**CIND 110 – Assignment 4**

Brendan Dagys

1.1)

Candidate 1-itemsets

MINIMUM SUPPORT: 2/10

|  |  |
| --- | --- |
| **Item** | **Fraction** |
| milk | 5/10 |
| bread | 4/10 |
| cookies | 2/10 |
| eggs | 4/10 |
| butter | 2/10 |
| coffee | 3/10 |
| juice | 3/10 |

Sufficient 1-itemsets

|  |  |
| --- | --- |
| **Item** | **Fraction** |
| milk | 5/10 |
| bread | 4/10 |
| cookies | 2/10 |
| eggs | 4/10 |
| butter | 2/10 |
| coffee | 3/10 |
| juice | 3/10 |

Candidate 2-itemsets

|  |  |
| --- | --- |
| **Items** | **Fraction** |
| milk, bread | 4/10 |
| milk, cookies | 1/10 |
| milk, eggs | 3/10 |
| milk, butter | 0 |
| milk, coffee | 0 |
| milk, juice | 1/10 |
| bread, cookies | 1/10 |
| bread, eggs | 3/10 |
| bread, butter | 0 |
| bread, coffee | 0 |
| bread, juice | 0 |
| cookies,eggs | 1/10 |
| cookies, butter | 1/10 |
| cookies, coffee | 0 |
| cookies, juice | 0 |
| eggs, butter | 0 |
| eggs, coffee | 1/10 |
| eggs, juice | 0 |
| butter, coffee | 0 |
| butter, juice | 1/10 |
| coffee, juice | 1/10 |

Sufficient 2-itemsets

|  |  |
| --- | --- |
| **Items** | **Fraction** |
| milk, bread | 4/10 |
| milk, eggs | 3/10 |
| bread, eggs | 3/10 |

Candidate 3-itemsets

|  |  |
| --- | --- |
| **Items** | **Fraction** |
| milk, bread, eggs | 3/10 |

Sufficient 3-itemsets

|  |  |
| --- | --- |
| **Items** | **Fraction** |
| milk, bread, eggs | 3/10 |

1.2)

milk, bread 🡪 eggs (confidence = 0.75)

milk, eggs 🡪 bread (confidence = 1)

bread, eggs 🡪 milk (confidence = 1)

2.1)

INFORMATION GAIN CALCULATIONS TO DETERMINE ROOT NODE:

CITY

0.0771056

GENDER

0.00162

EDUCATION

0.4912765

AGE

0.3864224

FIRST NODE: EDUCATION

HIGH SCHOOL AND GRADUATE BECOME LEAVES, NEW INFORMATION GAIN CALCULATIONS FOR NEXT SPLIT:

AGE

0.6500

GENDER

0.04839

CITY

0.04839

SPLIT EDUCATION = COLLEGE ON AGE (4 WAYS), ALL ARE LEAVES, END OF ALGORITHM

Education

**YES** **YES**

**YES YES YES NO**

High School

20 - 30

31 - 40

51 - 60

41 - 50

College

Graduate

AGE

3.1)

**ITERATION 1**

RID: 1

0

6

= 7.2111

RID: 2

3

3

5

RID: 3

6

0

4

RID: 4

= 6.3246

2

2

RID: 5

= 7.2111

4

0

RID: 6

2

= 6.3246

= 6.3246

CLUSTER 1: 1, 6

RID: 2 and RID: 4 had ties for minimum distance. Tiebreaker cluster was chosen at random

CLUSTER 2: 2, 3

CLUSTER 3: 4, 5

New cluster locations

|  |  |  |
| --- | --- | --- |
|  | **X** | **Y** |
| **Cluster 1** | 8 | 5 |
| **Cluster 2** | 3.5 | 4 |
| **Cluster 3** | 2 | 7 |

**ITERATION 2**

RID: 1

1

4.5

= 6.7082

RID: 2

= 3.1623

1.5

= 4.2426

RID: 3

= 6.0828

1.5

3

RID: 4

= 6.0828

= 2.5

1

RID: 5

= 6.7082

= 4.2720

1

RID: 6

1

= 4.9244

= 6.0828

CLUSTER 1: 1, 6

CLUSTER 2: 2, 3

CLUSTER 3: 4, 5

New cluster locations

|  |  |  |
| --- | --- | --- |
|  | **X** | **Y** |
| **Cluster 1** | 8 | 5 |
| **Cluster 2** | 3.5 | 4 |
| **Cluster 3** | 2 | 7 |

END OF CLUSTERING (locations do not change)

3.2)

Clustering is an unsupervised machine learning algorithm, whereas classification is supervised. What this means is that in clustering, we do not know beforehand what the categories will be. The algorithm determines these, and then we use them to classify further observations.

In classification, we already know the outcome possibilities, and we are trying to come up with a formula that will correctly predict other observations.