Classification
15	{'System': 'SK', 'mount': 'F', 'Price': (99, 149), 'Con': (1.9, 2.8), 'Snow': (2.2, 2.5), 'Ice': (2.5, 4.0), 'Dur': (3.8, 4.0)}	Ext - [0, 14]	T
16	{'System': 'SK', 'mount': 'F', 'Price': (140, 149), 'Con': (1.9, 2.6), 'Snow': (2.3, 2.5), 'Ice': (3.3, 4.0), 'Dur': (3.4, 3.8)}	Ext - [0, 15]	T
17	{'System': 'SK', 'mount': 'F', 'Price': (149, 215), 'Con': (1.9, 2.3), 'Snow': (2.5, 3.8), 'Ice': (4.0, 4.8), 'Dur': (2.3, 3.8)}	Ext - [0, 16]	T

2.3 Conclusions

Now we have achieved an accuracy of 100%. Remarkably, this result might not fully represent the true performance of our model. To obtain a more reliable measure of accuracy, it's essential to employ cross-validation techniques. Let's proceed to implement cross-validation and assess the model's performance more accurately.

3 Task 1.3

A 5-fold cross-validation approach was utilized to evaluate the model's performance. The following accuracy scores were obtained for each fold:

Fold	**Train Indices**	**Test Indices**
1	[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]	[0, 1, 2]
2	[0, 1, 2, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]	[3, 4, 5]
3	[0, 1, 2, 3, 4, 5, 9, 10, 11, 12, 13, 14, 15, 16]	[6, 7, 8]
4	[0, 1, 2, 3, 4, 5, 6, 7, 8, 12, 13, 14, 15, 16]	[9, 10, 11]
5	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 15, 16]	[12, 13, 14]

The accuracies for each fold were as follows:

- Fold 1: Accuracy = 66.67%
- Fold 2: Accuracy = 0.0%
- Fold 3: Accuracy = 66.67%
- Fold 4: Accuracy = 66.67%
- Fold 5: Accuracy = 66.67%

The average accuracy across all folds was 53.33%.

4 Task 1.4

We define a new type of interval pattern structure where the value for attribute J of object I is given by $w(I, J)$, and the pattern structure value is: $[w(I, J), \infty)$. The intersection on such intervals is defined as:

$$[a1, \infty) \cap [a2, \infty) = [\min(a1, a2), \infty)$$

Using this method, example number 15 was classified as **True**.
To determine the most suitable interval pattern structure for our dataset, we performed 5-fold cross-validation on both methods. The results are as follows:

Fold	Train Indices	Test Indices	Accuracy
1	[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]	[0, 1, 2]	66.67%
2	[0, 1, 2, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]	[3, 4, 5]	0.0%
3	[0, 1, 2, 3, 4, 5, 9, 10, 11, 12, 13, 14, 15, 16]	[6, 7, 8]	66.67%
4	[0, 1, 2, 3, 4, 5, 6, 7, 8, 12, 13, 14, 15, 16]	[9, 10, 11]	66.67%
5	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 15, 16]	[12, 13, 14]	66.67%

Average Accuracy: 53.33%

5 Task 1.5

Similarly, we define another interval pattern structure with the pattern structure value $[w(I, J), \infty)$. However, the intersection on these intervals is defined as:

$$[a1, \infty) \cap [a2, \infty) = [\max(a1, a2), \infty)$$

Using this second method, the classification of example number 15 resulted in **Indeterminate**.

Fold	Train Indices	Test Indices	Accuracy
1	[3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]	[0, 1, 2]	33.33%
2	[0, 1, 2, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]	[3, 4, 5]	0.0%
3	[0, 1, 2, 3, 4, 5, 9, 10, 11, 12, 13, 14, 15, 16]	[6, 7, 8]	33.33%
4	[0, 1, 2, 3, 4, 5, 6, 7, 8, 12, 13, 14, 15, 16]	[9, 10, 11]	33.33%
5	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 15, 16]	[12, 13, 14]	0.0%

Average Accuracy: 20.0%

Conclusion

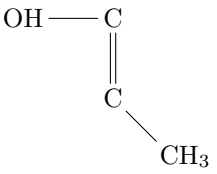
The cross-validation results demonstrate a notable difference in the performance of the 'minimal' and 'maximal' half-interval methods. The 'minimal' method achieved an average accuracy of 53.33%, suggesting a more effective prediction

capability compared to the 'maximal' method, which achieved an average accuracy of only 20.0%. These findings indicate that the 'minimal' half-interval intersection method is more suitable for the given dataset in terms of predictive accuracy.

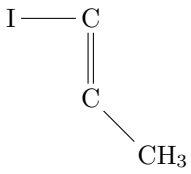
6 Task 2

6.1 Pattern concept for Positive examples

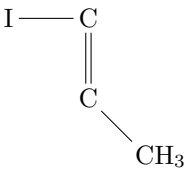
12 - 13 - 14 - 23 - 24 - 34



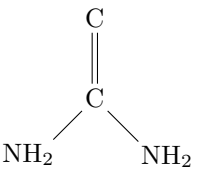
12



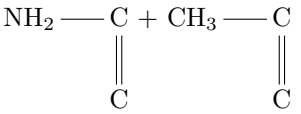
13



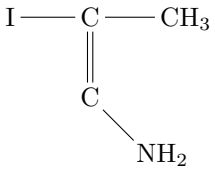
14



23

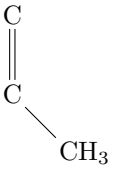


24

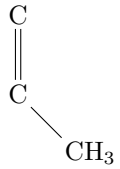


34

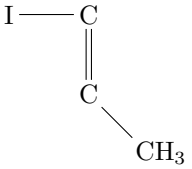
123 - 124 - 134 - 234



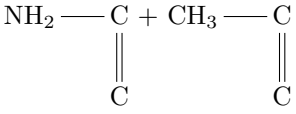
123



124

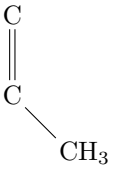


134



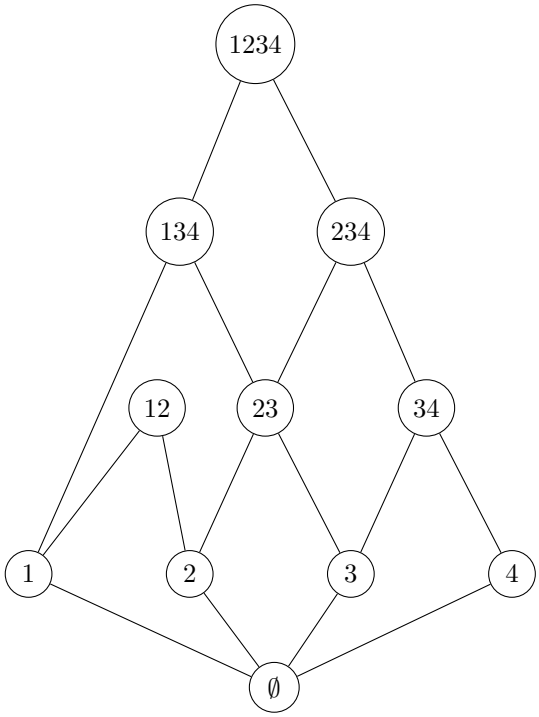
234

1234



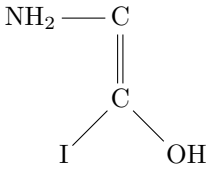
1234

6.2 Positive lattice

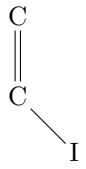


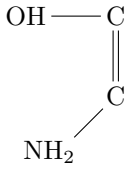
6.3 Pattern concept for Negative examples

56-57-67

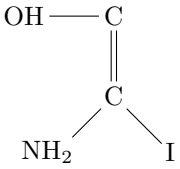


56



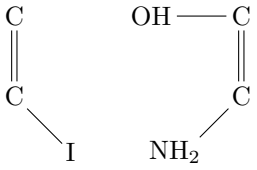


57



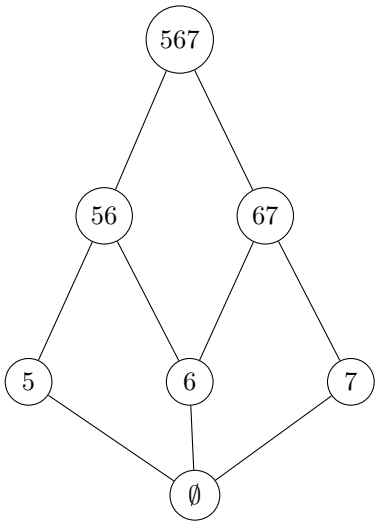
67

567



567

6.4 Negative lattice



6.5 Conclusion

The investigation into the molecular structures using the JSM method has led to the formation of two distinct hypotheses. The minimal positive hypothesis, denoted as H^+, comprises the set {1, 2, 3, 4}. This hypothesis is established due to the absence of analogous elements within the negative dataset. Conversely, the minimal negative hypothesis, represented by H^-, is composed of the set {5, 6, 7}, again validated by the lack of corresponding counterparts in the positive dataset.

Consequently, the classification of the test examples is as follows based on the delineated hypotheses:

- $G8$ is classified as positive (+), since H^+ is present within $G8$ and H^- is not; $H^+ \in G8$ and $H^- \notin G8$.
- $G9$ is classified as negative (−), owing to the inclusion of H^- within $G9$ and the exclusion of H^+; $H^- \in G9$ and $H^+ \notin G9$.
- $G10$ yields an indeterminate classification, as it fits both the positive and negative hypotheses; hence, $H^+ \in G10$ and $H^- \in G10$, leading to an undefined outcome.
- $G11$ aligns with the positive category (+), given that H^+ is found within $G11$ and H^- is not; thus, $H^+ \in G11$ and $H^- \notin G11$.