

Data Mining Project 1: Data Preprocessing

Report

Kiran Nanjundaswamy
U00833551

The input file for this report is pldata.txt file with K= 15

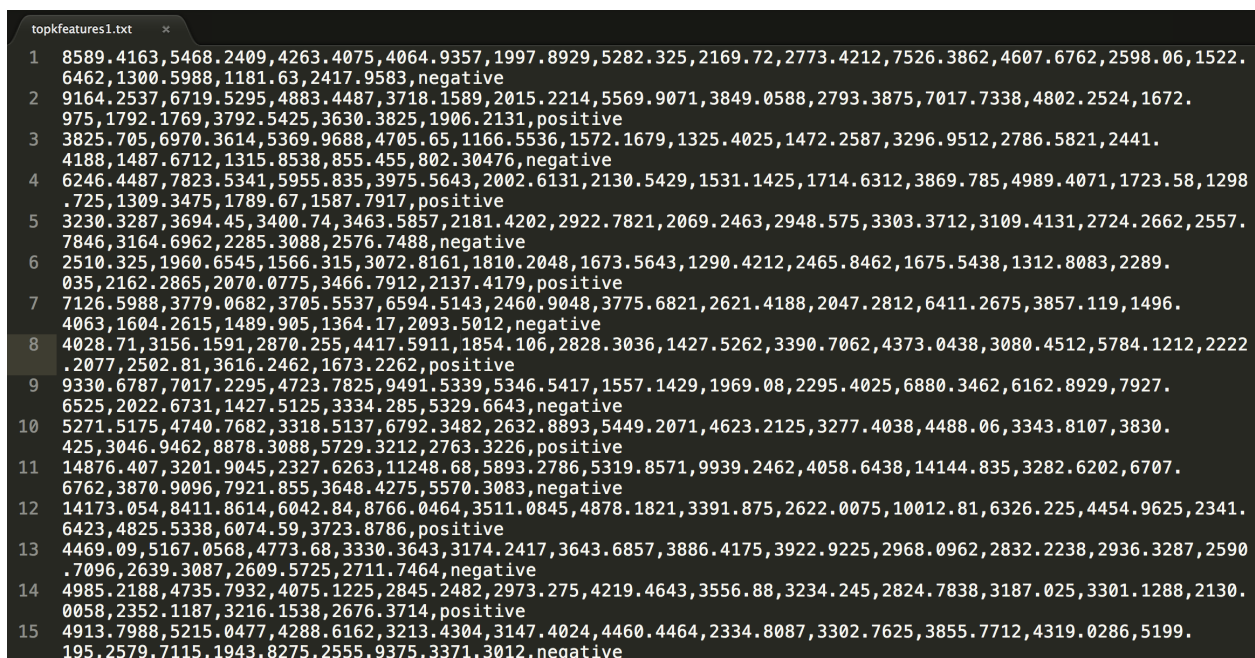
Task 1: Use a method implemented in weka to select K features.

For this task I had to import the pldata.txt and use weka features to extract K features. Since weka does not allow for text files to be used along with its packages, I converted the .txt file to .csv file in java and used weka.core.converters package of weka to convert the file to .arff (file format which is understood by weka)

Next I had to extract K features and produce topkfeatures1.txt file. Since there was no specification as to the method of filtering the data I have used REMOVE Attribute filter.

Remove filter will accept a range of values which are to be removed from the data set and remove them. It does not do any further changes to the data, nor does it reorder them in any way. I felt this was a good feature to work with as it returns a part of the RAW dataset which was inputted.

Below is the snapshot of the text file output which was produced by Remove filter.



```
topkfeatures1.txt
1 8589.4163,5468.2409,4263.4075,4064.9357,1997.8929,5282.325,2169.72,2773.4212,7526.3862,4607.6762,2598.06,1522.6462,1300.5988,1181.63,2417.9583,negative
2 9164.2537,6719.5295,4883.4487,3718.1589,2015.2214,5569.9071,3849.0588,2793.3875,7017.7338,4802.2524,1672.975,1792.1769,3792.5425,3630.3825,1906.2131,positive
3 3825.705,6970.3614,5369.9688,4705.65,1166.5536,1572.1679,1325.4025,1472.2587,3296.9512,2786.5821,2441.4188,1487.6712,1315.8538,855.455,802.30476,negative
4 6246.4487,7823.5341,5955.835,3975.5643,2002.6131,2130.5429,1531.1425,1714.6312,3869.785,4989.4071,1723.58,1298.725,1309.3475,1789.67,1587.7917,positive
5 3230.3287,3694.45,3400.74,3463.5857,2181.4202,2922.7821,2069.2463,2948.575,3303.3712,3109.4131,2724.2662,2557.7846,3164.6962,2285.3088,2576.7488,negative
6 2510.325,1960.6545,1566.315,3072.8161,1810.2048,1673.5643,1290.4212,2465.8462,1675.5438,1312.8083,2289.035,2162.2865,2070.0775,3466.7912,2137.4179,positive
7 7126.5988,3779.0682,3705.5537,6594.5143,2460.9048,3775.6821,2621.4188,2047.2812,6411.2675,3857.119,1496.4063,1604.2615,1489.905,1364.17,2093.5012,negative
8 4028.71,3156.1591,2870.255,4417.5911,1854.106,2828.3036,1427.5262,3390.7062,4373.0438,3080.4512,5784.1212,2222.2077,2502.81,3616.2462,1673.2262,positive
9 9330.6787,7017.2295,4723.7825,9491.5339,5346.5417,1557.1429,1969.08,2295.4025,6880.3462,6162.8929,7927.6525,2022.6731,1427.5125,3334.285,5329.6643,negative
10 5271.5175,4740.7682,3318.5137,6792.3482,2632.8893,5449.2071,4623.2125,3277.4038,4488.06,3343.8107,3830.425,3046.9462,8878.3088,5729.3212,2763.3226,positive
11 14876.407,3201.9045,2327.6263,11248.68,5893.2786,5319.8571,9939.2462,4058.6438,14144.835,3282.6202,6707.6762,3870.9096,7921.855,3648.4275,5570.3083,negative
12 14173.054,8411.8614,6042.84,8766.0464,3511.0845,4878.1821,3391.875,2622.0075,10012.81,6326.225,4454.9625,2341.6423,4825.5338,6074.59,3723.8786,positive
13 4469.09,5167.0568,4773.68,3330.3643,3174.2417,3643.6857,3886.4175,3922.9225,2968.0962,2832.2238,2936.3287,2590.7096,2639.3087,2609.5725,2711.7464,negative
14 4985.2188,4735.7932,4075.1225,2845.2482,2973.275,4219.4643,3556.88,3234.245,2824.7838,3187.025,3301.1288,2130.0058,2352.1187,3216.1538,2676.3714,positive
15 4913.7988,5215.0477,4288.6162,3213.4304,3147.4024,4460.4464,2334.8087,3302.7625,3855.7712,4319.0286,5199.195,2579.7115,1943.8275,2555.9375,3371.3012,negative
```

Task 2: Implement entropy based method to discretize all genes and to select the top K genes with highest information gain, listed in decreasing information gain order. Use 3 bins for each gene for this task

For this task we were not supposed to use weka packages and had to be implemented in java. To find the information gain, I divided each gene into 2 halves by calculating $(\text{highestValue} - \text{lowestValue})/2$. Then selected a split point in each half of the Gene attribute. I computed the information gain with each split and found the best split for the Gene which gave me the highest information gain.

This was assigned to the gene attribute as its information gain and we have 3 bins to split the data. To discretize the data I validated each value in the Gene against each bin to see which bin the value fell into. The 3 bins were assigned the values 'a', 'b' & 'c'. This was saved in entropydata.txt file.

The information gain for each gene was sorted. For the top K genes the info gain value along with the Bins was stored in entropybin.txt file and only top K genes were stored in topkfeatures2.txt file. Below are the snapshots of each of the files.

Entropybins.txt

```
entropybins.txt
1
2 G249: Info Gain: 0.261276 ; Bins: (-INF,1380.510], 6.0, 37.0; (1380.510, 2864.446], 9.0, 3.0; (2864.446,+INF], 7.0, 0.0;
3 G765: Info Gain: 0.246840 ; Bins: (-INF,800.751], 10.0, 40.0; (800.751, 1640.884], 8.0, 0.0; (1640.884,+INF], 4.0, 0.0;
4 G1772: Info Gain: 0.235159 ; Bins: (-INF,86.625], 21.0, 12.0; (86.625, 183.019], 1.0, 22.0; (183.019,+INF], 0.0, 6.0;
5 G267: Info Gain: 0.196975 ; Bins: (-INF,577.409], 5.0, 29.0; (577.409, 1254.397], 4.0, 9.0; (1254.397,+INF], 13.0, 2.0;
6 G493: Info Gain: 0.183695 ; Bins: (-INF,495.090], 8.0, 37.0; (495.090, 1050.693], 8.0, 2.0; (1050.693,+INF], 6.0, 1.0;
7 G377: Info Gain: 0.174188 ; Bins: (-INF,173.417], 0.0, 14.0; (173.417, 422.202], 12.0, 23.0; (422.202,+INF], 10.0, 3.0;
8 G245: Info Gain: 0.170802 ; Bins: (-INF,591.159], 6.0, 32.0; (591.159, 1287.403], 5.0, 6.0; (1287.403,+INF], 11.0, 2.0;
9 G822: Info Gain: 0.166058 ; Bins: (-INF,1146.514], 7.0, 34.0; (1146.514, 2329.416], 7.0, 5.0; (2329.416,+INF], 8.0, 1.0;
10 G1582: Info Gain: 0.158480 ; Bins: (-INF,125.227], 22.0, 22.0; (125.227, 263.489], 0.0, 11.0; (263.489,+INF], 0.0, 7.0;
11 G1892: Info Gain: 0.156959 ; Bins: (-INF,215.588], 11.0, 38.0; (215.588, 319.469], 6.0, 2.0; (319.469,+INF], 5.0, 0.0;
12 G138: Info Gain: 0.153160 ; Bins: (-INF,928.664], 14.0, 8.0; (928.664, 2019.048], 8.0, 19.0; (2019.048,+INF], 0.0, 13.0;
13 G1423: Info Gain: 0.151195 ; Bins: (-INF,543.406], 9.0, 36.0; (543.406, 1102.854], 5.0, 3.0; (1102.854,+INF], 8.0, 1.0;
14 G43: Info Gain: 0.146353 ; Bins: (-INF,1273.648], 9.0, 5.0; (1273.648, 2954.281], 13.0, 20.0; (2954.281,+INF], 0.0, 15.0;
15 G652: Info Gain: 0.144748 ; Bins: (-INF,541.555], 21.0, 19.0; (541.555, 1130.683], 1.0, 14.0; (1130.683,+INF], 0.0, 7.0;
16 G66: Info Gain: 0.139064 ; Bins: (-INF,1479.408], 5.0, 30.0; (1479.408, 2075.326], 7.0, 6.0; (2075.326,+INF], 10.0, 4.0;
```

Topkfeatures2.txt

```
topkfeatures2.txt
1 552.65875,389.765,87.71,471.16731,624.42625,282.27857,475.27885,560.58375,137.39,51.045,519.07875,309.49375,1926.3962,354.97375,976.26625,negative
2 2314.9488,1779.9875,53.1475,1712.2096,1342.9675,761.37857,1648.4596,4421.815,78.20875,237.20125,505.0675,1644.4262,1288.305,326.11875,1933.5863,positive
3 137.97125,164.7325,20.735,196.43654,338.46,217.13333,209.21923,442.3925,95.155,19.395,409.10375,333.625,535.72625,47.57375,415.31125,negative
4 987.79875,723.40125,27.58,597.73846,660.61,321.62024,576.64038,1325.3925,36.2775,39.16875,558.35875,973.16875,901.29,188.55875,438.51625,positive
5 584.7325,361.3125,129.9575,434.2,147.90625,104.64167,464.30962,340.535,82.85375,59.29625,1037.6375,124.02,2451.5912,518.66375,817.34125,negative
6 2030.8775,1386.8225,56.0275,1472.2481,387.2425,196.59524,1453.3442,452.80625,77.24625,116.4075,434.575,325.045,1022.6538,312.80125,1815.515,positive
7 228.28625,154.055,79.39125,199.05192,149.5225,195.65,214.58269,143.92375,99.305,39.99625,1371.3825,126.43,1501.8213,314.7875,737.7075,negative
8 3886.9925,1781.3175,45.74625,1731.3596,598.4975,388.40714,1849.6173,1253.5438,54.2325,19.58375,696.59375,541.5325,884.3625,192.6625,1074.2125,positive
```

Entropydata.txt

```
entropydata.txt
1 |a,a,b,a,b,b,a,a,b,a,a,b,a,a,negative
2 |b,c,a,c,c,c,c,c,a,b,a,c,b,a,b,positive
3 |a,a,a,a,a,b,a,a,a,a,a,a,a,a,negative
4 |a,a,a,b,b,b,a,b,a,a,a,b,a,a,a,positive
5 |a,a,b,a,a,a,a,a,a,a,b,a,b,a,a,negative
6 |b,b,a,c,a,b,c,a,a,a,a,a,a,b,positive
7 |a,a,a,a,a,b,a,a,a,a,b,a,b,a,a,negative
8 |c,c,a,c,b,b,c,b,a,a,a,a,a,a,positive
9 |a,a,b,a,a,a,a,a,b,a,c,a,c,c,a,negative
10 |c,c,a,c,b,c,c,b,a,b,a,b,b,a,c,positive
11 |a,a,c,a,a,c,a,a,c,a,c,a,c,c,a,negative
12 |c,c,a,c,c,c,c,c,a,c,b,c,b,a,b,positive
13 |a,a,b,a,a,b,a,a,a,a,b,a,b,a,a,negative
14 |b,b,a,c,c,b,c,b,a,a,a,c,b,a,c,positive
15 |a,a,b,b,a,b,b,a,a,a,b,a,b,b,b,negative
16 |a,a,a,a,b,b,a,a,a,a,a,b,a,c,positive
17 |a,a,a,a,b,a,a,a,a,b,a,c,a,a,negative
18 |a,a,a,a,b,c,a,a,a,a,a,a,a,c,positive
19 |a,a,b,a,a,a,a,a,a,b,a,b,b,a,negative
```

Task3: Compute the correlation coefficients between the numerical-valued genes selected by Task1 and the genes select by Task2. Let the K^2 pairs of genes be in increasing order of similarity

For this task I used the values of each gene attributed computed by Task1 and Task2. To find the correlation, I am using a package called PearsonsCorrelation which will accept an array of x and y values and return the correlation coefficient. This was stored in the file correlationgenesdata.txt. Since we have taken $K=15$, The result file has 225 records.

```
correlationgenes.txt
1 |Correlation Coefficient of G15 & G1423 : -0.18418554756313538
2 |Correlation Coefficient of G11 & G377 : -0.1761195282555163
3 |Correlation Coefficient of G11 & G822 : -0.17234161551643676
4 |Correlation Coefficient of G11 & G1892 : -0.16795249239546234
5 |Correlation Coefficient of G15 & G822 : -0.16472767953627854
6 |Correlation Coefficient of G11 & G1423 : -0.15472298591870595
7 |Correlation Coefficient of G12 & G1423 : -0.1429007927312393
8 |Correlation Coefficient of G5 & G1423 : -0.14241291565600256
9 |Correlation Coefficient of G5 & G822 : -0.13672646981484643
10 |Correlation Coefficient of G12 & G822 : -0.13251238331227816
11 |Correlation Coefficient of G11 & G493 : -0.13078061246669936
12 |Correlation Coefficient of G15 & G249 : -0.11631364698776857
13 |Correlation Coefficient of G15 & G1892 : -0.10857936483035845
14 |Correlation Coefficient of G15 & G245 : -0.10518912695607792
15 |Correlation Coefficient of G5 & G249 : -0.10289923124464238
16 |Correlation Coefficient of G9 & G66 : -0.10077904631678948
17 |Correlation Coefficient of G15 & G765 : -0.0979106822392443
18 |Correlation Coefficient of G11 & G249 : -0.0964900880046042
19 |Correlation Coefficient of G3 & G249 : -0.09385154237730803
```