Project 5 - Feature Selection

Blair Gemmer

CSCI 548 - Pattern Recognition

Spring 2013

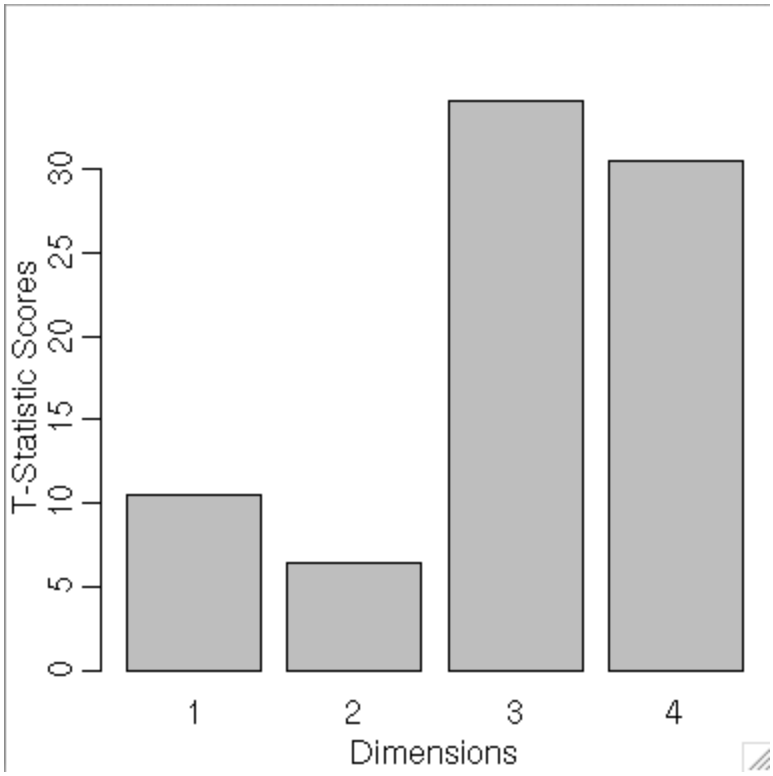

Figure 1. Iris Dataset t-statistic scores by dimension.

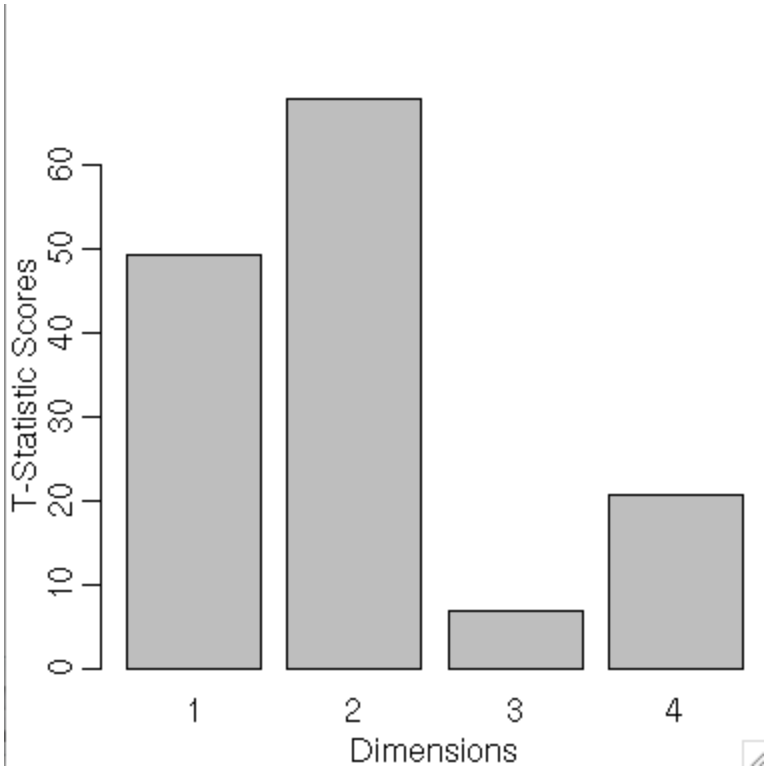

Figure 2. Fruit Dataset t-statistic scores by dimension.

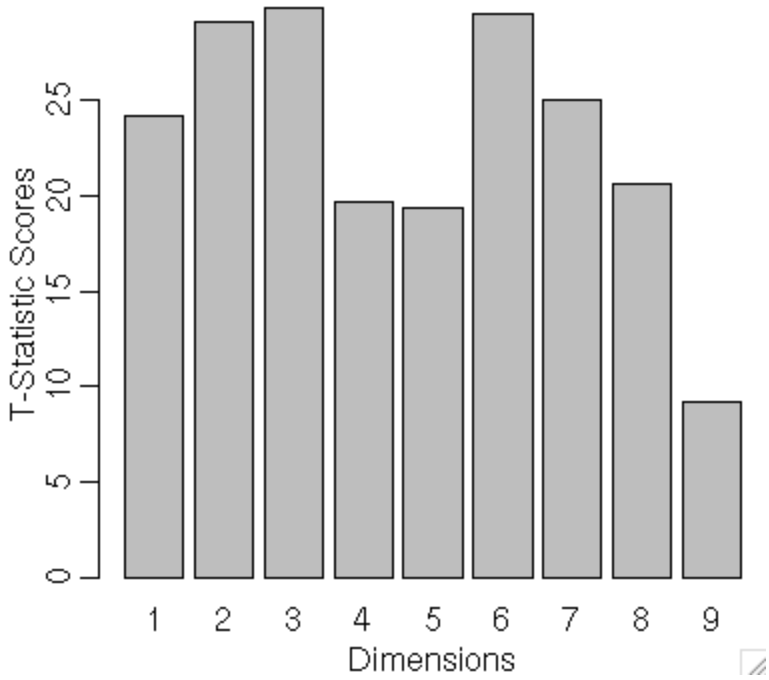

Figure 3. Tumor Dataset t-statistic scores by dimension.

Dataset	Complete Dataset	Reduced Dataset	# of Reduced	Reduced Dimensions
Iris	96.37%	94.74%	2	{3,4}
Fruit	93.2%	95.2%	3	{1,2,4}
Tumor	94.86%	97.71%	6	{1,2,3,6,7,8}

Figure 4. Comparison of performance of 3 datasets, based on all dimensions vs. reduced dimensions after optimization by picking top performing dimensions. Found top performers by scoring with t-test and using a threshold to determine the top set of dimensions. Performance was based on Linear Discriminant Analysis on the full dataset and then on the reduced dataset. The 4th column represents the number of reduced dimensions that we used after LDA. Threshold = 20 for all 3 experiments. Used 75% of each dataset to train LDA.

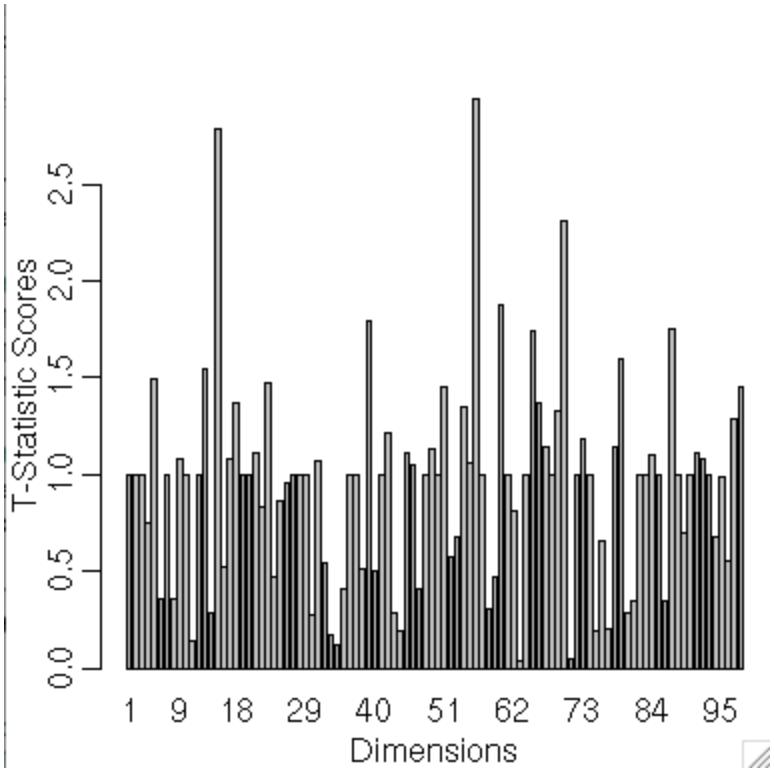

Figure 5. Mouse Dataset t-statistic scores by dimension.

Fold	Performance
1	100%
2	0%
3	100%
4	100%
5	0%
6	100%
7	0%
8	100%
9	100%
10	100%
11	100%
12	100%
13	0%
14	100%
15	100%
16	100%
17	100%
18	0%
19	100%
20	100%
Final Average Performance	**75%**

Figure 6. Performance of the mouse dataset, using Linear Discriminant Analysis after performing the Cross-Validation Leave One Out algorithm, which drops one of the elements from each experiment (rather than a set percentage), then performs cross-validation, and finally LDA on the reduced dataset. Threshold = 1.78. Final Average performance was 75%.

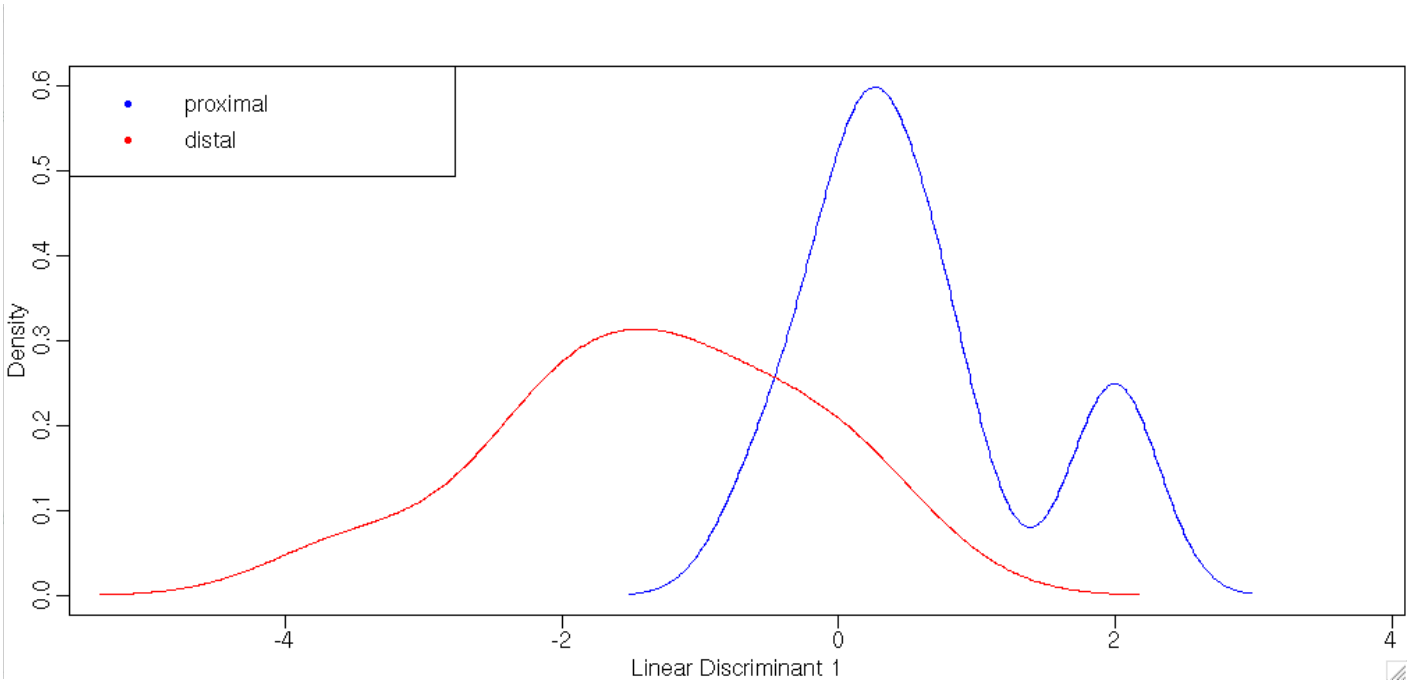

Figure 7. Distributions of Proximal and Distal Mouse Metagenomic Data Projected Onto First Linear Discriminant. LDA performed using all data as training set and just those dimensions selected using T-test approach. Just those dimensions were projected onto the Linear Discriminant for visualization purposes. The classifier operates at 60% accuracy when using leave-one-out cross-validation.

Dataset	Complete Dataset	Reduced Dataset
Iris		
J1	7.53	16.28
J2	40.42	22.23
J3	31.83	19.38
Fruit		
J1	1.19	2.68
J2	32.98	31.7
J3	21.7	21.5
Tumor		
J1	2.3	2.54
J2	5.999	5.988
J3	5.11	5.05

Figure 8. Performance of 3 datasets using the J1, J2, and J3 divergence measures. These 3 divergence measures use Within-Class Scatter Matrix (Sw), the Between-Class Scatter Matrix (Sb), and the Global Covariance Matrix (Sm) to find a score for each of the 3 datasets. Threshold for all 3 datasets was 20. Training sets were 75% of original datasets.

Dataset	T-Test Score (Full Dataset)	T-Test Score (Reduced)	Floating Search Score	Dimensions Used	Full J-Scores	Reduced J-Scores
Iris	96.37%	94.74%	94.37%	{3,4}	{7.53, 40.42, 31.83}	{16.28, 22.23, 19.38}
Fruit	93.2%	95.2%	94.8%	{1,2,4}	{1.19, 32.98, 21.7}	{2.68, 31.7, 21.5}
Tumor	94.86%	97.71%	97.71%	{1,2,3,6,7,8}	{2.3, 5.999, 5.11}	{2.54, 5.988, 5.05}
Mouse	75%	75%	NA	{15,39,56,60,70}	NA	NA
Sperm		92%	80%	{1,2,4,6,7}	{1.01, 1.04, .13}	{1.02, 1.07, .12}

Figure 9. Performance of 3 datasets using Floating Search Algorithm vs. T-Test Score and then running Linear Discriminant Analysis. The right column displays the dimensions used by both algorithms before running LDA. Also included is the performance of the Mouse genomic dataset using Leave-One-Out Cross-Validation vs. Floating Search Algorithm. The performance of Floating Search was impeded by too many zero values in the mouse dataset. The last row indicates the Sperm Fertility dataset from WHO 2010. Threshold of Iris, Fruit, and Tumor was 20, while Mouse was 1.78, and Sperm was 1. Training sets were 75% of original datasets.

Final Notes

If you are having any problem displaying the plots (since there are so many displaying simultaneously), I wrote the different datasets in separate functions that can be commented out so you can run each one separately. This also helps with reading the console output, as there is a lot of information!