CS6220-01 Data Mining Techniques – Fall 2016

Assignment 4

Qian Cao

**Part 1 Filter Method**

**Show the result:**

Part A: Features listed in descending order according to the |r| value

Feature 5 has an |r| of 0.436921797517

Feature 14 has an |r| of 0.368269040809

Feature 15 has an |r| of 0.368223721497

Feature 17 has an |r| of 0.366025114237

Feature 8 has an |r| of 0.352141261364

Feature 23 has an |r| of 0.351349925535

Feature 27 has an |r| of 0.341042614938

Feature 2 has an |r| of 0.308810814577

Feature 21 has an |r| of 0.299049007432

Feature 32 has an |r| of 0.290782911347

Feature 35 has an |r| of 0.266092789733

Feature 3 has an |r| of 0.195732390555

Feature 29 has an |r| of 0.156904332677

Feature 26 has an |r| of 0.153095989944

Feature 20 has an |r| of 0.137636220589

Feature 18 has an |r| of 0.113944729764

Feature 33 has an |r| of 0.0931737325674

Feature 9 has an |r| of 0.0877730096245

Feature 1 has an |r| of 0.0697950519202

Feature 11 has an |r| of 0.0568764889229

Feature 22 has an |r| of 0.0566051682402

Feature 12 has an |r| of 0.0421168839866

Feature 34 has an |r| of 0.0388096478213

Feature 7 has an |r| of 0.0352947753379

Feature 16 has an |r| of 0.0314779448057

Feature 36 has an |r| of 0.0308552371644

Feature 30 has an |r| of 0.0208294541251

Feature 19 has an |r| of 0.017931425398

Feature 28 has an |r| of 0.0156062347576

Feature 10 has an |r| of 0.0130053709987

Feature 4 has an |r| of 0.00921358145379

Feature 31 has an |r| of 0.00895519480113

Feature 25 has an |r| of 0.00777974308474

Feature 24 has an |r| of 0.0055078660728

Feature 13 has an |r| of 0.00217858402781

Feature 6 has an |r| of 9.81377865178e-05

**Question: Why would one be interested in the absolute value of r rather than the raw value?**

**Answer:** For Cov(X,Y) = E[(X-E[X])(Y-E[Y])] = E[XY]-E[X]E[Y], we are only interested in the difference between X and Xmean, Y and Ymean. The negative sign is not important here, which only shows one of X, Y is bigger than its mean and the other is smaller.

Part B: Values of m and Avg LOOCV accuracy

M: 1, LOOCV Accuracy: 596/846, 70.4% Correctly Classified

M: 2, LOOCV Accuracy: 673/846, 79.6% Correctly Classified

M: 3, LOOCV Accuracy: 696/846, 82.3% Correctly Classified

M: 4, LOOCV Accuracy: 704/846, 83.2% Correctly Classified

M: 5, LOOCV Accuracy: 707/846, 83.6% Correctly Classified

M: 6, LOOCV Accuracy: 707/846, 83.6% Correctly Classified

M: 7, LOOCV Accuracy: 729/846, 86.2% Correctly Classified

M: 8, LOOCV Accuracy: 740/846, 87.5% Correctly Classified

M: 9, LOOCV Accuracy: 757/846, 89.5% Correctly Classified

M: 10, LOOCV Accuracy: 749/846, 88.5% Correctly Classified

M: 11, LOOCV Accuracy: 763/846, 90.2% Correctly Classified

M: 12, LOOCV Accuracy: 765/846, 90.4% Correctly Classified

M: 13, LOOCV Accuracy: 752/846, 88.9% Correctly Classified

M: 14, LOOCV Accuracy: 754/846, 89.1% Correctly Classified

M: 15, LOOCV Accuracy: 769/846, 90.9% Correctly Classified

M: 16, LOOCV Accuracy: 759/846, 89.7% Correctly Classified

M: 17, LOOCV Accuracy: 770/846, 91.0% Correctly Classified

M: 18, LOOCV Accuracy: 782/846, 92.4% Correctly Classified

M: 19, LOOCV Accuracy: 776/846, 91.7% Correctly Classified

**M: 20, LOOCV Accuracy: 783/846, 92.6% Correctly Classified**

M: 21, LOOCV Accuracy: 771/846, 91.1% Correctly Classified

M: 22, LOOCV Accuracy: 765/846, 90.4% Correctly Classified

M: 23, LOOCV Accuracy: 757/846, 89.5% Correctly Classified

M: 24, LOOCV Accuracy: 748/846, 88.4% Correctly Classified

M: 25, LOOCV Accuracy: 743/846, 87.8% Correctly Classified

M: 26, LOOCV Accuracy: 744/846, 87.9% Correctly Classified

M: 27, LOOCV Accuracy: 741/846, 87.6% Correctly Classified

M: 28, LOOCV Accuracy: 736/846, 87.0% Correctly Classified

M: 29, LOOCV Accuracy: 723/846, 85.5% Correctly Classified

M: 30, LOOCV Accuracy: 719/846, 85.0% Correctly Classified

M: 31, LOOCV Accuracy: 709/846, 83.8% Correctly Classified

M: 32, LOOCV Accuracy: 706/846, 83.5% Correctly Classified

M: 33, LOOCV Accuracy: 699/846, 82.6% Correctly Classified

M: 34, LOOCV Accuracy: 700/846, 82.7% Correctly Classified

M: 35, LOOCV Accuracy: 686/846, 81.1% Correctly Classified

M: 36, LOOCV Accuracy: 699/846, 82.6% Correctly Classified

**Question: Which value of m gives the highest LOOCV classification accuracy, and what is the value of this optimal accuracy?**

**Answer:** M: 20, LOOCV Accuracy: 783/846, 92.6%

**Part 2 Wrapper method**

Iteration 0, Selected Features : {} LOOCV Accuracy: 0/846

Iteration 1 Selected Features : { 21 } LOOCV Accuracy: 634/846

Iteration 2 Selected Features : { 21 11 } LOOCV Accuracy: 711/846

Iteration 3 Selected Features : { 21 11 19 } LOOCV Accuracy: 752/846

**Part 3**

{ce, bcdf, ae, abc, d, adf, cdef, ace, ad, bcf}

C1

|  |  |
| --- | --- |
| itemset | sup |
| a | 5 |
| b | 3 |
| c | 6 |
| d | 5 |
| e | 4 |
| f | 4 |

L1

|  |  |
| --- | --- |
| itemset | sup |
| a | 5 |
| b | 3 |
| c | 6 |
| d | 5 |
| e | 4 |
| f | 4 |

C2

|  |  |
| --- | --- |
| {ab} | 1 |
| {ac} | 2 |
| {ad} | 2 |
| {ae} | 2 |
| **{bc}** | **3** |
| {bd} | 1 |
| {bf} | 2 |
| {cd} | 2 |
| **{ce}** | **3** |
| **{cf}** | **3** |
| {de} | 1 |
| **{df}** | **3** |
| {ef} | 1 |

L2

|  |  |
| --- | --- |
| **{bc}** | **3** |
| **{ce}** | **3** |
| **{cf}** | **3** |
| **{df}** | **3** |

bce, bcf, cef, cdf

C3

|  |  |
| --- | --- |
| {bcf} | 2 |
| {cdf} | 2 |
| {cef} | 1 |

L3-empty

Maximal Frequent Itemset: {bc}, {ce}, {cf}, {df}

Choose {ce}:

c->e: s=0.3, c=0.5

e->c: s=0.3, c=0.75

Total number of transactions is 10; minsup is 3, so s is always 0.3

The maximal sup in L1 is 6 so the minimal confidence is larger than 3/6=0.5

**Part 4**

a:8 b:6 c:6 d:4 e:4 f:1

for minsup=2, only choose a,b,c,d,e

|  |  |  |
| --- | --- | --- |
| TID | Item bought | Frequent Items |
| 1 | {a,b,e} | {a,b,e} |
| 2 | {a,b,c,d} | {a,b,c,d} |
| 3 | {a,c,d} | {a,c,d} |
| 4 | {a,c,e} | {a,c,e} |
| 5 | {b,c,f} | {b,c} |
| 6 | {a} | {a} |
| 7 | {a,b,c} | {a,b,c} |
| 8 | {b,d,e} | {b,d,e} |
| 9 | {a,c} | {a,c} |
| 10 | {a,b,d,e} | {a,b,d,e} |

The FP-tree is on next page.

