Building a classification model based on miRNA data

Jiahua Chen

Canada Research Chair, Tier I
Department of Statistics University of British Columbia

Based on Joint Analysis with Mr. Yi Huang
as our contribution to a CIHR project of
Drs. Cathie Garnis, School of Dentistry

BIRS Workshop 14w5011

Outline

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

Content

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

Content

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

Disclaimer

- Up front, this talk does not contain any new results in terms of statistical methodology.
- Our main contribution is to investigate how to make use of miRNA data to predict/diagnosis the disease status of a future patient.
- We look for your comments on how to proceed, present what we have done, and what we are confused about.

Information from literature

- There have been many research reports that some high throughput genomic, proteomic or other similar data are statistically significantly different between cancer patients and normal control.
  - Cruz and Wishart (2004),
  - Delen et al. (2005),
  - Menden et al. (2013).
- These successes prompt others to follow suit, including my collaborator.

A CIHR funded research project

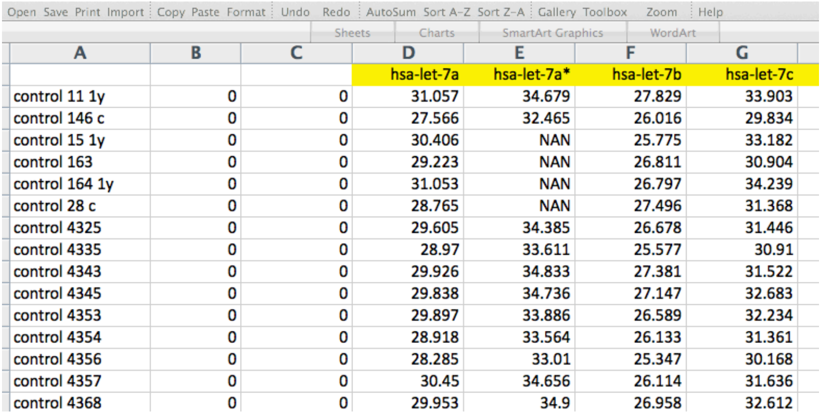
- We are participating in a research project which has been funded by CIHR.
- The goal is to determine if microRNA signatures derived from patient serum samples can be used to predict disease progression and recurrence.
- Success in this project will significantly improve disease management for those diagnosed with oral cancer or oral premalignant lesions

Content

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

A peek of the data

The following is a section of Excel File containing partially processed data:



A	B	C	D	E	F	G
			hsa-let-7a	hsa-let-7a*	hsa-let-7b	hsa-let-7c
control 11 1y	0	0	31.057	34.679	27.829	33.903
control 146 c	0	0	27.566	32.465	26.016	29.834
control 15 1y	0	0	30.406	NAN	25.775	33.182
control 163	0	0	29.223	NAN	26.811	30.904
control 164 1y	0	0	31.053	NAN	26.797	34.239
control 28 c	0	0	28.765	NAN	27.496	31.368
control 4325	0	0	29.605	34.385	26.678	31.446
control 4335	0	0	28.97	33.611	25.577	30.91
control 4343	0	0	29.926	34.833	27.381	31.522
control 4345	0	0	29.838	34.736	27.147	32.683
control 4353	0	0	29.897	33.886	26.589	32.234
control 4354	0	0	28.918	33.564	26.133	31.361
control 4356	0	0	28.285	33.01	25.347	30.168
control 4357	0	0	30.45	34.656	26.114	31.636
control 4368	0	0	29.953	34.9	26.958	32.612

The biological process

- Serum samples are collected from individuals, case or control.
- The targeted expression level is measured based on the real-time PCR
- Roughly, the device counts the number of doubling cycles it takes for the target miRNA until its abundance in the sample attains a specific level.
- The abundance itself is measured based on the strength of the light the sample reflects.

CENSORSHIP

- If it takes 35 cycles or more for the abundance to reach the level, the measurement is considered as censored.
  The strength of the reflected light after 35 cycles will exceed the threshold value even if we start with plain water.
- A higher CT reading indicates a lower initial abundance of the target miRNA.
- We replace all censored values by 35. The effect of this practice, if any, would be negative toward the usefulness of the classifier to be built.

Some specifics of the data set

- My presentation will be based on 105 miRNA readings on 48 cases and 51 controls.
- Additional samples are not included because they have only 13 of these 105 miRNAs measured.

Normalization

- Denote 105 miRNA readings as a vector x.
- Because the growth rate of the miRNA may differ from person to person, it is recommended to have x value "normalized".
- One way to normalize is to locate a most suitable miRNA in the list and subtract its value from all other miRNA readings.
- We refer the normalized value as $\check{x}$.

Transformation

- One may transform the CT value into direct measurements of the miRNA abundance: $\tilde{x} = \exp(-\check{x} \log 2)$.
- The data analysis can be done on either x, $\check{x}$ or $\tilde{x}$, regardless the scientific justification.

Content

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

Content

1 Introduction
- Scientific Motivation
- Explanation of the miRNA data

2 Classification
- General introduction to classification
- Variable selection and lasso
- Assessment of performance

3 Data analysis
- Applying Lasso logistic regression
- Logistic regression with additive effects
- Classification tree

An abstract classifier

- Suppose we have an abstract population made of sample units, so that each unit has an attribute vector x and a status value Y.
- A classifier is a function of x taking values in the domain of Y.
- If Y is either "cancer" or "control", the classifier takes value 1 or 0.

A logistic regression

- When we postulate a logistic linear regression on the target population, we are assuming there exists a vector β such that

$$\log \frac{\mathsf{P}(Y = 1 \mid x[1:p])}{\mathsf{P}(Y = 0 \mid x[1:p])} = \beta_0 + \sum_{j=1}^{p} \beta[j]\, x[j]. \qquad (1)$$

- Note that the probability statement is population dependent.

Notational convention

- We use p for the dimension of x.
- We will use a general notation s for a subset of $1:p$.
- If $s = \{2, 6, 8, 30\}$,

  $$x[s] = (x[2], x[6], x[8], x[30])^T;$$

  and $x[0, s]$ is $x[s]$ with additional component $x[0]$.
- Similarly, $x[3]$ means the third component of vector x.

A LOGISTIC REGRESSION AS A CLASSIFIER

- Suppose we have built a logistic regression for a population:

$$\log \frac{\mathsf{P}(Y = 1 \mid x[1:p])}{\mathsf{P}(Y = 0 \mid x[1:p])} = \beta[0] + \sum_{j=1}^{p} \beta[j]\, x[j].$$

  That is, the vector β has been completely specified.

- Let x be the attribute vector (miRNA values) of a unit sampled from the same population with its Y value concealed.
- One possible classification rule is to classify the unit as $Y = 1$ when $\mathsf{P}(Y = 1 \mid x) > 0.5$ according to the built logistic model.

The role of population

- The value of $\beta[0]$ is dependent on how the sample is obtained from the population.

- In applications, we often take samples retrospectively to build a logistic model.

- Even if the model assumption is correct, and sample size is infinite, the value of $\hat{\mathrm{P}}(Y = 1 \mid x)$ is not the probability of a new patient admitted to the same clinical has cancer.

- The classification rule on the last slide is questionable from this point of view. Yet this can be a starting point of statistical analysis.

Content

1 Introduction

- Scientific Motivation
- Explanation of the miRNA data

2 Classification

- General introduction to classification
- Variable selection and lasso
- Assessment of performance

3 Data analysis

- Applying Lasso logistic regression
- Logistic regression with additive effects
- Classification tree

Variable selection

- To make use of logistic model, we must have β estimated with sufficient accuracy.
- When the number of attributes, p, is large, the model can always fit the data set very well yet it has little predictive value.
- Including only a selective few attributes in the model helps. This leads to “variable selection” issue.

Model fitting

- Suppose we have a size n sample from the target population.
- Given a subset s, the log-likelihood function conditional on $x[s]$ values is given by

$$\ell_n(\beta[0, s];\ x[s]) = \sum_{i=1}^{n} \{y_i \log \mathsf{P}(y_i; x_i[s]) + (1 - y_i) \log \mathsf{P}(1 - y_i; x_i[s])\}.$$

- Fitting this model usually means to find a $\hat{\beta}[0, s]$ at which this likelihood is maximized.

Computational issue

- The numerical task of fitting the above model can be carried out with a R-function very quickly and reliably.
- We usually judge the fitness of the model built on the specific $x[s]$ based on $\ell_n(\hat{\beta}[0, s] \mid x[s])$.
- When $p = 105$ and size of s is 5, there are over 96.5 million such subsets/models. Computation of $\hat{\beta}[0, s]$ for all of them is infeasible.

Computationally effective regularization method

- A class of regularization methods have been recently proposed which helps to cut down the amount of computation, and lead to a sensible compromise.
- LASSO is likely the most popular one. It works at finding the maximum point of

  $$\ell_n(\beta[0:p] \mid x[1:p]) - \lambda\Big|\beta[1:p]\Big|$$

  for some positive constant λ, where $|\beta[1:p]| = \sum_{j=1}^{p} |\beta[j]|$.
- When λ decreases from infinity to 0, $\hat{\beta}_\lambda[1:p]$ contains practically increasing number of non-zero entries, starting from $\hat{\beta}_\infty = \mathbf{0}$.

Outcome of lasso

- When lasso is applied, a sequence of nested and successive subsets of attributes (miRNAs) will be produced.
- Each of them offers a classification rule for future observations.
- Which one of them is the best? This answer depends on the definition of best.
- Lasso itself does not have a generically recommended λ value.
- This leads to the issue of tuning the Lasso.

Content

1 Introduction
- Scientific Motivation
- Explanation of the miRNA data

2 Classification
- General introduction to classification
- Variable selection and lasso
- Assessment of performance

3 Data analysis
- Applying Lasso logistic regression
- Logistic regression with additive effects
- Classification tree

Measure the performance

- To decide which subset of attribute, s, is the best for building a classifier, we need a way to measure its performance.
- We use cross-validation in this project.
- We now consider the case where the logistic regression is used to build a classifier.

Cross-validation.

- Let a subset of attributes, s, be **given**.
- We randomly divide the data set into training set and test set.
- We fit a model based on $(x[s], y)$ in the training set.
- The resulting classifier is applied to units in the test set. Compute performance measurements.
- Repeat the "divide-fit-classify" many many times to obtain the "mean" performance measurements.

Performance measure

- We use sensitivity, specificity, and their average to jointly judge the classifier based on s.
- One dilemma is: when the data set changes, $\hat{\beta}[s]$ changes. Hence, the judgement is not on a specific classifier, but is on s.
- Another dilemma: can we interpret the "mean" performance measurements straightforwardly?

Content

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

Content

1 Introduction
- Scientific Motivation
- Explanation of the miRNA data

2 Classification
- General introduction to classification
- Variable selection and lasso
- Assessment of performance

3 Data analysis
- Applying Lasso logistic regression
- Logistic regression with additive effects
- Classification tree

First step of using lasso

- The following miRNAs are selected by lasso based on normalized CT: $\check{x}$.

hsa-miR-23a	hsa-miR-346	hsa-miR-342-3p	hsa-miR-205	hsa-miR-33a
hsa-miR-582-5p	hsa-miR-125b	hsa-miR-497	hsa-miR-28-3p	hsa-miR-10a
hsa-miR-616	hsa-miR-142-3p	hsa-miR-200c	hsa-miR-29b-1	hsa-miR-365
hsa-miR-654-3p	hsa-miR-490-3p	hsa-miR-744	hsa-miR-934	hsa-miR-888
hsa-miR-1909				

- As an example of how to read the table, model s_5 is made of first five miRNAs in this table.

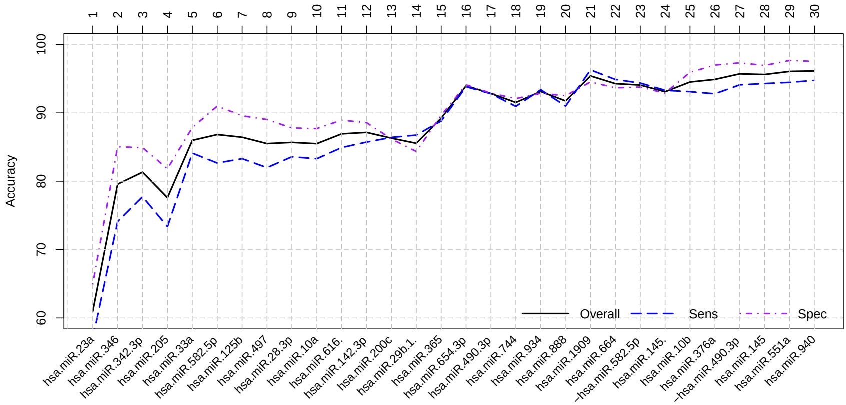
Performance of s_5

- We repeat the following steps 1000 times.

  A: randomly select 16 + 16 units to form a test set. Use the rest of units to fit a logistic regression model on $x[s_5]$*.*

  B. classify each unit in the test set. Record the numbers of correctly classified cases and controls.

Performance of $s_1, s_2, \ldots$ based on $\check{x}$



Best performance measurements obtained

- The best performances achieved:

	Sensitivity	Specificity	Overall
x	91.49	95.36	93.43
$\check{x}$	96.29	94.53	95.41
$\tilde{x}$	84.90	87.09	85.99

- The high values are from models with over 20 miRNAs.
- I would prefer the model with 6 miRNAs which is a local maximum.

Are we successful?

- If the cross-validation performance is a good indication of future precision, then a 6 normalized miRNAs can make a very good classifier.
- These 6 miRNAs differ moderately from those identified by the analysis of our dentistry collaborators.
- Should we recommend this specific classifier to be validated with their future observations? I would be happy to hear from you.

Content

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

Can we do even better?

- Perhaps the log-odds is not linear in $\check{x}$ nor linear in $\tilde{x}$. but is linear in functions of $\check{x}$.
- This consideration leads to the generalized additive model (GAM):

$$\log \frac{\mathsf{P}(Y = 1 \mid x[s])}{\mathsf{P}(Y = 0 \mid x[s])} = \beta[0] + \sum_{j \in s} f_j(x[j]), \qquad (2)$$

for some unspecified function f_j.

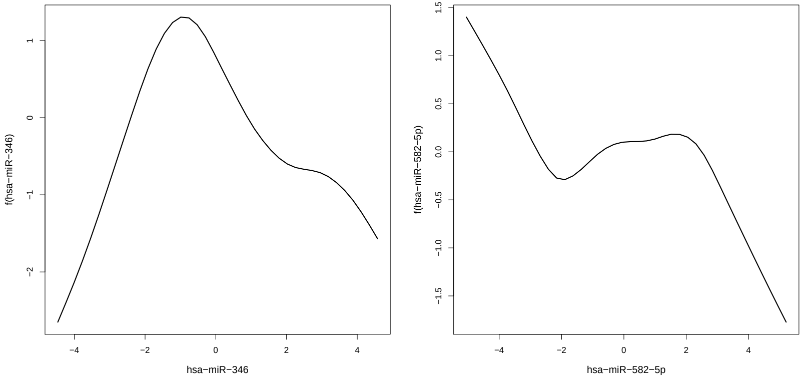
Fitting GAM

- We now must find a most suitable f_j based on data.
- If f_j is allowed to take any form, the result is definitely overfitting.
- Under some smoothness restriction/penalty, the choice is a cubic spline.
- A cubic spline is a piece-wise polynomial of order 3 over the range of $x[j]$, differentiable to order 3 at knots.

Fitting GAM

- We now must find a most suitable f_j based on data.
- If f_j is allowed to take any form, the result is definitely overfitting.
- Under some smoothness restriction/penalty, the choice is a cubic spline.
- A cubic spline is a piece-wise polynomial of order 3 over the range of $x[j]$, differentiable to order 3 at knots.

Examples of fitted splines



Performance of GAM

- The predictive precision of the best GAMs are:

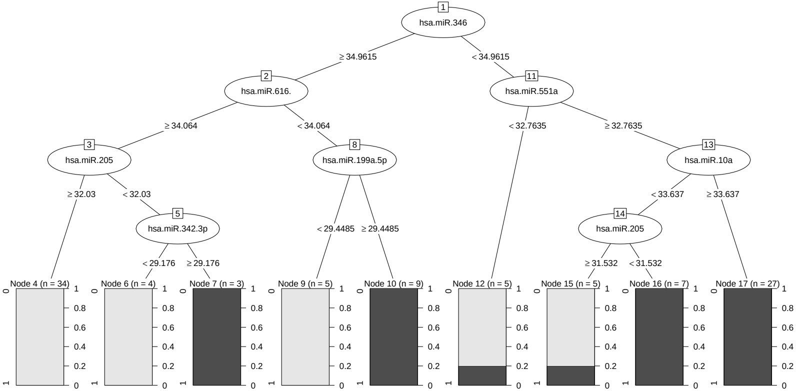
	Sensitivity	Specificity	Overall
x	90.31	84.69	87.50
$\check{x}$	77.81	80.63	79.22
$\tilde{x}$	73.13	82.50	77.81

- The flexibility of GAM does not seem to help.

Content

1. Introduction
   - Scientific Motivation
   - Explanation of the miRNA data
2. Classification
   - General introduction to classification
   - Variable selection and lasso
   - Assessment of performance
3. Data analysis
   - Applying Lasso logistic regression
   - Logistic regression with additive effects
   - Classification tree

Can classification tree have a better performance?



- Instead of giving an introduction to classification tree, let me show one fitted to our data.

Interpreting and assessing a classification tree

- Classification tree is easy for understanding and presentation.
- When growing a classification tree, we also automatically perform variable selection: miRNAs used in splitting nodes are those to be kept.
- The predictive precision of classification tree, however, is hard to assess: the classification trees built on different training set uses different miRNAs to split nodes.
- Consequently, we can only assess the "classification tree" procedure.

The performance of the classification tree procedure

	Sensitivity	Specificity	Overall
x	71.84	77.04	74.44
$\check{x}$	69.24	71.87	70.55
$\tilde{x}$	70.10	73.14	71.62

- Almost to our relief, the tree method does not work too well.

Conclusions

- The simplistic logistic regression with variables selected via lasso using cross-validation seems to work well.
- Other methods, already tried or to be tried, may not help a lot.
- Do the cross-validation “sensitivity” and “specificity” estimate the true “sensitivity” and “specificity” consistently?

Conclusions

- The simplistic logistic regression with variables selected via lasso using cross-validation seems to work well.
- Other methods, already tried or to be tried, may not help a lot.
- Do the cross-validation "sensitivity" and "specificity" estimate the true "sensitivity" and "specificity" consistently?
- Ultimately, are we confident enough to recommend a "confirmation study"?
- What would be your sample-size formula?

Last slide

Thank you.

Cruz, J. A. and Wishart, D. S. (2006).
Applications of machine learning in cancer prediction and prognosis.
Cancer informatics, 2:59.