
Feature Selection for High-Dimensional Genomic Microarray Data

Eric P. Xing, Michael I. Jordan, & Richard M. Karp
@ Division of Computer Science, UC Berkeley

Presented by Degui Zhi @ UCSD
Apr 30th 2002

Microarray – biology in 1 slide


- Human genome contains ~30K genes
- Not all genes are expressed at same time
- A **microarray** is a systematic way to test the expression levels of thousands of genes in a single experiment
 - A “snapshot” or a state vector of a cell/tissue
- Multiple microarrays: compare expression levels under different conditions:
 - Binary (normal vs. diseased; treated vs. untreated)
 - Multi-class (different types of cancers, populations)
 - Continuous (time course; dose response)

Experiments: Leukemia data set

- Data are drawn from Golub et al. 1999.
- 7130 genes in a microarray (6817 in Golub's)

Leukemia Type	Training set	Test set	Total
ALL	27 (71%)	20 (59%)	47
AML	11	14	25
Total	38	34	72

- Training set: same tissue, same age, and same lab
- Test set: different tissue, different age, and different lab

72 experiments  7130 genes

3

Prediction by Golub et al.

- **Feature selection:** rank by Pearson correlation to class label
 - 1100 genes with significant correlation
 - 50 gene with highest correlation used for classification
- **Classification:** simple linear classifier, $\text{sign}(\mathbf{w} \cdot \mathbf{x})$
 - Training: cross validation success for 36 of the 38 samples
 - Test: success for 29 of the 34 independent samples
 - The other samples are uncertain due to lack of significance
 - Predictors made of top 10-200 genes all can be trained to make no mistake.

4

Feature selection

- The concept to learn is $F \rightarrow \{0,1\}$
- $|F|$ is too large, so we want to find a small but informative subset $G \subseteq F$ and learn $G \rightarrow \{0,1\}$ instead

2 popular approaches for feature selection

- Wrapper
 1. Find a feature subset G ;
 2. Optimizing the classifier C for G , measure the error $\epsilon(C(G))$
 3. Find $G = \operatorname{argmin} \epsilon(C(G))$
- Filter
 - Find a feature subset G independent of any classifier C .

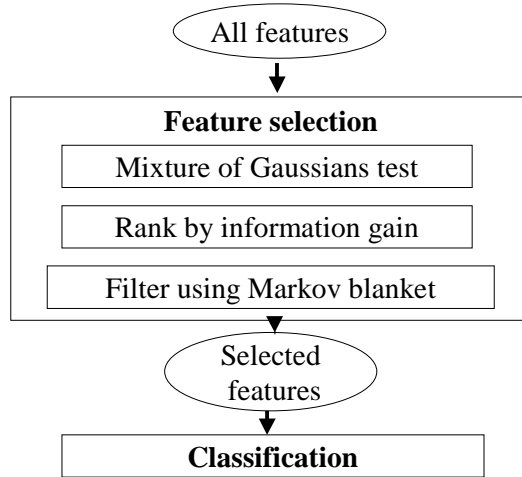
7

Heuristic for feature selection

Biological knowledge (assumption)	Feature selection filter
Gene expression is 'on' or 'off'	Testing of bimodal distribution
Not all genes respond to a single event	Ranking by information gain
Genes are highly redundant	Filter using Markov blanket

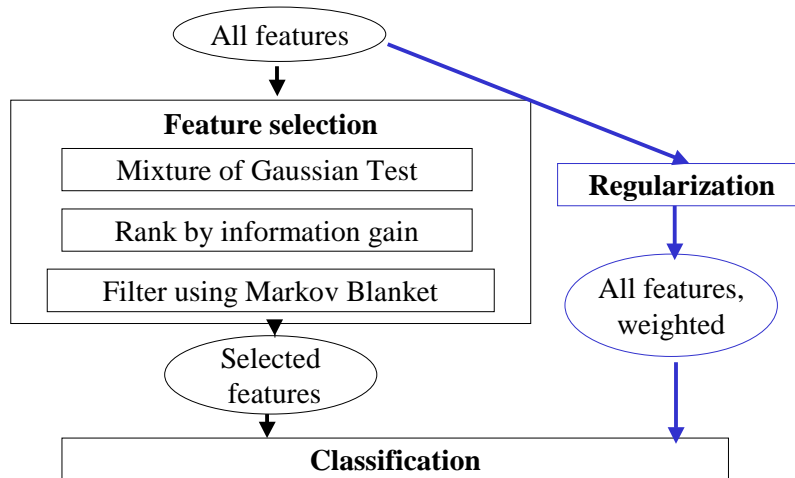
8

Outline of procedure



9

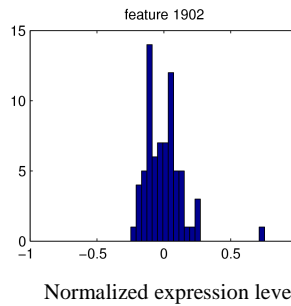
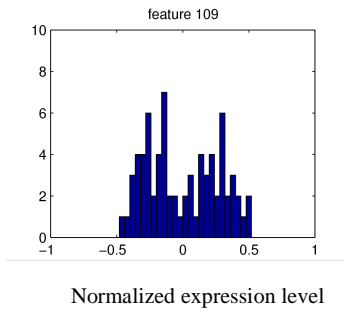
Outline of procedure



10

Feature modeling: bimodal

- Heuristic: a feature with discriminative power should have bimodal distribution
- A simple bimodal model: mixture of 2 univariate Gaussians



11

Gaussian mixtures

- For a feature F , we have measurements $\mathbf{f}=\{f_1, \dots, f_N\}$
- Mixture of K univariate Gaussians with parameter $\boldsymbol{\theta} = \{(\pi_k, \mu_k, \sigma_k)\}$ $1 \leq k \leq K$ π_k is class prior
- The likelihood of f_n to the k -th Gaussians is:

$$P(f_n | \boldsymbol{\theta}, k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{(f_n - \mu_k)^2}{2\sigma_k^2}\right\}$$

and

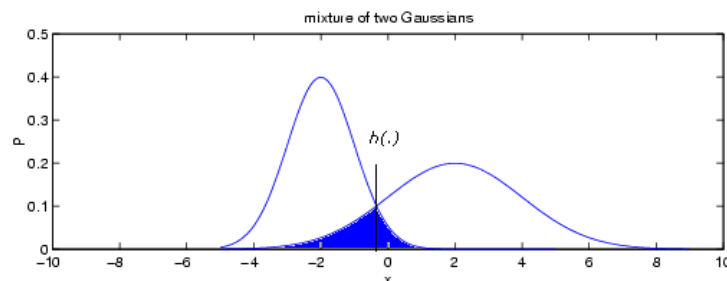
$$P(f_n | \boldsymbol{\theta}) = \sum_k \pi_k P(f_n | \boldsymbol{\theta}, k)$$

- Learn $\boldsymbol{\theta}$ from the sample data using EM

12

Mixture overlap

- The mixture overlap ϵ is the minimal error achievable by any classifier $h(\cdot)$ on this EM-trained Gaussian mixture model
- ϵ can be used as a measure of the discriminability of feature F
- $h(\cdot)$ can be used to quantize continuous value f_i , which is used for later filters that are based on information theory.

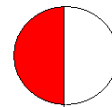


13

Feature selection via information gain

- For a *reference partition* $Q = \{S_0, S_1\}$ of the training set S
- Entropy of this partition

$$H(Q) = - \sum_{c \in \{0,1\}} P(S_c) \log P(S_c)$$



- Test on feature F induces partition $E = \{E_0, E_1\}$.
- Partition Q projected onto E_k forms subpartition

$$Q_k = \{S_c \cap E_k, c \in \{0,1\}\}$$

- Entropy of Q_k

$$H(Q_k) = - \sum_{c=1}^C P(S_c | E_k) \log P(S_c | E_k)$$



14

Ranking by information gain

- The information gain due to F w.r.t. the reference partition is




$$I(Q|E) = H(Q) - H(Q|E)$$

$$= H(Q) - \sum_{k=1}^K (P(E_k)H(Q_k))$$

$$H(\text{red}) = -2 * [1/2 * \log(1/2)] = 1$$

$$H(\text{red/blue}) = (1/2) * H(\text{red}) + (1/2) * H(\text{blue}) = 0.9188$$

$$H(\text{red/blue}) = 1 \quad H(\text{red}) = 0$$

- Ranking by infogain:  <  < 

15

Redundant feature

Let \mathbf{F} be the full feature set and $\mathbf{G} \subseteq \mathbf{F}$. Let C be the class label.

- A feature F_i is redundant in \mathbf{G} if the classification results are same with or without it.

- That is conditional independence:

$$P(C|\mathbf{G} - \{F_i\}) = P(C|\mathbf{G})$$

for all values of the features in \mathbf{G} .

- To be more precise, if there is a subset $\mathbf{M} \subseteq \mathbf{G}$, F_i not in \mathbf{M} , but

$$P(C|\mathbf{M} - \{F_i\}) = P(C|\mathbf{M})$$

- \mathbf{M} is called a **Markov blanket** of F_i

16

Markov blanket filtering

Proposition 2

For a complete feature set \mathbf{F} , let \mathbf{G} be a subset of \mathbf{F} , and $\mathbf{G}' = \mathbf{G} - \{F_i\}$. If $\mathbf{M} \subseteq \mathbf{G}$ is a Markov Blanket of F_i , then

$$\Delta(\mathbf{G}', \mathbf{F}) = \Delta(\mathbf{G}, \mathbf{F})$$

where Δ is any divergence function between 2 pdf's.

If Δ is the expected KL divergence

$$\Delta(\mathbf{G}, \mathbf{F}) = \mathbb{E}_{\mathbf{F}}\{D(P(C | \mathbf{F}) || P(C | \mathbf{G}))\}$$

Implication

Once we find a Markov blanket F_i in \mathbf{G} , we can safely remove F_i from \mathbf{G} without increasing the divergence

Algorithm: Iteratively remove a feature if it has a Markov blanket.

17

Approximate Markov blanket

Practically, we only search for Markov blankets of limited size.

It is still expensive to find an exact Markov blanket.

Observation:

If \mathbf{M} is really a Markov blanket for F_i , then for any feature value f_i ,

$$D(P(C | \mathbf{M}, F_i = f_i) || P(C | \mathbf{M})) = 0$$

Approximation:

Find \mathbf{M} so that the following quality (expected KL divergence) is small

$$\mathcal{D}(F_i | \mathbf{M}) = \mathbb{E}_{f_i}\{D(P(C | \mathbf{M}, F_i = f_i) || P(C | \mathbf{M}))\}$$

18

Approximate MB algorithm

Initialize

$\mathbf{G} = \mathbf{F}$

Iterate

For each feature $F_i \in \mathbf{G}$,

let $\mathbf{M}(F_i)$ be the set of k features $F_j \in \mathbf{G} - \{F_i\}$
having highest correlation with F_i

Compute $\delta(F_i | \mathbf{M}(F_i))$ for each $F_i \in \mathbf{G}$

Choose the $F_i = \operatorname{argmin}_F \delta(F_i | \mathbf{M}(F_i))$

Update $\mathbf{G} := \mathbf{G} - \{F_i\}$

19

Classification algorithms

3 classifiers are applied after filtering

- Multivariate Gaussian classifier
- Logistic regression
- K nearest neighbor

20

Gaussian classifier

- A Gaussian classifier is a generative classifier assuming data distributed as a mixture of c Gaussians
- The model θ consists of a prior probability π_c for each class c , having class-conditional density $N(\mu_c, \Sigma_c)$
- For binary case $C=2$, the ratio of posterior probabilities is

$$r = \frac{P(y=1|\mathbf{x},\theta)}{P(y=0|\mathbf{x},\theta)} = \frac{c_1 \exp\{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_1(\mathbf{x}-\boldsymbol{\mu})/2|\boldsymbol{\Sigma}_1|^{-2}\}}{c_0 \exp\{-(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_0(\mathbf{x}-\boldsymbol{\mu})/2|\boldsymbol{\Sigma}_0|^{-2}\}}$$

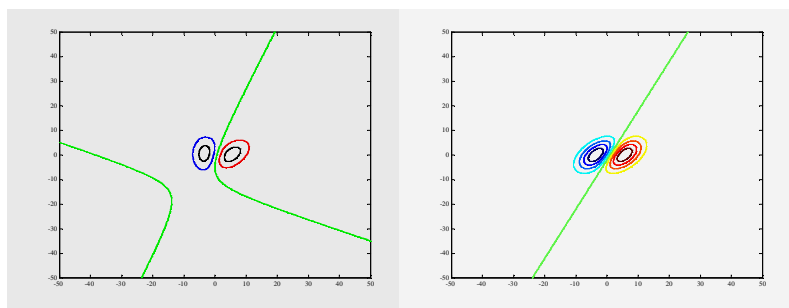
Theorem:
$$\log r = \frac{1}{2} \mathbf{x}' \boldsymbol{\Sigma} \mathbf{x} - \mathbf{p}' \mathbf{x} - \gamma$$

where $\boldsymbol{\Sigma}$, \mathbf{p} , γ are functions of model parameters π_c, μ_c, Σ_c .

21

Gaussian classifier – decision boundary

- The decision boundary is a quadratic surface in the feature space



22

Logistic regression

- Logistic regression is a discriminative classifier. The parameter $\boldsymbol{\theta}$ is a weight vector for \mathbf{x} .

$$p(y=1 | \mathbf{x}) = \frac{1}{1 + \exp\{-\boldsymbol{\theta}' \mathbf{x}\}}$$

- Geometrically, this classifier corresponds a sigmoid-shape ramp at the edge of the decision hyperplane.
- $\boldsymbol{\theta}$ can be estimated by stochastic gradient ascent:

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \rho(y_n - \hat{y}_n) \mathbf{x}_n$$

where

$$\hat{y}_n = \frac{1}{1 + \exp\{-\boldsymbol{\theta}' \mathbf{x}_n\}}$$

23

Regularization vs. Feature selection

Problem: Too many features implies too much complexity in the hypothesis space

Solutions:

- Feature selection: reduces the number of features
- Regularization: constrains the norm of parameters.

- Regularization can cope with overfitting
- Feature selection is easier to compute and interpret

24

Regularization vs. Feature selection

In a maximum likelihood setting, learning θ given data set \mathcal{D}

- Without regularization,

$$\hat{\theta} = \arg \max \{l(\theta | \mathcal{D})\}$$

- With regularization,

$$\hat{\theta} = \arg \max \{l(\theta | \mathcal{D}) - \lambda \|\theta\|\}$$

L1 or L2 norm



Regularization parameter

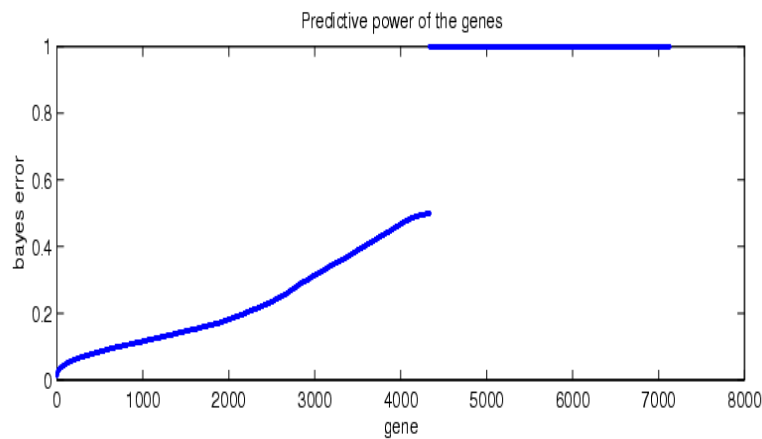


- In case of L2 norm, we obtain the stochastic gradient to estimate parameter

$$\theta := \theta + \rho((y_n - \hat{y}_n)\mathbf{x}_n - \lambda\theta)$$

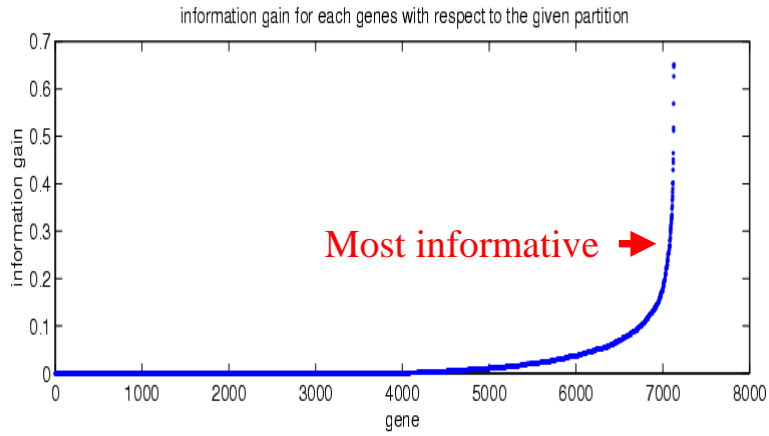
25

Filtering result: mixture overlap



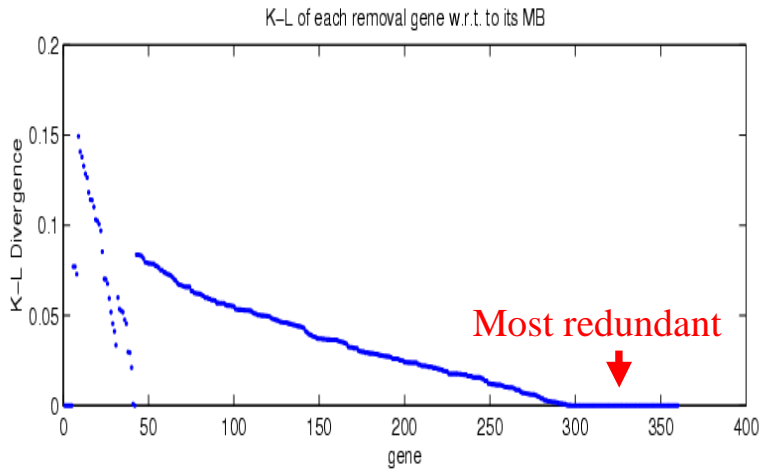
26

Filtering result: information gain



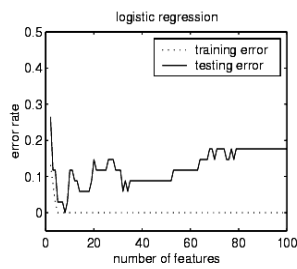
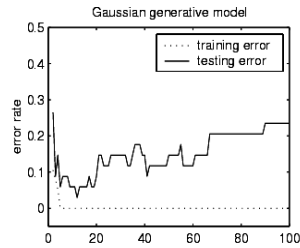
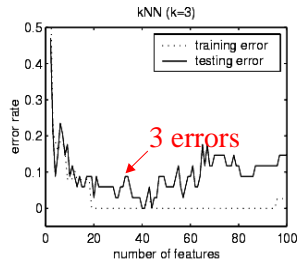
27

KL of top 360 genes by infogain



28

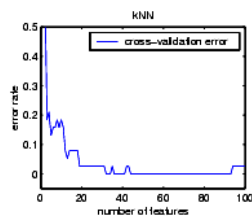
Classification result



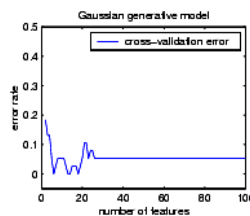
Best performance with feature selection (# errors)			
Classifier	$ G $	Training	Test
kNN	40	0	0
Gaussian	12	0	1
LR	8	0	0

Leave-one-out cross validation

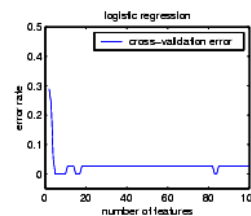
- Best number of features to be used for a classifier is chosen by minimizing the leave-one-out cross validation error



(a)



(b)



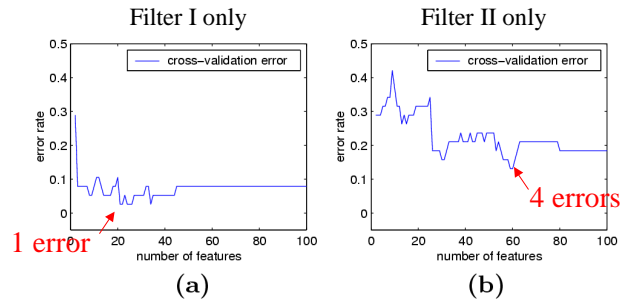
(c)

- Choose smallest $|G|$ that gives 0 error and the test error rates are:

kNN	Gaussian	Logistic Regression
5.9% (2/34)	8.8% (3/34)	0

Filter I vs. Filter II

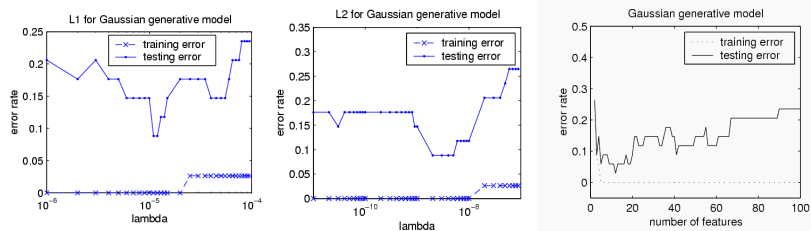
- Leave-one-out cross validation error for logistic regression with



31

Feature selection vs. regularization

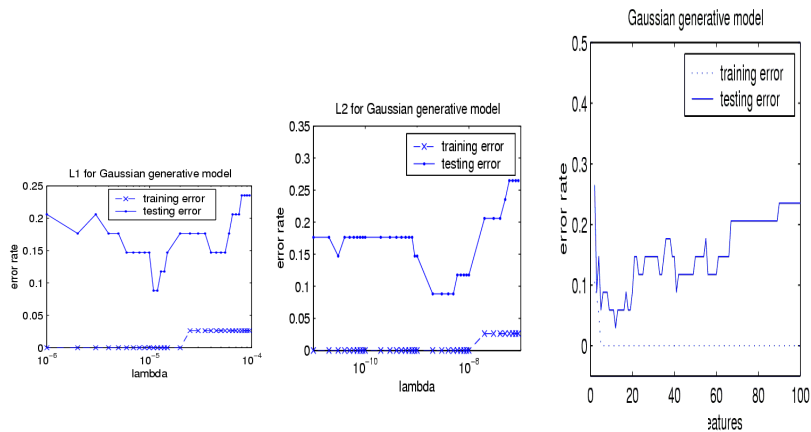
- Gaussian classifier using all features with L1 or L2 penalties.



32

Feature selection vs. regularization

- Gaussian classifier using all features with L1 or L2 penalties.



Summary and thoughts

- Microarray data create challenge to machine learning due to many features but few samples
- Feature selection or regularization is preferred before learning
- A series of filtering feature selection methods based on information theory are proposed in this paper
- The experimental design could better if:
 - Put more focus on feature selection
 - Compare with other's work
 - Use information theory based classifier
 - Redundant features could be used for cross validation

