How Many Genes Are Needed for a Discriminant Microarray Data Analysis?

Wentian Li and Yaning Yang

Lab of Statistical Genetics

Rockefeller University

[poster, CAMDA'00, Dec 18-19, 2000]

contact information:
wli@linkage.rockefeller.edu
http://linkage.rockefeller.edu/wli/

Abstract

The analysis of the second data set (from the MIT group) is a discriminant analysis or supervised learning. Among thousands of genes whose expression level is measured, not all are needed for discriminant analysis: a gene may either not contribute to the separation of two types of tissues/cancers, or it may be redundant because it is highly correlated with other genes. There are two theoretical frameworks in which variable selection (or gene selection in our case) can be addressed. The first is model selection, and the second is model averaging. We have carried out model selection using Akaike information criterion and Bayesian information criterion within logistic regression (discrimination, predictor, classifier) to determine the number of genes that provide the best model. These model selection criteria set an upper limit of 22-25 and 12-13 genes for the sample size of this data (38), and the best model consists of only one (no.4847, zyxin) or two genes. We have also carried out model averaging over the best single-gene logistic predictors using three different weights: maximized likelihood, prediction rate on training set, and equal weight. We have observed that the performance of these weighted predictors on the testing set is gradually reduced as more genes are included, but a clear cutoff that separates good and bad prediction performance is not found.

Notations: $\{x_i\}$ (i=1,2, ... 7129) (log) mRNA expression level of gene i. y cancer type (0 for ALL - acute lymphoblastic leukemia, 1 for AML, acute myeloid leukemia).

Discriminator/Classifier: logistic regression

$$P(y=1) = \frac{1}{1+e^{-a_0-\sum_i a_i x_i}}$$

Model selection: $\hat{L}$ for maximized likelihood, K is the number of parameters in the model (number of genes + 1), N (=38) the sample size, we select the model with the minimum

$$AIC = -2\log(\hat{L}) + 2K, or$$

$$BIC = -2\log(\hat{L}) + \log(N)K$$

Model averaging: single-gene discriminators are averaged with weight $\{w_j\}$ to obtain the final discriminator/classifier:

$$P(y=1) = \sum_j w_j \left(\frac{1}{1+e^{-a_0-a_j x_j}}\right)$$

Two Null Models

random guess of y with a 0.5 probability: $L = 0.5^N$, $-2\log(L) = 52.68$, AIC=BIC=52.68. No parameter is used in the model ($K = 0$).

random guess of y with an estimated probability: estimated P(y=1) is 11/38. $L = (11/38)^{11}(27/38)^{27}$, $-2\log(L) = 45.73$, AIC= 47.73, BIC= 49.37. One parameter is estimated.

these two null models can set a limit on the number of genes used:

$$0 + 2(p+1) < 52.68/47.73 \quad \text{AIC}$$

$$0 + \log(38)(p+1) < 52.68/49.365 \quad \text{BIC}$$

which leads to

$$p < 25.34/22.86 \quad \text{AIC}$$

$$p < 13.48/12.57 \quad \text{BIC}$$

by BIC, the number of genes used should not be more than 12 for these 38 sample points

results in AIC

type	K	-2log($\widehat{L}$)	AIC	△AIC
#1 g4847 (zyxin)	2	≈ 0	4.000	0
#2 g1882 (CST3 cystatin C)	2	6.973	10.973	6.973
#3 g3320 (leukotriene..)	2	10.914	14.914	10.914
#4 g5039 (LEPR leptin rec)	2	11.355	15.355	11.355
#5 g6218 (ELA2 elastatse 2)	2	11.459	15.459	11.459
#6 g2020 (FAH ..)	2	12.103	16.103	12.103
#7 g1834 (CD33 antigen)	2	12.226	16.226	12.226
#8 g760 (cystatin A)	2	13.104	17.104	13.104
#9 g1745 (LYN v-yes-1..)	2	13.151	17.151	13.15
#10 g5772 (c-myb)	2	14.723	18.723	14.723
#100 g2833(AF1q)	2	27.215	31.215	27.21
#200 g3312(ATR)	2	30.841	34.841	30.841
g1834+g2267	3	**0.004**	6.004	2.004
g5039+g5772	3	**0.008**	6.008	2.008
sum of top 2 (g4847+g1882)	3	**0.029**	6.029	2.029
sum of top 5	6	**0.011**	12.011	8.011
sum of top 10	11	**0.002**	22.002	18.002
sum of top 22	23	**0.001**	46.001	42.001
sum of top 37	38	**0.001**	76.001	72.001
fixed prob	1	45.728	47.728	43.728
random guess	0	52.679	52.679	48.679

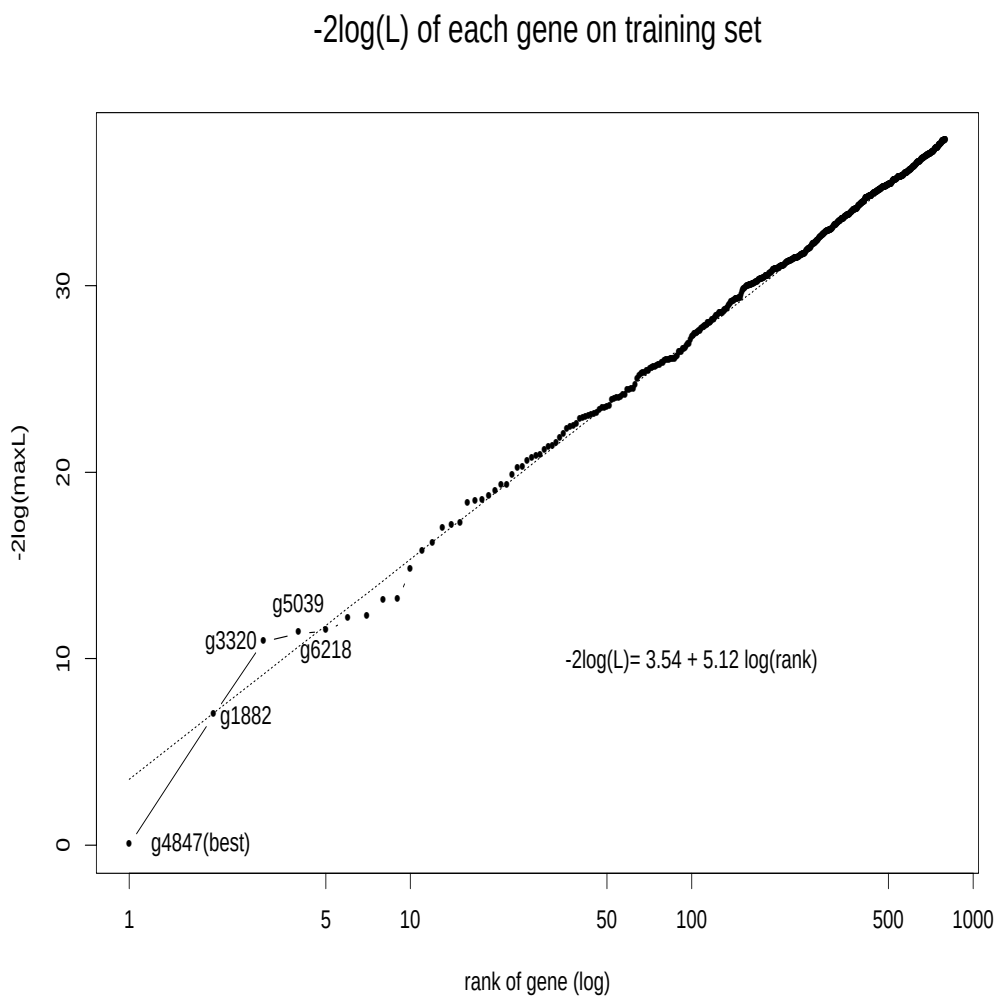
results in BIC and prediction rates

type	BIC	△BIC	p_{train}	p_{test}
#1 g4847 (zyxin)	7.275	0	**38/38**	31/34
#2 g1882 (CST3 cystatin C)	14.248	6.973	36/38	**32/34**
#3 g3320 (leukotriene..)	18.190	10.915	35/38	27/34
#4 g5039 (LEPR leptin rec)	18.630	11.355	36/38	22/34
#5 g6218 (ELA2 elastatse 2)	18.734	11.459	34/38	22/34
#6 g2020 (FAH ..)	19.378	12.103	36/38	25/34
#7 g1834 (CD33 antigen)	19.501	12.226	35/38	31/34
#8 g760 (cystatin A)	20.379	13.104	35/38	**32/34**
#9 g1745 (LYN v-yes-1..)	20.426	13.151	33/38	28/34
#10 g5772 (c-myb)	21.998	14.723	35/38	27/34
#100 g2833(AF1q)	34.490	27.215	30/38	28/34
#200 g3312(ATR)	38.117	30.842	29/38	21/34
g1834+g2267	10.917	3.642	**38/38**	22/34
g5039+g5772	10.921	3.646	**38/38**	26/34
sum of top 2 (g4847+g1882)	10.942	3.667	**38/38**	32/34
sum of top 5	21.837	14.562	**38/38**	24/34
sum of top 10	40.016	32.741	**38/38**	31/34
sum of top 22	83.666	76.391	**38/38**	27/34
sum of top 37	138.229	130.954	**38/38**	21/34
fixed prob	49.365	42.090	27/38	20/34
random guess	52.679	45.404	19/38	17/34

Model Selection Results

Models with 37, 22, 10, 5, 2 genes all fit the training set perfectly (prediction rate is 100%). By the model selection criterion, models with less number of parameters are selected if the data-fitting performance is the same. This leads to the logistic regression model with **one gene (gene 4847, zyxin**, "a component of adhesion plaques that has been suggested to perform regulatory functions at these specialized regions of the plasma membrane"). Models with **two genes** are also rather good.

What about other genes besides g4847?



Zipf's law

$-2\log(\hat{L})$ vs. log(rank) is a straight line, or $\hat{L}$ vs. rank (r) is a power-law function:

$$\hat{L} \sim \frac{1}{r^{\alpha}}$$

where $\alpha \approx 2.56$. This type of rank-frequency plot is studied extensively by George Zipf (1902-1950), though many of Zipf's original plots have $\alpha \approx 1$.

More information on the Zipf's law can be found at *http://linkage.rockefeller.edu/wli/zipf/*

Are training and testing sets consistent?

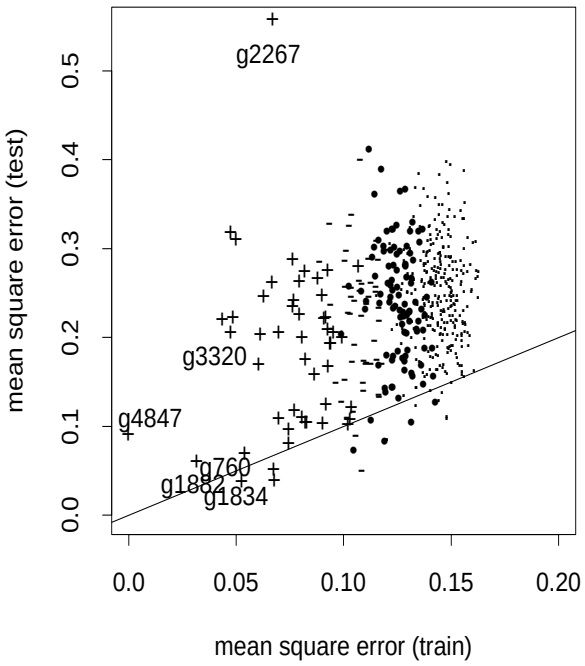
x: error in training set
y: error in testing set

left: mean square error
right: 0/1 error (prediction error)

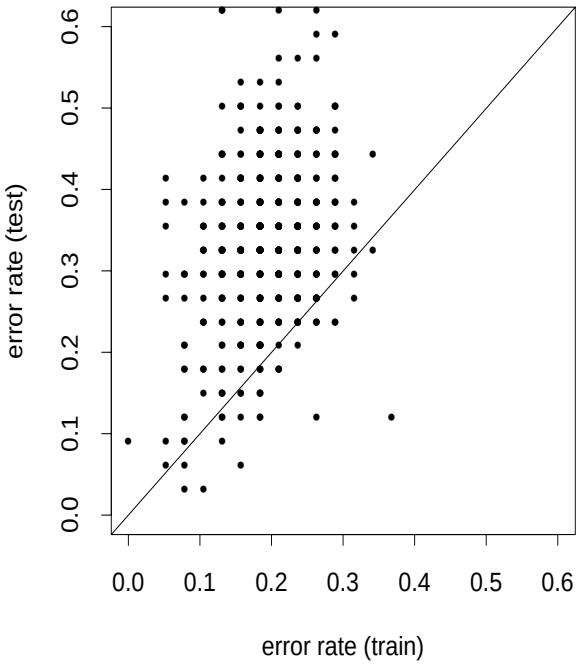
diagonal line: same error in training and in testing set

g1882 and g760 are doing better (than g4847) on the testing set

test vs train (mean square error)



test vs train (0/1 error)



How to choose weights in model averaging?

(1) Maximum likelihood weight: $w_j \propto \widehat{L_j}$

Bayesian framework suggests that the weight should be proportional to $exp(-BIC/2)$, or the maximum likelihood $\widehat{L}$ if all models have the same number of parameters, because (D for available data, $\tilde{D}$ for new data):

$$p(\tilde{D}|D) = \sum_M p(\tilde{D}|M)p(M|D) = \sum_M p(\tilde{D}|M) \cdot \frac{p(D|M)p(M)}{p(D)}$$

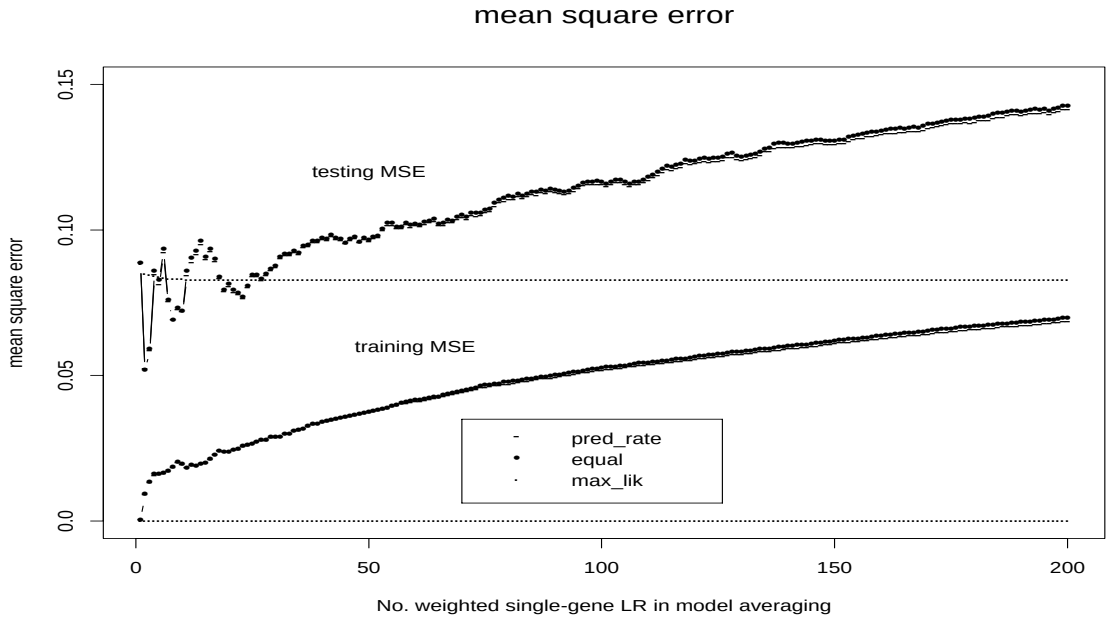
$$\propto \sum_M p(\tilde{D}|M) \cdot e^{-BIC/2} \propto \sum_M p(\tilde{D}|M) \cdot \hat{L}_M$$

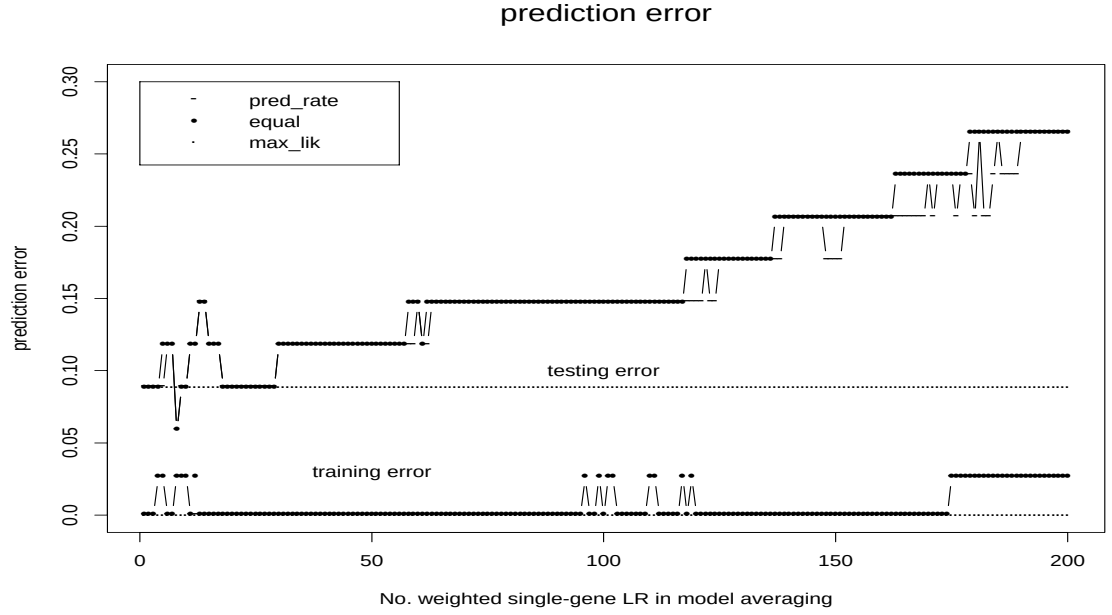
(2) Prediction rate weight: $w_j \propto p_{train}$

(3) Equal weight: $w_j \propto 1$

Error as a function of number of models being averaged

upper graph: mean square error; lower graph: 0/1 error
upper half: testing set; lower half: training set
dot: maximum L weight; dash: prediction rate weight; star: equal weight





Effective number of genes in model averaging

Because of the weight, adding one model (one single-gene logistic regression) does not mean this gene is fully used. The number of models added can be infinity, but the "effective" number of genes is finite. One definition of the effective number of genes is to treat the first dominant model as contributing 1 gene, the second contributing w_2/w_1 genes, etc.

In the maximum likelihood weighting scheme, for our data set, $\{w_j/w_1\}$ are 1, 0.031, 0.0043, 0.0034, 0.0032, 0.0024, 0.0022, 0.0014, 0.006... The sum of the top ten terms is $\sum_{j=1}^{10} w_j/w_1 \approx 1.05$. In other words, **the effective number of genes in this model averaging is less than 2.**

Conclusion

How many genes are needed ... in this dataset?

in model selection framework

A: one or two genes

in model averaging framework

A: any number of single-gene predictor can be averaged, but the effective number of genes is less than 2 in the maximum-likelihood / Bayesian / Akaike weighting scheme.