Washington State University
School of Electrical Engineering and Computer Science
CptS 315 – Introduction to Data Mining
Online

Ananth Jillepalli

Homework 1

Name: Nam Jun Lee

Student Number: 11606459

Question 1. Consider the following market-basket data, where each row is a basket and shows the list of items that are part of that basket.

1. {A,B,C}
2. {A,C,D,E}
3. {A,B,F,G,H}
4. {A,B,X,Y,Z}
5. {A,C,D,P,Q,R,S}
6. {A,B,L,M,N}

a) What is the absolute support of item set {A, B} ?

Absolute support is the number of baskets containing all items in.

Therefore, row 1, 3, 4, and 6 have absolute support of item set {A,B}.

Hence, 4 baskets.

b) What is the relative support of item set {A, B} ?

Relative support is the fraction of baskets that contain items in.

Total baskets: 6

Baskets that contain items in set{A,B}: 4

Hence, Baskets that contain items in set{A,B} / Total baskets = 4 / 6 = 0.667.

c) What is the confidence of association rule A ⇒ B ?

Confidence of association rule is the probability of j given i_1, …, i_k that also contain j.

A is in: 6

B is also in: 4

Hence, 4 / 6 = 0.667.

Question 2. Answer the below questions about storing frequent pairs using triangular matrix and tabular method.

a) **Suppose we use a triangular matrix to count pairs and the number of items n = 20. If we store this triangular matrix as a ragged one-dimensional array Count, what is the index where count of pair (7, 8) is stored?**

Triangular Matrix as One-Dimensional Array find pair formula:

$$(I - 1)\ (n - I / 2) + j + I$$

Count of pair (7,8): I = 7, j = 8

Number of items: n = 20

So, (7 – 1) * (20 – 7 / 2) + 8 + 7 = 114.

b) **Suppose you are provided with the prior knowledge that only ten percent of the total pairs will have a non-zero count. In this case, which method among triangular matrix and tabular method should be preferred and why?**

Tabular method should be preferred. Because tabular approach beats triangular matrix only when at most 1/3 of all pairs have a nonzero count. Hence, only ten percent of the total pairs will have a non-zero count, then tabular method more preferred than triangular matrix.

Question 3. This question is about the PCY algorithm for counting frequent pairs of items. Suppose we have six items numbered 1, 2, 3, 4, 5, 6. Consider the following twelve baskets.

1. {1,2,3} 2. {2,3,4} 3. {3,4,5} 4. {4,5,6} 5. {1,3,5} 6. {2,4,6} 7. {1,3,4} 8. {2,4,5}
9. {3,5,6} 10. {1,2,4} 11. {2,3,5} 12. {3,4,6}

Suppose the support threshold is 4. On the first pass of the PCY algorithm, we use a hash table with 11 buckets, and the set {i, j} is hashed to i × j mod 11.

a) **By any method, compute the support for each item and each pair of items.**

Absolute support:

{1} = 4, {2} = 6, {3} = 8, {4} = 8, {5} = 6, {6} = 4

{1,2} = 2, {1,3} = 3, {1,4} = 2, {1,5} = 1, {1,6} = 0

{2,3} = 3, {2,4} = 4, {2,5} = 2, {2,6} = 1

{3,4} = 4, {3,5} = 4, {3,6} = 2

{4,5} = 3, {4,6} = 3

{5,6} = 2

Relative support: Total baskets: 12 / Baskets that contain items in set: {?}

{1} = 4/12 = 0.33, {2} = 6/12 = 0.5, {3} = 8/12 = 0.67, {4} = 8/12 = 0.67, {5} = 6/12 = 0.5, {6} = 4/12 = 0.33

{1,2} = 2/12 = 0.17, {1,3} = 3/12 = 0.25, {1,4} = 2/12 = 0.17, {1,5} = 1/12 = 0.0834, {1,6} = 0/12 = 0

{2,3} = 3/12 = 0.25, {2,4} = 4/12 = 0.33, {2,5} = 2/12 = 0.17, {2,6} = 1/12 = 0.0834

{3,4} = 4/12 = 0.33, {3,5} = 4/12 = 0.33, {3,6} = 2/12 = 0.17

{4,5} = 3/12 = 0.25, {4,6} = 3/12 = 0.25

{5,6} = 2/12 = 0.17

b) Which pairs hash to which buckets?

The set {i, j} is hashed to i × j mod 11:

{1,2} = 2 mod 11 = 2

{1,3} = 3 mod 11 = 3

{1,4} = 4 mod 11 = 4

{1,5} = 5 mod 11 = 5

{1,6} = 6 mod 11 = 6

{2,3} = 6 mod 11 = 6

{2,4} = 8 mod 11 = 8

{2,5} = 10 mod 11 = 10

{2,6} = 12 mod 11 = 1

{3,4} = 12 mod 11 = 1

{3,5} = 15 mod 11 = 4

{3,6} = 18 mod 11 = 7

{4,5} = 20 mod 11 = 9

{4,6} = 24 mod 11 = 2

{5,6} = 30 mod 11 = 8

c) Which buckets are frequent?

we use a hash table with 11 buckets from 0 to 10.

Count of pairs hashed: support{?} + support{?}

Bucket 0 = 0

Bucket 1 = {2,6} + {3,4} = 1 + 4 = 5

Bucket 2 = {1,2} + {4,6} = 2 + 3 = 5

Bucket 3 = {1,3} = 3

Bucket 4 = {1,4} + {3,5} = 2 + 3 = 5

Bucket 5 = {1,5} = 1

Bucket 6 = {1,6} + {2,3} = 0 + 3 = 3

Bucket 7 = {3,6} = 2

Bucket 8 = {2,4} + {5,6} = 4 + 2 = 6

Bucket 9 = {4,5} = 3

Bucket 10 = {2,5} = 2

A bucket is frequent if its counts is at least the support threshold (hash to bucket > 4):

Hence, 1, 2, 4, and 8 buckets are frequent.

d) Which pairs are counted on the second pass of the PCY algorithm?

Second pass of the PCY algorithm, we only count pairs of frequent items that also hash to a frequent bucket. Therefore, 1, 2, 4, and 8 buckets are frequent.

Bucket 1: {2,6}, {3,4} hashed

Bucket2: {1,2}, {4,6} hashed

Bucket4: {1,4}, {3,5} hashed

Bucket8: {2,4}, {5,6} hashed

Hence, pairs {2,6}, {3,4}, {1,2}, {4,6}, {1,4}, {3,5}, {2,4}, {5,6} are counted on the second pass of the PCY algorithm.

Question 4. read the following paper and write a brief summary of the main points in at most ONE page. You can skip the theoretical parts.

The main point of this article describes the local document fingerprint algorithm and its complexity. It detects this copy by accurately identifying the copied content, including the copy, within a large document. In current society, there are cases where the original copy is used to benefit. For example, students plagiarize their homework on the web for their grades, and companies reuse digital copies. To prevent such document plagiarism, a local document fingerprint program is running, but it is easy to compare against the entire document, but it is difficult to detect partial copies. So, the algorithm used is winnowing. The winnowing algorithm can select a fingerprint from the hash of the k-gram and distinguish a copy through the spacing between the fingerprints and the consecutive hashes. It works well with MOSS, a service that mainly detects plagiarism in programming tasks, and by reporting the overlap ratio between the two documents, it can be seen that this service has dramatically reduced many plagiarism cases now.