Supervised classification and tunnel vision

David J. Hand
Imperial College

d.j.hand@ic.ac.uk

Paris, November 2002

PART I: Tunnel vision

The standard supervised classification paradigm:

> *Given a set of objects, from each of which a vector of measurements has been taken, and for each of which its class membership is known, construct a rule which will allow one to assign new objects to classes using only their measurement vector*

Huge amount of work in different areas:

- statistics
- pattern recognition
- machine learning
- operations research
- data mining

Evolution of ***methods*** - largely driven by progress in computer technology

- LDA
- QDA
- logistic DA
- nearest neighbour and kernel nonparametric methods
- trees
- PPR
- MARS
- ANN
- SVM
- Ensemble models
    - model averaging
    - boosting
    - bagging

and of ***understanding***

- overfitting
- generalisation

Emphasis on the ***accuracy*** of the rule:

This is a simplistic view of many real problems

Many issues in addition to accuracy

- interpretation
- performance measure
- drift
- out of date data
- sample bias
- handle incomplete data?
- etc

These other issues can swamp the 'accuracy' improvements

Illustrate for consumer credit scoring, but similar points apply elsewhere

'The tools [of credit scoring] are based on statistical and operational research techniques and are some of the most successful and profitable applications of statistical theory in the last 20 years..'

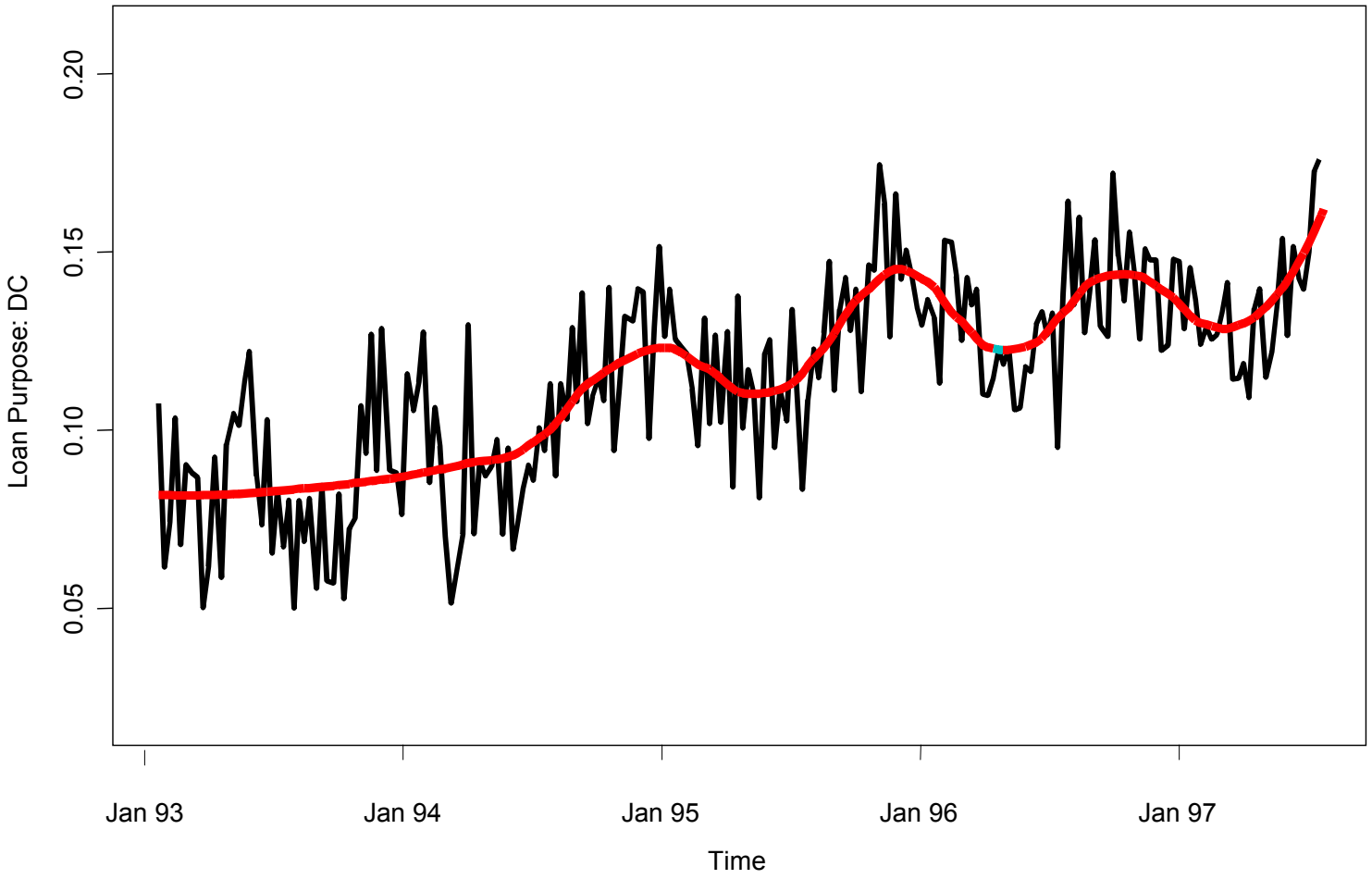
Crook, Edelman, and Thomas (1992)

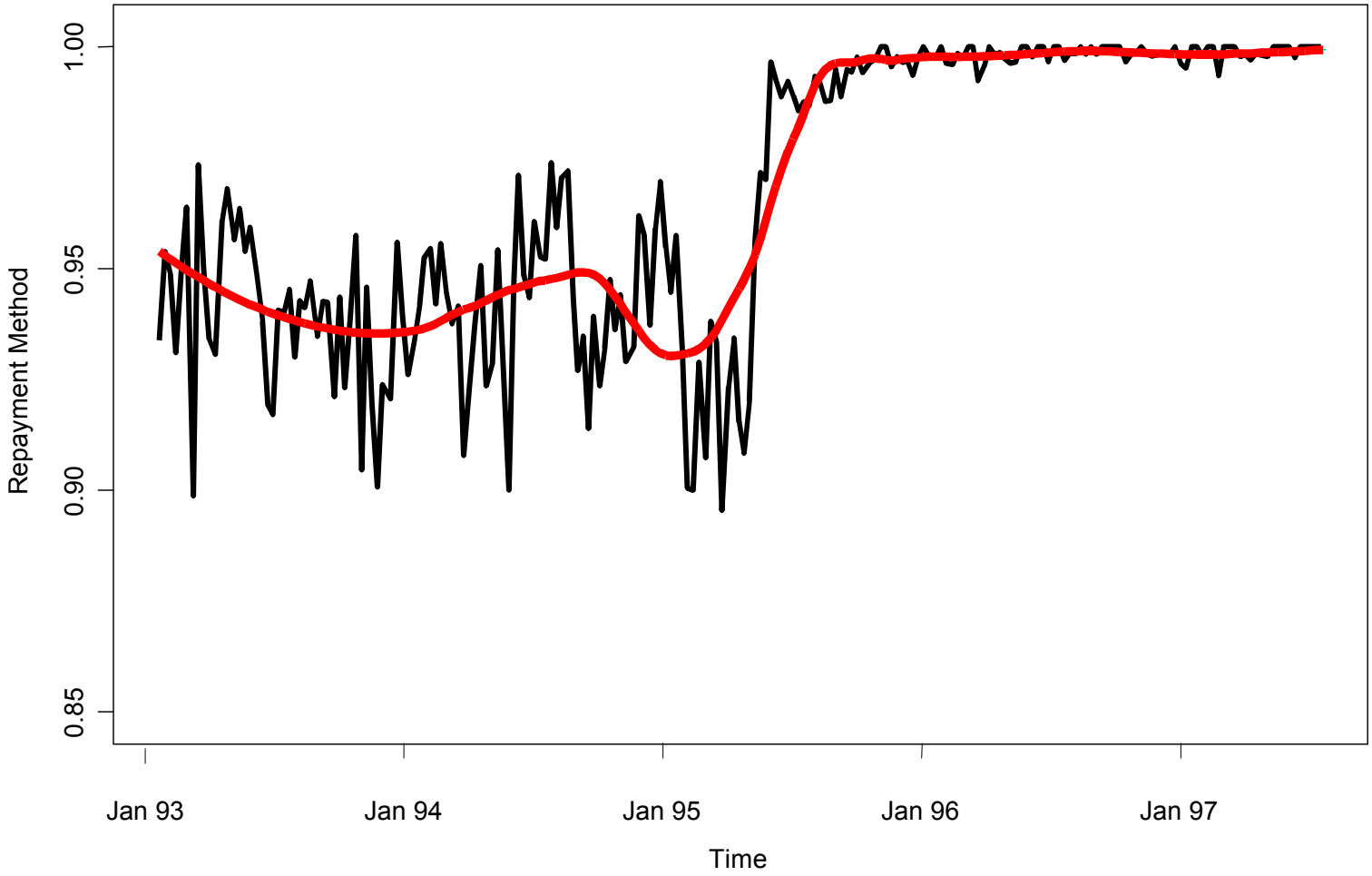
Many problems in this area fall into the classifier mould:

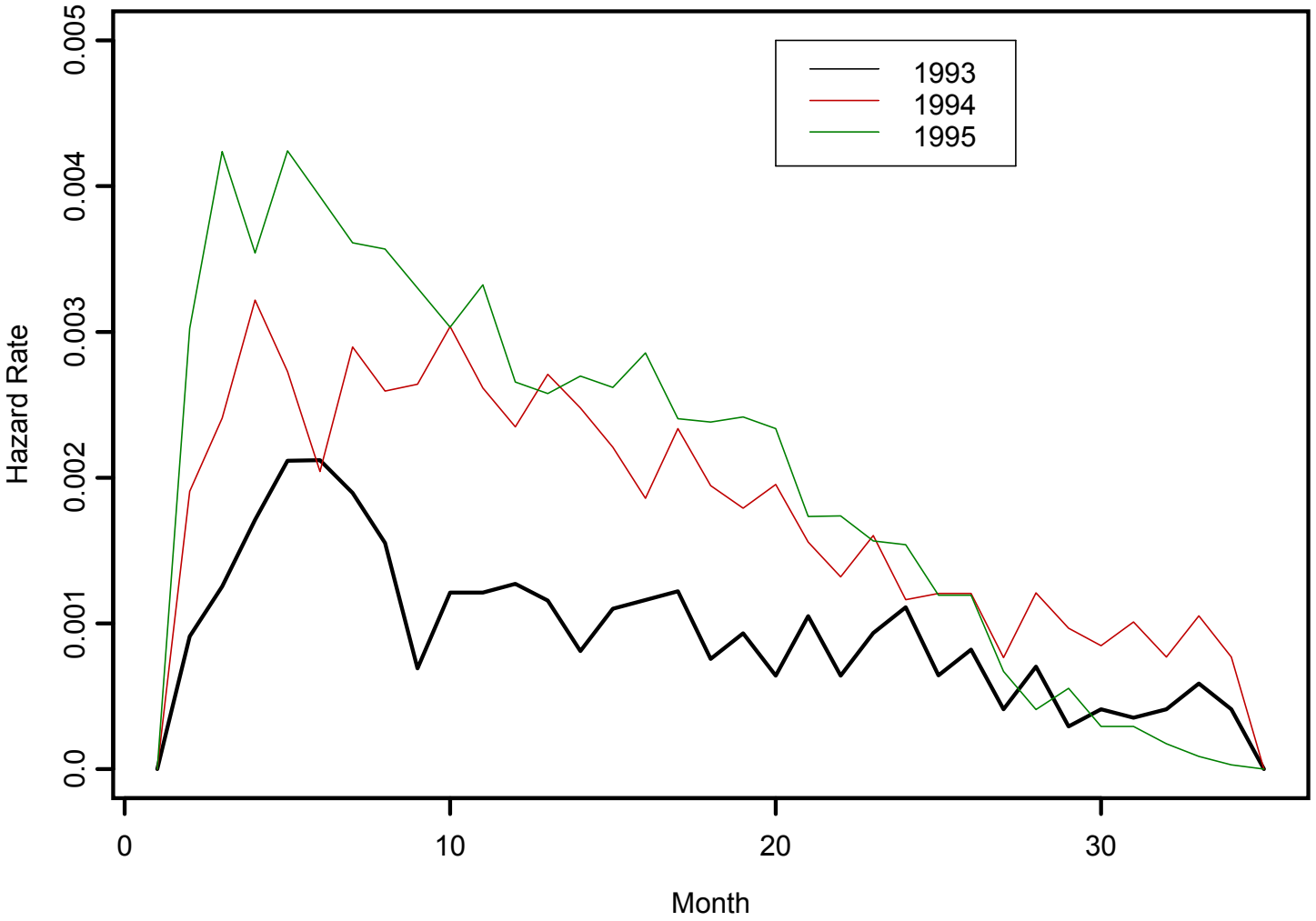
- default
- fraud
- churn
- extend loan
- etc.

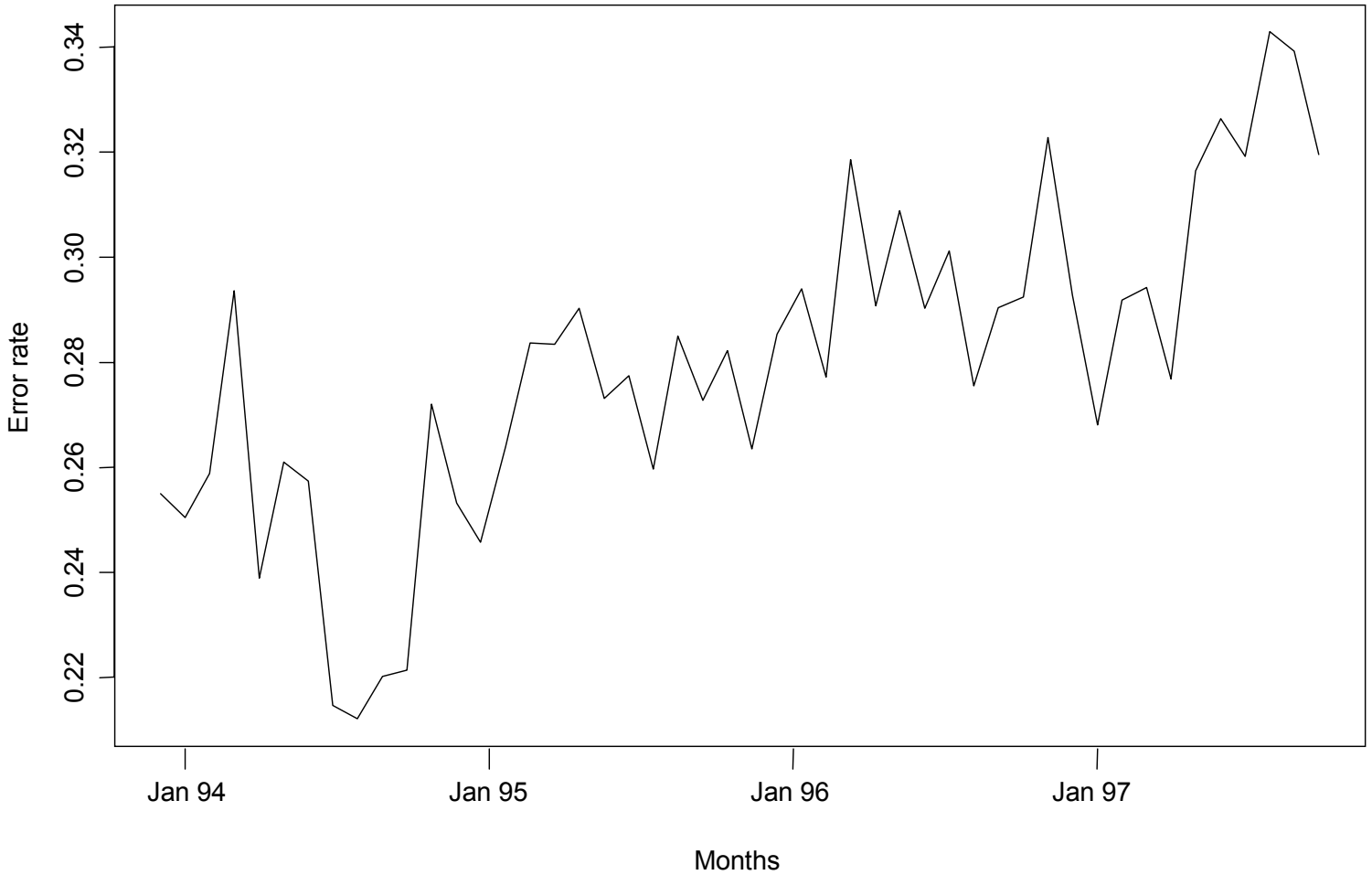
Population drift

- seasonality
- due to changing economic climate
- due to customers dropping out
- due to changing sales and marketing strategies
- etc.


Loan Purpose: DC
0.05
0.10
0.15
0.20
Jan 93
Jan 94
Jan 95
Jan 96
Jan 97
Time


Repayment Method
1.00
0.95
0.90
0.85
Jan 93
Jan 94
Jan 95
Jan 96
Jan 97
Time


1993
1994
1995
Hazard Rate
0.0
0.001
0.002
0.003
0.004
0.005
0
10
20
30
Month


Error rate
0.22
0.24
0.26
0.28
0.30
0.32
0.34
Jan 94
Jan 95
Jan 96
Jan 97
Months

- Standard paradigm: use test data from same distribution as design data
- *But future data are not from this distribution*
- Must wait to discover class labels in design set:
  - 2 years for a bank loan?
  - 25 years for a mortgage?

    → *rules are out of date before you start using them*

    → their performance gets even worse over time

Tiny differences between rule construction methods are irrelevant in the context of variations in the populations

Is the effort and expense of developing new rule worthwhile?

Is all the research on improved classification methods relevant?

In times of change, learners inherit the Earth, while the learned find themselves beautifully equipped to deal with a world that no longer exists

Eric Hoffer

Reject inference

Observe customer behaviour:

- some stay good
- some go bad

But only observe true class for those accepted

- selectivity bias
- bads more likely to be rejected than goods

e.g. Number of weeks since last CCJ

Level (weeks)	% G in whole pop	% G in accepts	Ratio .
1-26	22.7	44.6	1.964
27-52	30.4	46.9	1.542
53-104	31.4	49.2	1.567
105-208	37.8	55.2	1.460
209-312	42.6	63.1	1.481
> 312	55.6	69.2	1.245

Reject inference is the term used for trying to predict what would be the good/bad class of the rejects and then using these predictions in building an improved classification rule, assessing the current rule's performance, etc.

What's the aim ? – is it to build a rule which predicts including 'policy rules' etc.?

In practice, things are further complicated by attrition: customers accepted, but don't take up product

$f(g \mid y)$ = Prob (applicant with vector *y* is good)

$$f(g \mid y) = f(g \mid y, a) f(a \mid y) + f(g \mid y, r) f(r \mid y)$$

$f(a \mid y)$ and $f(r \mid y)$ observed from past data

$f(g \mid y, a)$ observed from past data

$f(g \mid y, r)$ not observed

Original rule – based on characteristics *X*

New rule – based on characteristics *Y*

Case 1: $X \subset Y$

Case 2: $X \not\subset Y$ e.g. *age* used in original rule, but not available for new one

Case 1: $X \subset Y$

- leads to a partition of the *Y* space

→ extrapolate from accept region over reject region

Effectiveness depends on:

- the form of the model in the accept region also applying in reject region (not so critical, since only *thresholds* matter)
- accuracy of model built in accept region
- continuity assumption – that $f(g \mid y)$ does not change dramatically with *y*

This last can be problematic (recall that most characteristics are categorical):

e.g. Suppose original included *number of CCJs* (*NCCJs*), and (to make life simple) suppose that *NCCJs* > 0 $\Rightarrow$ high prob of bad.

Then original will reject all applicants with *NCCJs* > 0.

New has only those with *NCCJs* = 0, so that this highly predictive variable will not figure in new.

If extrapolate, then perhaps use confidence bounds: best and worst case situations

Case 2: $X \not\subset Y$

Possible that *every y* has some accepts and some rejects. For the rejects, we don't know $f(g \mid y, r)$ so for *no y* can we compute $f(g \mid y)$

Possible that $f(g \mid y, r)$ very different from $f(g \mid y, a)$ (indeed, if original is any good, this will be the case)

That is, one might expect characteristics in *X* / *Y* to be predictive of g/b

Some approaches:

1) Ignore the issue – not a good idea, as the example above shows

2) Assume all rejects are bad – may be OK if a more stringent rule is needed

3) Augmentation methods assume $f(g \mid y, r) = f(g \mid y, a)$

4) Mixture decomposition approach requires nothing whatsoever is known about $f(g \mid y)$ - but requires strong assumptions about the *forms* of the distributions of $f(y \mid g)$ and $f(y \mid b)$.

5) Accept some rejects

6) Use information from other sources – often rejects find products from other suppliers. Beware sample distortion, different good/bad definitions, etc.

Interpretation

Modern sophisticated tools may have smaller error rate

But they achieve this via complicated nonlinear decision surfaces

Many problems require a simple model which can be explained

Credit scoring problems are of two kinds:
- front end. e.g. loan decisions
- back end e.g. fraud detection

Front end must often be simple: popular choice GLMs
Back end can be complicated: ANNs, etc.

Improved accuracy of a non-interpretable model is pointless

Measuring performance

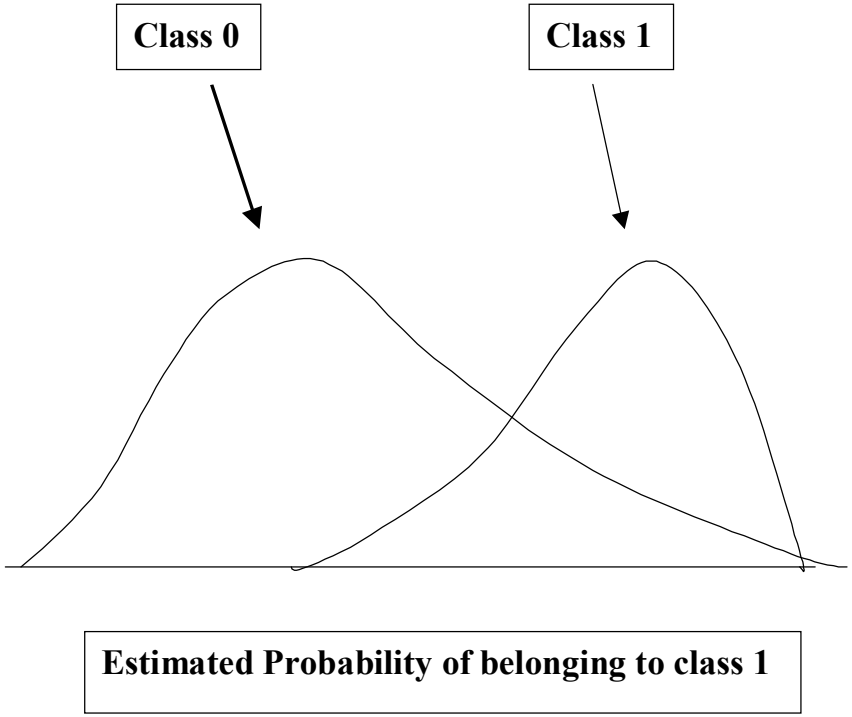
Most common choice: error rate

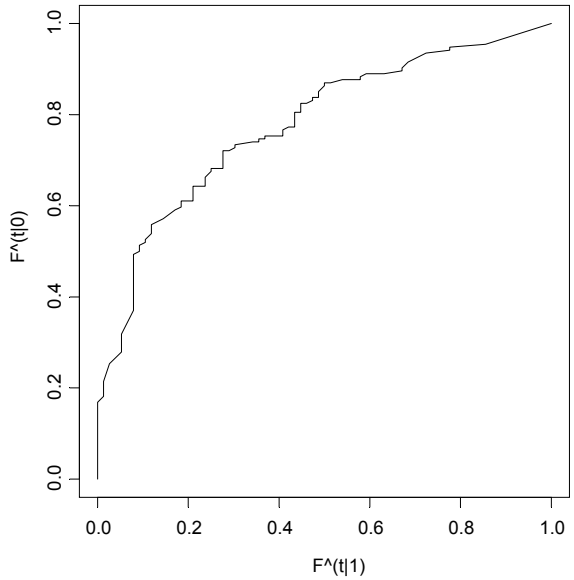
Error rate is seldom appropriate
- assumes equal misclassification costs

Use known costs - determined from the problem
- if you can determine them

Area under ROC curve / Gini / Mann-Whitney-Wilcoxon (*U*)

Area under ROC curve / Gini / Mann-Whitney-Wilcoxon (*U*)





Area under ROC curve / Gini / Mann-Whitney-Wilcoxon (*U*)

- avoids choice of misclassification costs
  - equivalent to weighted integral of error rate over all choices of cost ratio, but with a data derived weight function

Great deal of work on obtaining accurate estimates of error rate

Misplaced effort? Given that error rate is seldom of real interest?

Research opportunity: transfer the methods developed for error rate estimation to other performance measures

PART II: Local models (with PhD student Veronica Vinciotti)

Fundamental assumption typically made in statistical modelling:

The inference and prediction stages are separate

A statistical inference carries us from observations to conclusions about the populations sampled …No considerations of losses [consequences] is usually involved directly…The theory of statistical decision deals with the action to take on the basis of statistical information. Decisions are based on not only the considerations listed for inferences, but also on an assessment of the losses resulting from wrong decisions…

David Cox

…the inference problem is basic to any decision problem, because the latter can only be solved when the knowledge of [the probability model] … is completely specified … The person making the inference need not have any particular decision problem in mind. The scientist in his laboratory does not consider the decisions that may subsequently have to be made concerning his discoveries. His task is to describe accurately what is known about the parameters in question.

Dennis Lindley

But this is only true if the models are properly specified

- *Which they almost never are*
- *Often deliberately choose an improperly specified model*
  e.g. for interpretability

In an improperly specified model
- not seeking a model of the underlying data generating mechanism
- seeking a model which is *closest* to that mechanism
- where *closest* is measured by some criterion
- often (penalised) likelihood

Different criteria will yield different models (even for $n \rightarrow \infty$)

A given criterion will sum over contributions from each design set point
A given criterion will average over the entire data space

But in many problems we need a good model only in part of this space

Illustrate using credit scoring in retail banking:

(i) Interpretability is important (maybe legally required)
- we'll restrict to models linear in components of **x**

(ii) We are not interested in accuracy of $\hat{P}(0 \mid \mathbf{x})$ for all **x**

But only in the vicinity of $P(0 \mid \mathbf{x}) = t$

where $t = k_1 / (k_0 + k_1)$

with k_i the cost of misclassifying a class *i* customer

If the model is misspecified, forcing accuracy in places we don't need it may sacrifice accuracy where we want it

E.g. logistic discrimination based on likelihood

Model: $\hat{P}(0 \mid \mathbf{x}) = \left(1 + e^{-\boldsymbol{\beta}'\mathbf{x}}\right)^{-1}$

Criterion: $L = \sum_{i=1}^{n} \ln \hat{P}(c_i \mid x_i)$

- gives equal weights to all $\hat{P}(c_i \mid x_i)$

But we only need to know when $P(c \mid x) > t$ or $P(c \mid x) \leq t$

We only need an accurate estimate of $P(c \mid x) = t$

Likelihood is an inappropriate criterion for measuring closeness

Modify the criterion to emphasise regions of interest

[aim determines 'region of interest'
- region of interest determines criterion
- criterion determines model]

$\Rightarrow$ *locally weighted logistic regression* $\quad L = \prod_{i=1}^{n} \hat{P}(c_i \mid x_i)^{w_i}$

Various approaches:

(i) $w_i = \alpha \exp\left\{-\lambda\left(\hat{P}(0 \mid x_i) - t\right)^2\right\}$

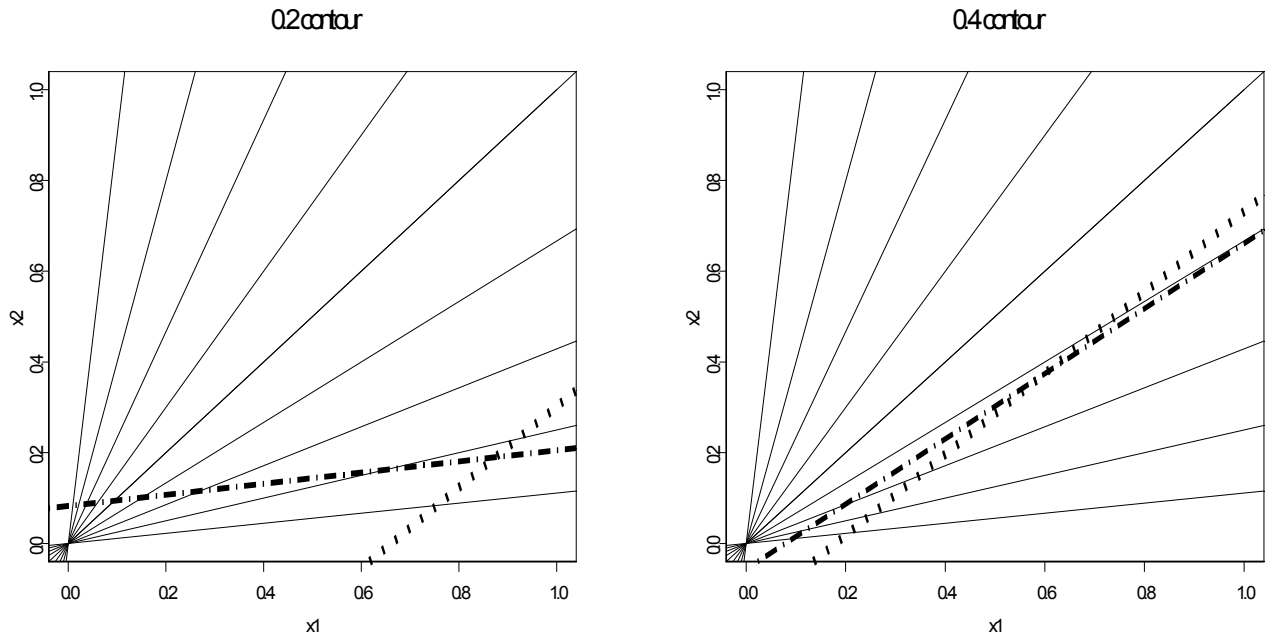
(ii) $w_i = \begin{cases} 1 & x_i \text{ amongst } k \text{ nearest points to } x \\ 0 & \text{otherwise} \end{cases}$

Iterative methods

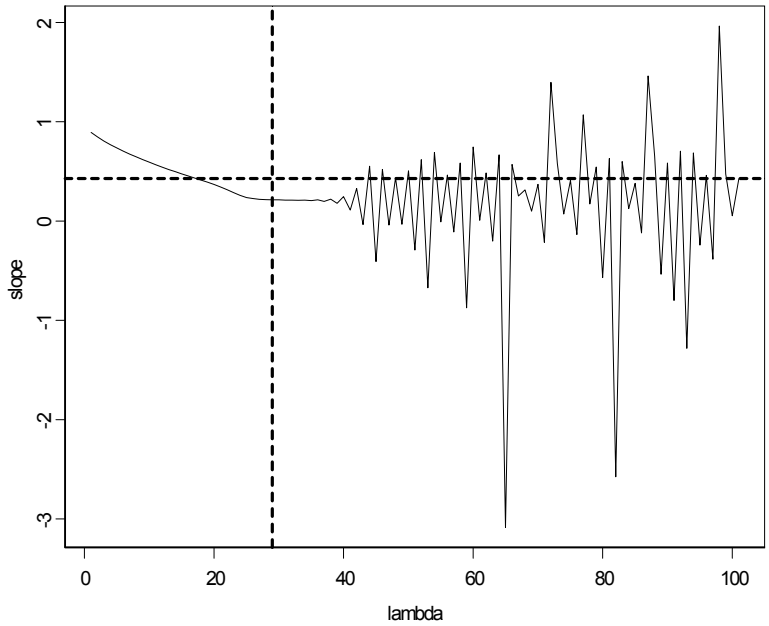
(a) Start with $\lambda = 0$ and gradually increase, evaluating cost-weighted misclassification rate at each value

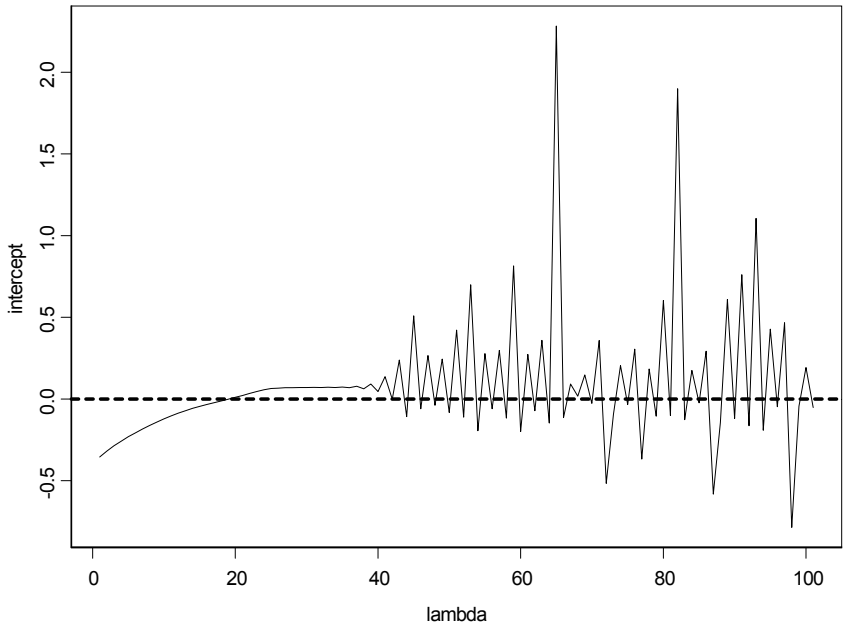
(b) Start with $k = n$ and gradually decrease, evaluating cost-weighted misclassification rate at each value

Example 1: Estimates of the $P(0 \mid \mathbf{x}) = 0.2$ and $P(0 \mid \mathbf{x}) = 0.4$ contours by standard logistic discrimination and one step local logistic discrimination, weights (i).

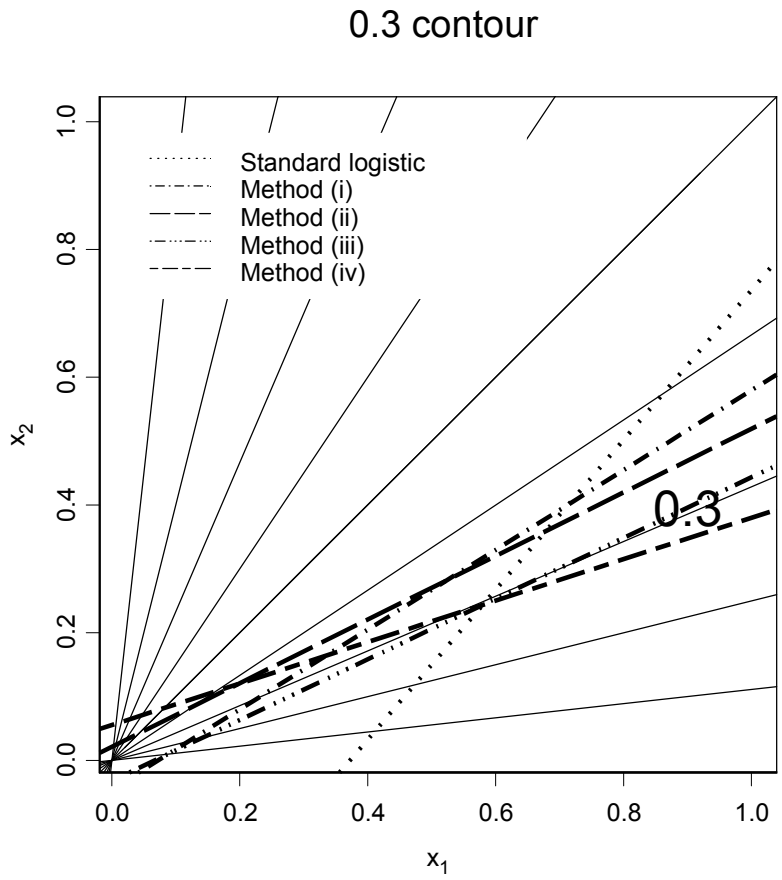


Gradually increase λ to find a good value:



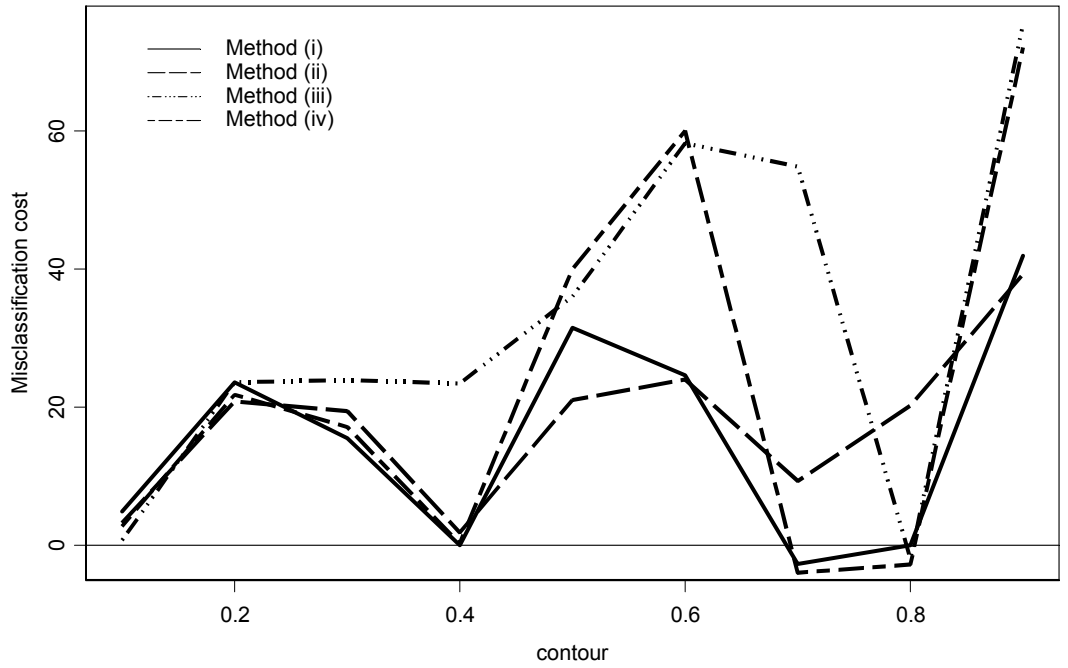


Estimates of the 0.3 contour by all methods



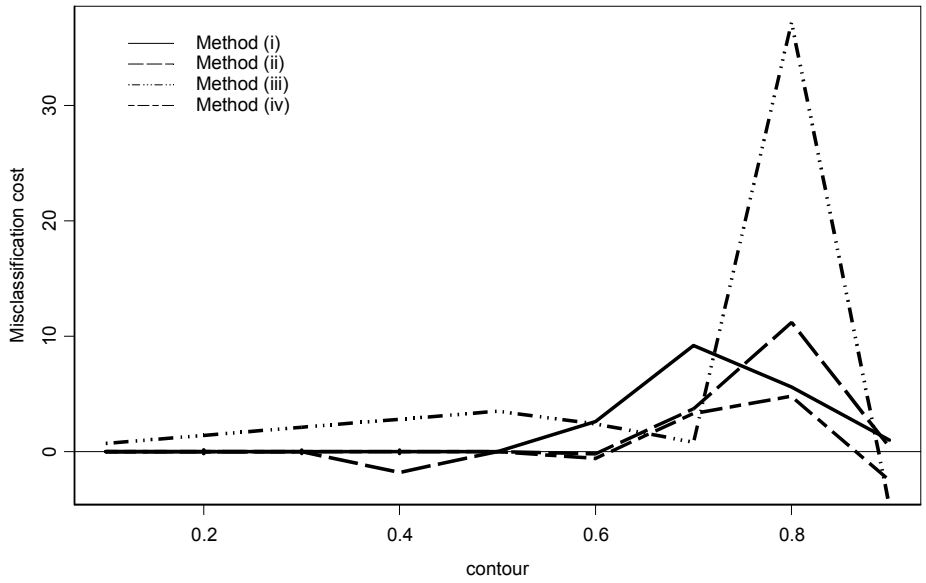
Example 2: Predicting earnings class (>,≤ $50,000 pa)

- 30,162 observations
- 15 variables (age, sex, marital-status, education, bank account state, etc).



Example 3: Unsecured personal loans

- 21,618 observations, 24-month term, two year period Jan 1995 to Dec 1996
- 16 variables describe the application for the loan.
- an account is defined as bad if it is at least one months in arrears
- 11% bads



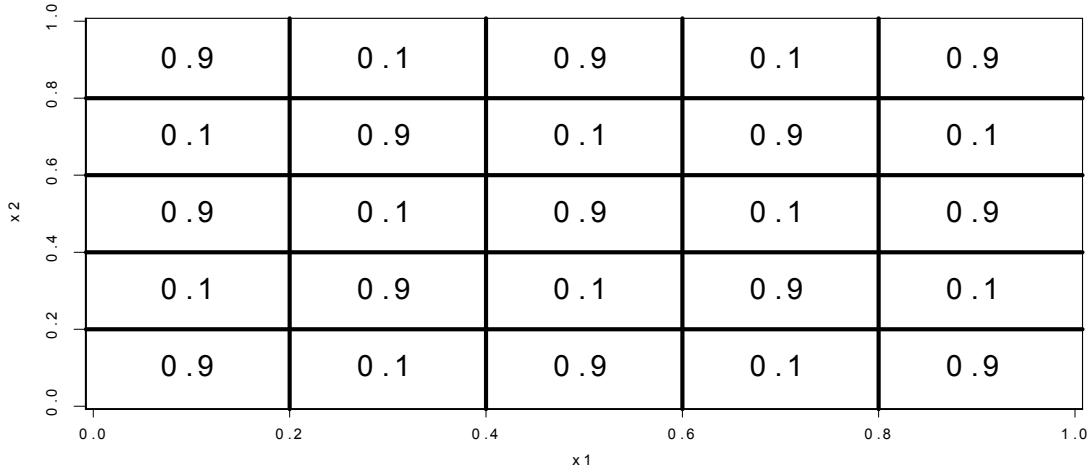
Example 4: k-nearest-neighbour

Two cases:

(1) misclassification costs $k_0 = k_1 = 0.5$, leading to $t = 0.5$

(2) misclassification costs $k_0 = 0.05$ and $k_1 = 0.95$, leading to $t = 0.95$

$P(0 \mid \mathbf{x})$ in a bivariate example.



Case (1): important to distinguish between the alternating squares

$\rightarrow$ use small k in k-nn

Case (2): all regions of the large square have true probability less than t

$\rightarrow$ small k is likely to lead to occasional local regions which have estimated probability above t

$\rightarrow$ the larger the value of k, the less likely this is to occur, and hence the lower the overall loss is likely to be

Case (1): $t = 0.5$ contour

***k* = 3**		
	True 1	**True 0**
Pred 1	1513	474
Pred 0	427	1586
Cost	450.5	

***k* = 51**	
True 1	**True 0**
776	892
1164	1168
1028	

Case (2): $t = 0.95$ contour

***k* = 3**		
	True 1	**True 0**
Pred 1	1814	1173
Pred 0	126	197
Cost	178.35	

***k* = 51**	
True 1	**True 0**
1940	2060
0	0
103*	

(* Bayes cost also equals 103)

Conclusions

1) Too much focus on narrow idealised version of real problems

2) The purpose of the classifier must drive the method