Machine learning for thyroid cancer diagnosis

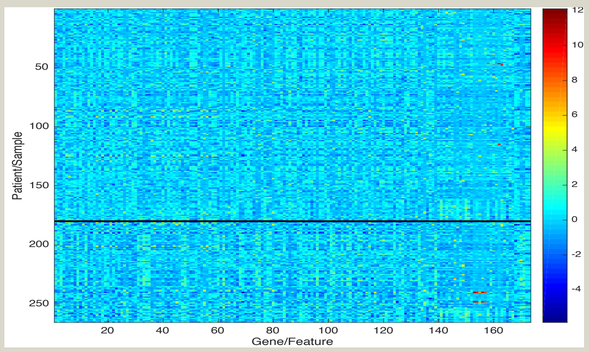
Rajiv Krishnakumar, Akash V. Maharaj, Raghu Mahajan
Department of Physics, Stanford University
CS229, Fall 2015

Abstract

We study the usefulness of gene-expression data in the diagnosis of thyroid cancer. Using logistic regression and support vector machines, we develop a classifier which gives similar performance to the currently best-known classifier, but uses a lot fewer features.

Structure of the Data

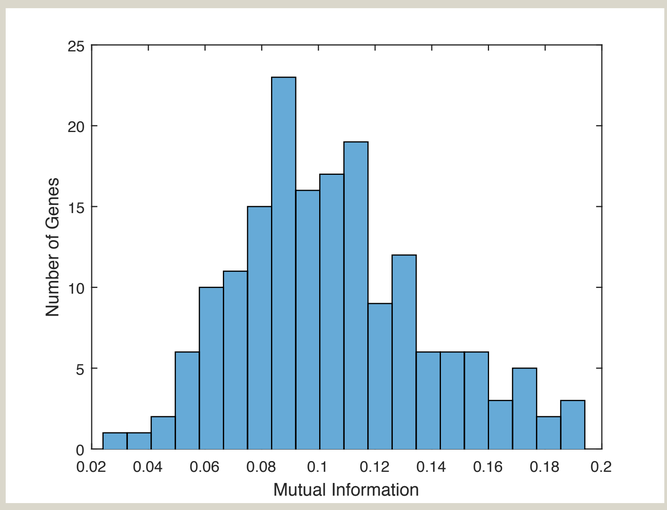
Number of patients in dataset	367
Result of biopsy (cytology)	
Benign	55
Malignant	47
Indeterminate	**265**
Post operative diagnosis of indeterminate tumors (ground truths)	
Benign	**180**
Malignant	**85**
Number of features (gene expressions)	
Raw data set	~ 25000
Pre-normalized and selected	**173**



Color map visualization of the 173 gene intensities for each of the 265 patients.

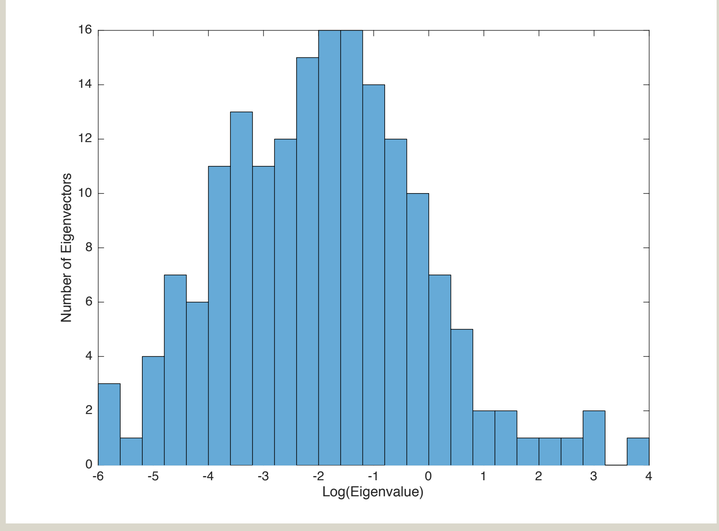
Mutual Information

With 173 features and 265 samples, smart feature selection is crucial to avoid overfitting. We computed the Mutual Information score of each gene to select the most relevant ones.



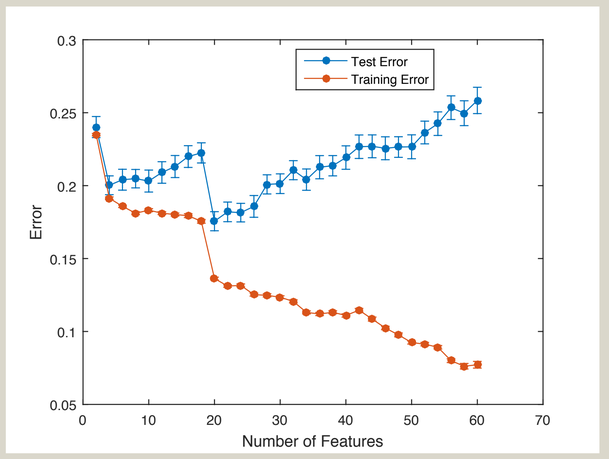
Principal Components Analysis

To have more confidence in our feature selection, we also did a principal components analysis (PCA). This allows us to work with more general linear subspaces within the feature space. **The predictive power of the resulting classifier is similar regardless of whether we use the mutual information statistic or PCA to select the subset of genes.**



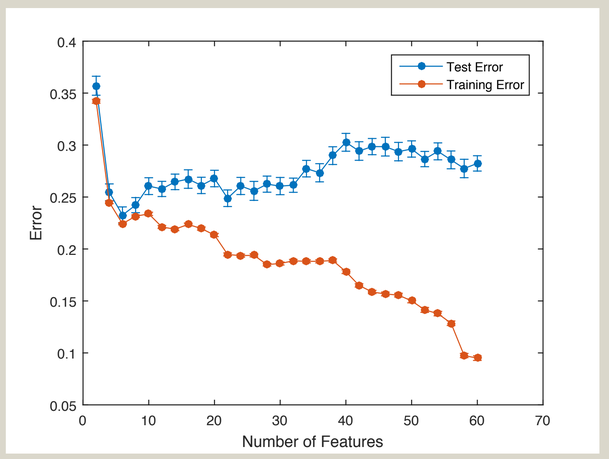
Logistic Regression

We trained different classifiers by tweaking the number of features kept and also the relative weights assigned to malignant and benign samples. It is very essential to correctly diagnose the patients who have a malignant tumor, so we impose a big penalty for misclassifying the malignant samples.



A plot of the test error and training error vs. the number of features chosen; 1:1 weighting; with features ordered using PCA.

With 8 features, we get a 38% error on malignant samples (bad) and a 9% error on benign samples.



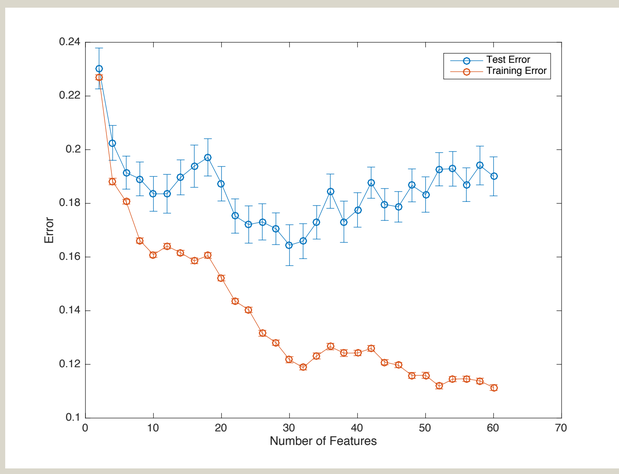
Similar plot as above but with a 1:4 weighting.

With 8 features, we get a 11% error on malignant samples and a 29% error on the benign samples.

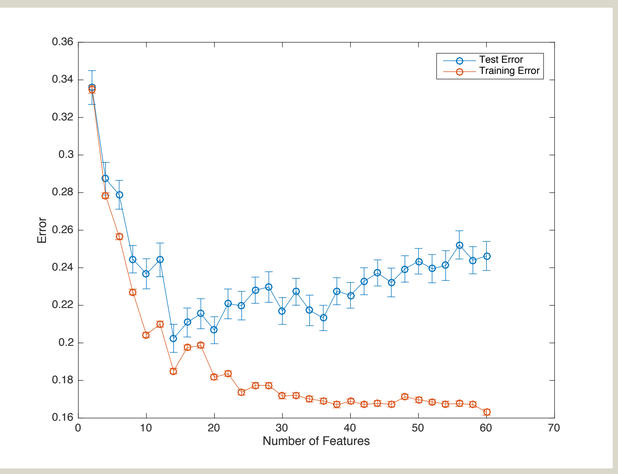
This performance is comparable to what is presented in Reference [1].

Support Vector Machines

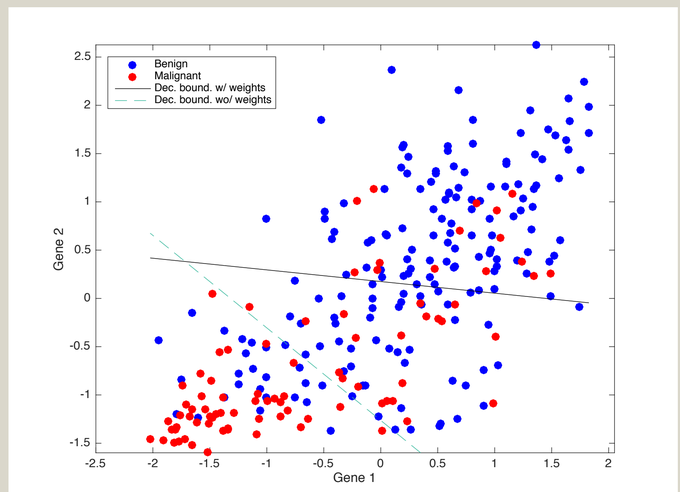
We repeat the classification using support vector machines, varying the number of features kept and the weighting (cost) parameter.



Learning curve for linear SVM 1:1 weighting.



Learning curve for linear SVM 1:4 weighting. The weighted SVM gives a 12% error on the malignant samples and a 22% error on the benign samples.



An illustration of the (linear) SVM decision boundary when training on the two top ranked genes (via Mutual Information).

Adding weights to penalize malignant misclassifications moves the decision boundary significantly.

References

[1] Alexander, Erik K, Kennedy, Giulia C, Baloch, Zubair W, Cibas, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. ***New England Journal of Medicine***, 367(8):705–715, 2012

[2] Howlader, N, Noone, AM, Krapcho, M, Neyman, N, Aminou, et al. Seer cancer statistics review, 1975–2008. ***Bethesda, MD: National Cancer Institute***, 19, 2011

[3] OnlineRef, 2012. URL http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34289

Thanks to Kevin Brennan and Olivier Gevaert @ Stanford Center for Biomedical Informatics Research