Machine Learning – Clustering (Unit 5)

Overview

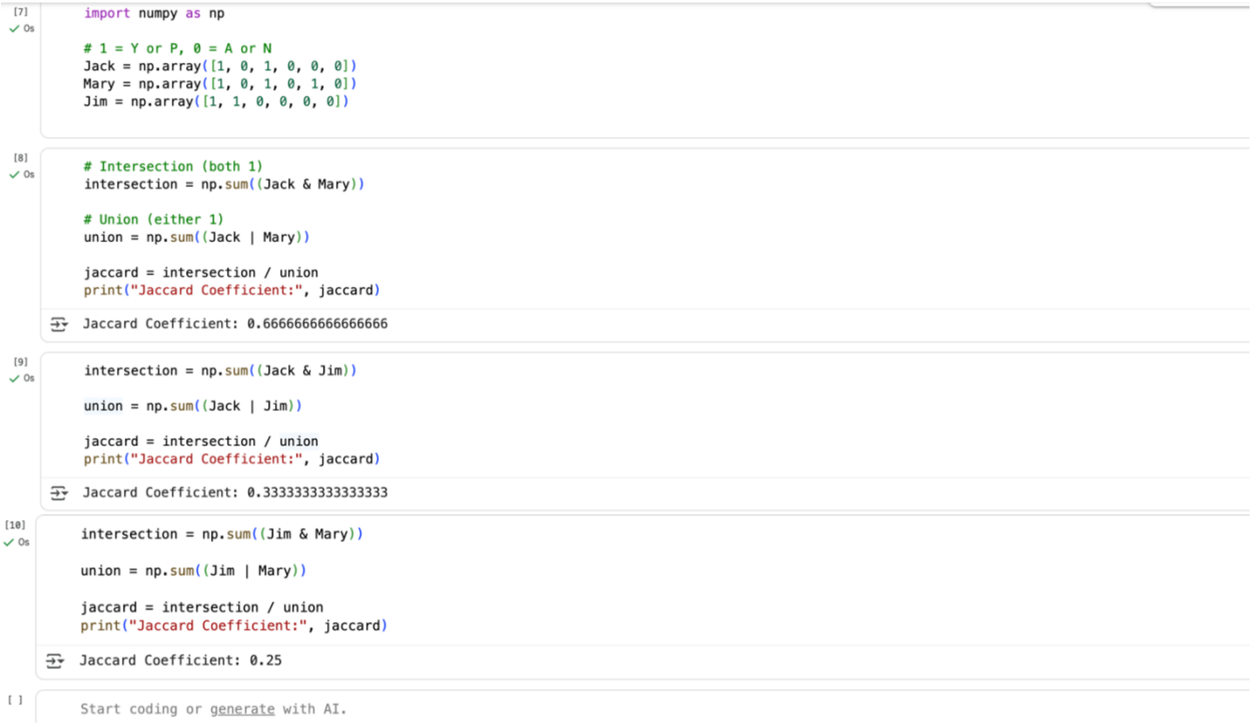
This unit introduced clustering for grouping similar data points, with applications in ML and pattern recognition (Tan et al., 2019). We explored distance measures, K-means, agglomerative clustering, evaluation methods, and common pitfalls.

What I Have Learned

I now better understand clustering and how distance metrics influence grouping. K-means is quick but limited, while agglomerative clustering offers more flexibility. The Jaccard Coefficient activity highlighted the importance of measuring similarity and evaluating results. This week boosted my confidence in using clustering practically.

Activity: Jaccard Coefficient Calculations

For this activity, I calculated the Jaccard coefficient to compare the similarity between three individuals based on their pathological test results. The Jaccard coefficient measures similarity between two sets by dividing the number of attributes they share in common by the total number of unique attributes they have (Tan et al., 2019).



```
import numpy as np

# 1 = Y or P, 0 = A or N
Jack = np.array([1, 0, 1, 0, 0, 0])
Mary = np.array([1, 0, 1, 0, 1, 0])
Jim = np.array([1, 1, 0, 0, 0, 0])
```

```
# Intersection (both 1)
intersection = np.sum((Jack & Mary))

# Union (either 1)
union = np.sum((Jack | Mary))

jaccard = intersection / union
print("Jaccard Coefficient:", jaccard)
```

```
Jaccard Coefficient: 0.6666666666666666
```

```
intersection = np.sum((Jack & Jim))

union = np.sum((Jack | Jim))

jaccard = intersection / union
print("Jaccard Coefficient:", jaccard)
```

```
Jaccard Coefficient: 0.3333333333333333
```

```
intersection = np.sum((Jim & Mary))

union = np.sum((Jim | Mary))

jaccard = intersection / union
print("Jaccard Coefficient:", jaccard)
```

```
Jaccard Coefficient: 0.25
```

```
Start coding or generate with AI.
```

The results were:

- **Jack and Mary:** 2 shared attributes / 6 total = **0.33**
- **Jack and Jim:** 2 shared / 6 = **0.33**
- **Jim and Mary:** 1 shared / 7 = **0.14**

Jack and Mary, and Jack and Jim each had 33% similarity, while Jim and Mary had only 14%. These low Jaccard coefficients show limited overlap, highlighting how useful this metric is for quantifying similarity in clustering and health data analysis (Tan et al., 2019).

Reference

Tan, P.N., Steinbach, M. and Kumar, V. (2019) *Introduction to Data Mining*. 2nd edn. Harlow: Pearson.