**Homework 3**

Auston Schendzielos (partner: Spencer Hanson)

1.

|  |  |
| --- | --- |
| **Highest Fold Ratio:** | **How Many Have It** |
| 800 | 34 |

|  |  |
| --- | --- |
| **Range** | **Count** |
| Val <= 2 | 411 |
| 2 < Val <= 4 | 285 |
| 8 < Val <= 16 | 942 |
| 16 < Val <= 32 | 3251 |
| 32 < Val <= 64 | 1129 |
| 64 < Val <= 128 | 611 |
| 128 < Val <= 256 | 258 |
| 256 < Val <= 512 | 112 |
| 512 < Val | 71 |
| **Total** | 7070 |

2.

b. (***Include answer in report***) List the top 50 genes with the highest S2N ratio along with their S2N values.

|  |  |  |
| --- | --- | --- |
| Top 50 S2N: |  |  |
| 1 | M55150\_at | 1.467641165 |
| 2 | X95735\_at | 1.444530875 |
| 3 | U50136\_rna1\_at | 1.421708308 |
| 4 | U22376\_cds2\_s\_at | 1.339307881 |
| 5 | M81933\_at | 1.204041677 |
| 6 | M16038\_at | 1.203217545 |
| 7 | M84526\_at | 1.201419811 |
| 8 | M23197\_at | 1.19597434 |
| 9 | U82759\_at | 1.19255629 |
| 10 | Y12670\_at | 1.184736877 |
| 11 | D49950\_at | 1.143703904 |
| 12 | M27891\_at | 1.133426544 |
| 13 | X59417\_at | 1.12463678 |
| 14 | X52142\_at | 1.122589116 |
| 15 | M28170\_at | 1.116755638 |
| 16 | X17042\_at | 1.105975049 |
| 17 | U05259\_rna1\_at | 1.103966367 |
| 18 | Y00787\_s\_at | 1.081994777 |
| 19 | M96326\_rna1\_at | 1.077189938 |
| 20 | U12471\_cds1\_at | 1.069730905 |
| 21 | U46751\_at | 1.064078049 |
| 22 | M80254\_at | 1.044394599 |
| 23 | M92287\_at | 1.043055722 |
| 24 | L13278\_at | 1.04203194 |
| 25 | U09087\_s\_at | 1.036256865 |
| 26 | L08246\_at | 1.03478374 |
| 27 | X74262\_at | 1.027773107 |
| 28 | M62762\_at | 1.023393307 |
| 29 | M31211\_s\_at | 1.022881216 |
| 30 | M28130\_rna1\_s\_at | 1.001285867 |
| 31 | D26156\_s\_at | 0.989097871 |
| 32 | M63138\_at | 0.983628011 |
| 33 | M31523\_at | 0.971227716 |
| 34 | M57710\_at | 0.967686169 |
| 35 | X15949\_at | 0.960090273 |
| 36 | M69043\_at | 0.958859986 |
| 37 | S50223\_at | 0.956733251 |
| 38 | U32944\_at | 0.954874872 |
| 39 | M81695\_s\_at | 0.95306583 |
| 40 | L47738\_at | 0.951380054 |
| 41 | M83652\_s\_at | 0.947504094 |
| 42 | X85116\_rna1\_s\_at | 0.94637181 |
| 43 | M11147\_at | 0.945755455 |
| 44 | Z15115\_at | 0.945199785 |
| 45 | M21551\_rna1\_at | 0.941980587 |
| 46 | M19045\_f\_at | 0.938075874 |
| 47 | X04085\_rna1\_at | 0.930498848 |
| 48 | L49229\_f\_at | 0.920745023 |
| 49 | X14008\_rna1\_f\_at | 0.914954098 |
| 50 | M91432\_at | 0.913520798 |

c. (***Include answer in report***) List the top 50 genes with the highest T-value along with their T-values.

|  |  |  |  |
| --- | --- | --- | --- |
| Top 50 T-Value: |  |  |  |
| 1 | M55150\_at |  | 8.091951183 |
| 2 | U22376\_cds2\_s\_at |  | 7.904300374 |
| 3 | X59417\_at |  | 6.803106346 |
| 4 | U50136\_rna1\_at |  | 6.435952103 |
| 5 | M31211\_s\_at |  | 6.294206502 |
| 6 | M28170\_at |  | 6.253971105 |
| 7 | U82759\_at |  | 6.243020468 |
| 8 | M92287\_at |  | 6.217365336 |
| 9 | U09087\_s\_at |  | 6.182424806 |
| 10 | U05259\_rna1\_at |  | 6.175126475 |
| 11 | M81933\_at |  | 6.164964999 |
| 12 | U12471\_cds1\_at |  | 6.146298795 |
| 13 | D26156\_s\_at |  | 6.097112519 |
| 14 | L13278\_at |  | 6.021341675 |
| 15 | X74262\_at |  | 6.016389062 |
| 16 | S50223\_at |  | 5.858625771 |
| 17 | X52142\_at |  | 5.833144157 |
| 18 | X95735\_at |  | 5.727643095 |
| 19 | L47738\_at |  | 5.715039288 |
| 20 | M31523\_at |  | 5.677543403 |
| 21 | Z15115\_at |  | 5.615471093 |
| 22 | M11147\_at |  | 5.588644932 |
| 23 | D49950\_at |  | 5.588063309 |
| 24 | X63469\_at |  | 5.583296318 |
| 25 | D38073\_at |  | 5.562983485 |
| 26 | U32944\_at |  | 5.442991882 |
| 27 | U26266\_s\_at |  | 5.42520538 |
| 28 | J05243\_at |  | 5.406254456 |
| 29 | L49229\_f\_at |  | 5.398203084 |
| 30 | X15949\_at |  | 5.389656222 |
| 31 | X17042\_at |  | 5.388623496 |
| 32 | M21551\_rna1\_at |  | 5.377292024 |
| 33 | M31303\_rna1\_at |  | 5.370570645 |
| 34 | Y08612\_at |  | 5.36665273 |
| 35 | U20998\_at |  | 5.36123935 |
| 36 | AF012024\_s\_at |  | 5.333337212 |
| 37 | X56411\_rna1\_at |  | 5.327391771 |
| 38 | Y12670\_at |  | 5.324928168 |
| 39 | U29175\_at |  | 5.322715508 |
| 40 | M91432\_at |  | 5.300882197 |
| 41 | HG1612-HT1612\_at |  | 5.276868788 |
| 42 | M13792\_at |  | 5.222356203 |
| 43 | D63874\_at |  | 5.206860895 |
| 44 | U72342\_at |  | 5.206761353 |
| 45 | X97267\_rna1\_s\_at |  | 5.203003458 |
| 46 | X76648\_at |  | 5.195203157 |
| 47 | U35451\_at |  | 5.144391195 |
| 48 | Z69881\_at |  | 5.143815953 |
| 49 | D63880\_at |  | 5.125101526 |
| 50 | M62762\_at |  | 5.122153112 |

d. (***Include answer in report***) How many genes are common between top 50 genes selected using S2N and those selected using T-value (i.e. exist in BOTH lists)? List them **(Answer: there should 31 genes in common)**

**The highlighted genes above are the genes that both lists have in common. There are 31.**

3.

Results for ALL using every gene

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ALL Precision** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.85 | 0.722222222 | 0.869565217 | 0.80952381 |
| k=5 | 0.85 | 0.722222222 | 0.869565217 | 0.80952381 |
| k=7 | 0.85 | 0.722222222 | 0.869565217 | 0.80952381 |
| k=9 | 0.85 | 0.722222222 | 0.869565217 | 0.80952381 |
| k=11 | 0.85 | 0.722222222 | 0.869565217 | 0.80952381 |
| **ALL Recall** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.62962963 | 0.619047619 | 0.666666667 | 0.653846154 |
| k=5 | 0.62962963 | 0.619047619 | 0.666666667 | 0.653846154 |
| k=7 | 0.62962963 | 0.619047619 | 0.666666667 | 0.653846154 |
| k=9 | 0.62962963 | 0.619047619 | 0.666666667 | 0.653846154 |
| k=11 | 0.62962963 | 0.619047619 | 0.666666667 | 0.653846154 |
| **ALL F1 Measures** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.723404255 | 0.666666667 | 0.754716981 | 0.723404255 |
| k=5 | 0.723404255 | 0.666666667 | 0.754716981 | 0.723404255 |
| k=7 | 0.723404255 | 0.666666667 | 0.754716981 | 0.723404255 |
| k=9 | 0.723404255 | 0.666666667 | 0.754716981 | 0.723404255 |
| k=11 | 0.723404255 | 0.666666667 | 0.754716981 | 0.723404255 |

Results for AML using every gene

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML Precision** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.15 | 0.277777778 | 0.130434783 | 0.19047619 |
| k=5 | 0.15 | 0.277777778 | 0.130434783 | 0.19047619 |
| k=7 | 0.15 | 0.277777778 | 0.130434783 | 0.19047619 |
| k=9 | 0.15 | 0.277777778 | 0.130434783 | 0.19047619 |
| k=11 | 0.15 | 0.277777778 | 0.130434783 | 0.19047619 |
| **AML Recall** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.428571429 | 0.384615385 | 0.75 | 0.5 |
| k=5 | 0.428571429 | 0.384615385 | 0.75 | 0.5 |
| k=7 | 0.428571429 | 0.384615385 | 0.75 | 0.5 |
| k=9 | 0.428571429 | 0.384615385 | 0.75 | 0.5 |
| k=11 | 0.428571429 | 0.384615385 | 0.75 | 0.5 |
| **AML F1 Measures** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.222222222 | 0.322580645 | 0.222222222 | 0.275862069 |
| k=5 | 0.222222222 | 0.322580645 | 0.222222222 | 0.275862069 |
| k=7 | 0.222222222 | 0.322580645 | 0.222222222 | 0.275862069 |
| k=9 | 0.222222222 | 0.322580645 | 0.222222222 | 0.275862069 |
| k=11 | 0.222222222 | 0.322580645 | 0.222222222 | 0.275862069 |

Results for ALL using Significant Genes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ALL Precision** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.75 | 0.771428571 | 0.72972973 | 0.714285714 |
| k=5 | 0.75 | 0.771428571 | 0.72972973 | 0.714285714 |
| k=7 | 0.75 | 0.771428571 | 0.72972973 | 0.714285714 |
| k=9 | 0.75 | 0.771428571 | 0.72972973 | 0.714285714 |
| k=11 | 0.75 | 0.771428571 | 0.72972973 | 0.714285714 |
| **ALL Recall** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.931034483 | 0.9 | 0.964285714 | 0.961538462 |
| k=5 | 0.931034483 | 0.9 | 0.964285714 | 0.961538462 |
| k=7 | 0.931034483 | 0.9 | 0.964285714 | 0.961538462 |
| k=9 | 0.931034483 | 0.9 | 0.964285714 | 0.961538462 |
| k=11 | 0.931034483 | 0.9 | 0.964285714 | 0.961538462 |
| **ALL F1 Measures** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.830769231 | 0.830769231 | 0.830769231 | 0.819672131 |
| k=5 | 0.830769231 | 0.830769231 | 0.830769231 | 0.819672131 |
| k=7 | 0.830769231 | 0.830769231 | 0.830769231 | 0.819672131 |
| k=9 | 0.830769231 | 0.830769231 | 0.830769231 | 0.819672131 |
| k=11 | 0.830769231 | 0.830769231 | 0.830769231 | 0.819672131 |

Results for AML using Significant Genes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AML Precision** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.25 | 0.228571429 | 0.27027027 | 0.285714286 |
| k=5 | 0.25 | 0.228571429 | 0.27027027 | 0.285714286 |
| k=7 | 0.25 | 0.228571429 | 0.27027027 | 0.285714286 |
| k=9 | 0.25 | 0.228571429 | 0.27027027 | 0.285714286 |
| k=11 | 0.25 | 0.228571429 | 0.27027027 | 0.285714286 |
| **AML Recall** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 1 | 1 | 1 | 0.833333333 |
| k=5 | 1 | 1 | 1 | 0.833333333 |
| k=7 | 1 | 1 | 1 | 0.833333333 |
| k=9 | 1 | 1 | 1 | 0.833333333 |
| k=11 | 1 | 1 | 1 | 0.833333333 |
| **AML F1 Measures** | Euclidean | Chebyshev | Manhattan | Cosine Similarity |
| k=3 | 0.4 | 0.372093023 | 0.425531915 | 0.425531915 |
| k=5 | 0.4 | 0.372093023 | 0.425531915 | 0.425531915 |
| k=7 | 0.4 | 0.372093023 | 0.425531915 | 0.425531915 |
| k=9 | 0.4 | 0.372093023 | 0.425531915 | 0.425531915 |
| k=11 | 0.4 | 0.372093023 | 0.425531915 | 0.425531915 |

The above table shows the F1 measures where k=11 of ALL for the first 4, and then AML for the second 4. As shown in the tables, the k values did not have an effect on the predictions.

Best Predictor: Any value of k, and the Manhattan proximites were the best indicators for ALL and AML in the preprocessed data. For the original data set, the Manhattan was best for the ALL, and the Chebyshev was the best for the AML.

Overall

Ultimately, the preprocessed data clearly outperformed the original dataset. However, getting the preprocessed data costs more time than the original.