

PROJECT DESCRIPTION:

Used GenomeTrainXY.txt to select 100 top-ranked genes based on f-test.

The f-test scores and the feature number (the line/feature number) is in TASK A.

TASK A:

Line number/feature number

707	inf
1564	inf
1870	127.668
4419	78.73881
4422	54.0313
3843	53.91094
4423	51.48979
2331	47.71333
738	47.39209
554	46.54245
118	45.78737
449	43.0127
89	42.45087
4030	39.51113
3562	37.41871
4369	36.43285
1142	36.10056
3748	35.98197
3407	35.50416
3442	35.21791
4182	35.19634
1381	34.91252
3275	34.82786
1895	34.07687
2766	33.69881
2127	33.09798
729	32.85991
2119	32.75626
619	32.65714
1775	32.1163
367	32.09961
3733	31.91542
1742	31.90604
628	31.69984

1907	31.50802
739	31.29053
3992	31.22396
3012	30.88887
4240	30.63713
56	30.49268
4424	30.41241
247	29.48935
2872	28.96412
85	28.86434
3845	28.84202
4009	28.77841
3299	28.59984
819	28.53358
2851	28.3424
277	28.30707
4279	28.30043
1223	28.23506
1978	28.06689
3044	27.86331
2427	27.70718
57	27.69494
2650	27.53566
202	27.39622
3311	27.37212
4425	27.24292
3337	27.22874
677	27.09534
1788	26.90586
2961	26.75032
1541	26.71111
1331	26.45016
70	26.06083
3074	25.99718
4421	25.98259
1955	25.60176
3448	25.57576
3554	25.51221
2609	25.42107
4157	25.1362
523	25.06119
3572	25.03471
312	24.77466
543	24.38024
393	24.32577

818	24.29688
1346	24.20278
1532	24.17139
3712	24.11453
4281	24.06737
124	23.9366
1845	23.91557
1647	23.76417
4265	23.75219
1896	23.69781
3985	23.54923
2974	23.53965
2878	23.49616
4433	23.3867
2732	23.3739
155	23.28198
3399	23.25837
993	23.20272
36	23.19291
2546	23.11643
579	22.98083

TASK B:

Used the above selected genes as the features, train the four classifiers

- SVM linear kernel
- linear regression
- KNN (k=3)
- centroid method

TASK C:

KNN Predicted Labels: [4, 1, 4, 3, 2, 1, 3, 4, 4, 1]

Centroid Predicted Labels: [4. 1. 4. 3. 2. 1. 3. 4. 4. 2.]

SVM Predicted Labels: [4. 1. 4. 4. 1. 1. 4. 4. 4. 1.]

LR Predicted Labels: [3.23658372 1.72891677 3.92335127 3.35910325
2.39582737 1.270337 2.52180066 2.81010537 3.52413477
2.26299648]