Mismatched models, wrong results, and dreadful decisions

-

on choosing appropriate data mining tools

David J. Hand

Imperial College London

KDD 2009, Paris

Data mining is a broad discipline

Many different aspects, including

- prediction of numerical values
- supervised classification
- unsupervised classification
- anomaly detection
- association analysis
- ....

The issues discussed apply to all aspects

But here focus on *supervised classification*

Because

- I know most about this area
- this area illustrates all the aspects
- it is not possible to cover everything

Basic problem:

Given a set of objects, with known characteristics and known class memberships

Construct a rule which will allow one to assign new objects to classes solely on the basis of their characteristics

Examples:

- medical diagnosis and prognosis
- fraud detection
- classifying astronomical objects
- customer value management
- web clickstream analysis
- etc

Many different models

Linear discriminant analysis, quadratic discriminant analysis, naive Bayes, regularised discriminant analysis, logistic regression, SIMCA, DASCO, logistic regression, perceptrons, neural networks, support vector machines, tree classifiers, random forests, nearest neighbour, kernel methods, ...

Obvious question: how to choose ?

Need a *performance criterion* or a *model score*

Roles of performance criteria

- to estimate parameters
  (e.g. regression coefficients, split points, ...)
- to choose model components
  (e.g. variables, transformations of variables, number of hidden nodes, number of trees,..)
- to evaluate models
  (e.g. to compare models, is it good enough, ...)

Performance criteria play a crucial role in data mining

Need simple numerical summaries

Which can be compared automatically

- ***e.g. cannot study millions of ROC curves by eye***

Poor choice of performance criterion yields a poor model

Example:

Error rate is inappropriate with unbalanced classes

For rare disease detection, *minimum error rate* is achieved by classifying everyone as healthy

Example:

Fraud detection criterion using account rather than transaction counts

		True class	
Predicted class		Legitimate	Fraud
	Legitimate	*a*	*b*
	Fraud	*c*	*d*

Different types of classification model performance criteria

1. Problem-based criteria
2. Model fit criteria
3. Classification accuracy criteria

1. Problem based criteria

- speed of construction
- speed of classification
- ability to handle very large data sets
- effectiveness on small n large p problems
- ability to cope with incomplete data
- interpretability
- ability to identify important aspects
- ease of dynamic updating

All problem and context dependent

Example:

Timeliness in fraud detection

nnnnfnnfnnnnnfnnnnnnnnnnnnf

f **= fraudulent transaction**

n **= legitimate transaction**

black **= not flagged as suspicious**

red **= flagged as suspicious**

2. Model fit criteria

Accuracy of probability estimates

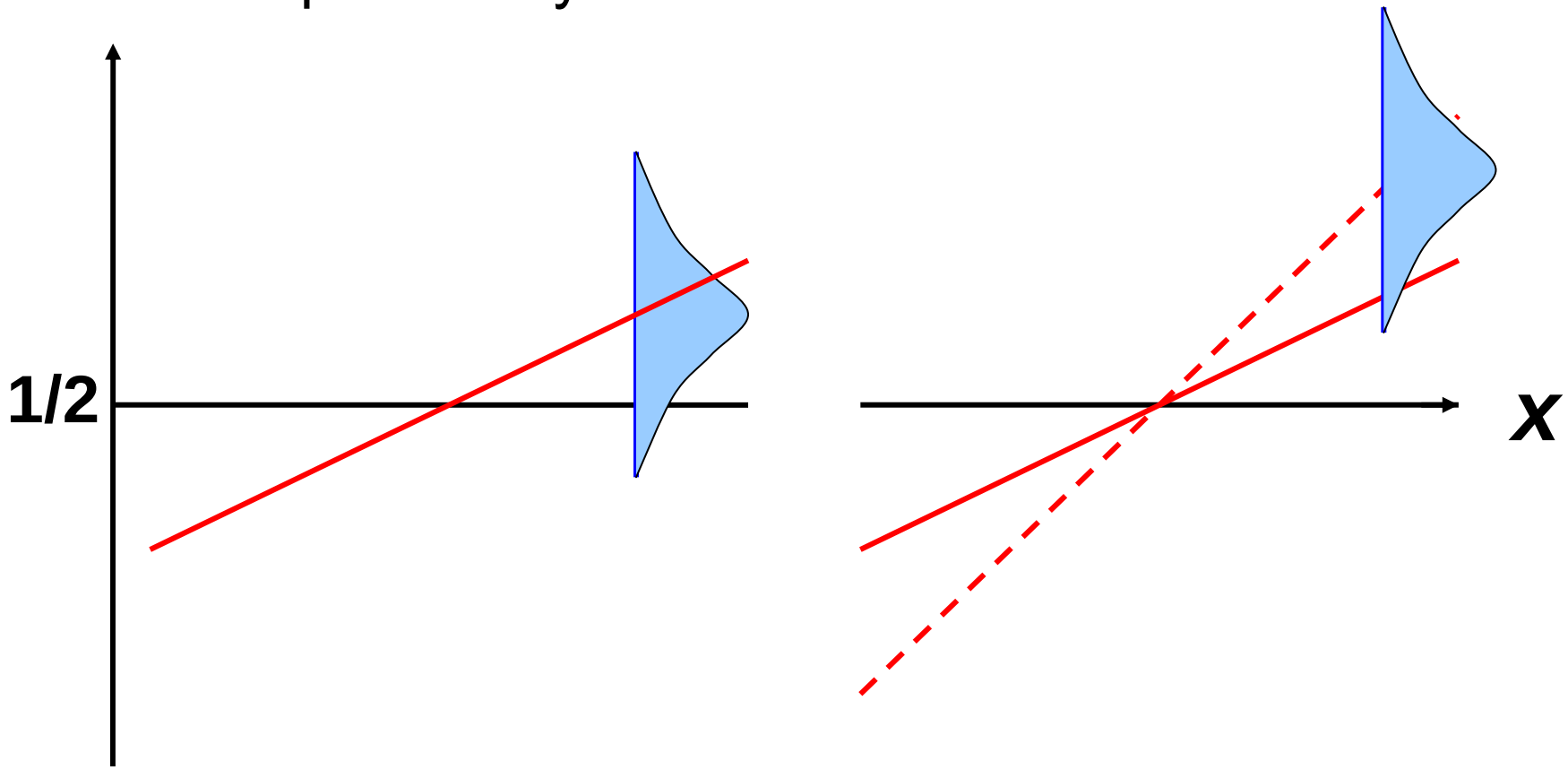
- log-likelihood: $\sum_{i=1}^{n}\left\{c_i \ln\left(p\left(1|x_i\right)\right)-\left(1-c_i\right)\ln\left(1-p\left(1|x_i\right)\right)\right\}$
- Brier score: $\sum_{i=1}^{n}\left(c_i - p\left(1|x_i\right)\right)^2$

But goodness of fit of the model is not the same as accuracy of prediction

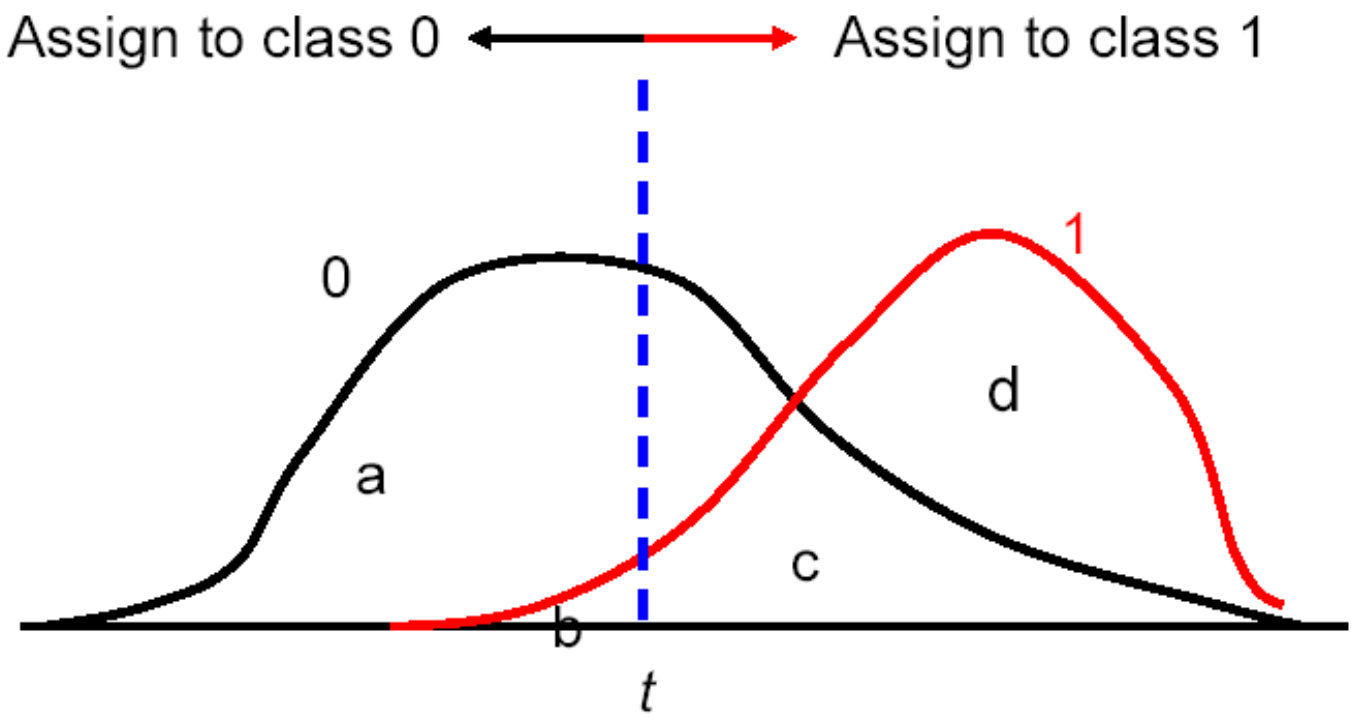
Poorer fit can mean greater accuracy

Example:

Estimated probability of class 1



3. Classification accuracy criteria



		True class	
		0	1
Predicted class	0	a	b
	1	c	d

		True class	
		0	1
Predicted class	0	a	b
	1	c	d

If class 1 are ‘cases’, ‘positives’, ‘abnormals’, etc

$\frac{d}{b+d}$ = ***sensitivity, recall*** $\frac{a}{a+c}$ = ***specificity***

$\frac{d}{c+d}$ = ***positive predicted value, precision***

$\frac{a}{a+b}$ = ***negative predicted value***

Straightforward relationships using class priors, π_0, π_1

Four counts
fixed priors, fixed total → two degrees of freedom

How to combine into single performance measure?

1) Error rate

$$\frac{d}{b+d}.\frac{b+d}{a+b+c+d}+\frac{a}{a+c}.\frac{a+c}{a+b+c+d}=\frac{b+c}{a+b+c+d}$$

$$Se.\pi_1 + Sp.\pi_0 \text{ = } \textbf{\textit{Error rate}}$$

		True class	
		0	1
Predicted class	0	a	b
	1	c	d

2) Chance adjusted proportion correct

$$\left(\frac{a+c}{a+b+c+d}\right)\Big/\left(\frac{b+d}{a+b+c+d}\right)\cdot\left(\frac{a+c}{a+b+c+d}\right)$$

3) F measure

$$\frac{2d}{(b+c)+2d}=\frac{2\times precision\times recall}{(precision+recall)}$$

		True class	
		0	1
Predicted class	0	a	b
	1	c	d

Different performance criteria are generally not monotonically related

But there are relationships

$$\left(1-AUC\right)/2 \le \min Er \le \sqrt{1-AUC}$$

$$KS \le 1-Er$$

$$KS \le AUC \le 1-\left(1-KS\right)^2/2$$

$$Er = Se.\pi_1 + Sp.\pi_0$$

But: all methods based on the classification table assume threshold t has been chosen

How to choose t ?

Let cost of misclassifying a class *i* case be c_i

Then overall loss due to misclassifications is

$$Q = c_0 \pi_0 \left(1 - F_0\left(t\right)\right) + c_1 \pi_1 F_1\left(t\right)$$

Given the costs, a user will always choose *t* to minimise the loss

The threshold *T* which minimises the loss is given by

$$T\left(c_0, c_1\right) \triangleq \arg\min_t \left\{ c_0 \pi_0 \left(1 - F_0\left(t\right)\right) + c_1 \pi_1 F_1\left(t\right) \right\}$$

Leading to a mapping between c_1 / c_0 and *T*

$$c_1 / c_0 = \pi_0 f_0\left(T\right) / \pi_1 f_1\left(T\right)$$

and minimum loss

$$Q = c_0 \pi_0 \left(1 - F_0\left(T\left(c_0, c_1\right)\right)\right) + c_1 \pi_1 F_1\left(T\left(c_0, c_1\right)\right)$$

= the cost weighted error rate

But what should we do if the (future) costs are not known?

(recall the airport anti-terrorist face recognition application)

Equivalent: what should we do if the (future) threshold is not known, since

$$c_1 / c_0 = \pi_0 f_0(T) / \pi_1 f_1(T)$$

Strategy (1):

Make default assumptions

Example 1: $c_0 = c_1 = 1$

Leading to ***error rate***

Example 2: $c_0 = \pi_1, \;\; c_1 = \pi_0$

Leading to ***Youden statistic*** = ***Kolmogorov-Smirnov statistic***

BUT

Costs should be chosen on the basis of the problem

Not picked for mathematical convenience !

Strategy 2:

Average over all possible costs

$$L = \int_0^\infty \int_0^\infty \left\{ c_0 \pi_0 \left(1 - F_0\left(T(\mathbf{c})\right)\right) + c_1 \pi_1 F_1\left(T(\mathbf{c})\right) \right\} u(c_0, c_1)\, dc_0\, dc_1$$

$$L = \int_0^1 \left\{ c \pi_0 \left(1 - F_0\left(T(c)\right)\right) + (1-c) \pi_1 F_1\left(T(c)\right) \right\} w(c)\, dc$$

with $c = \pi_1 f_1(T) \Big/ \left\{ \pi_0 f_0(T) + \pi_1 f_1(T) \right\}$

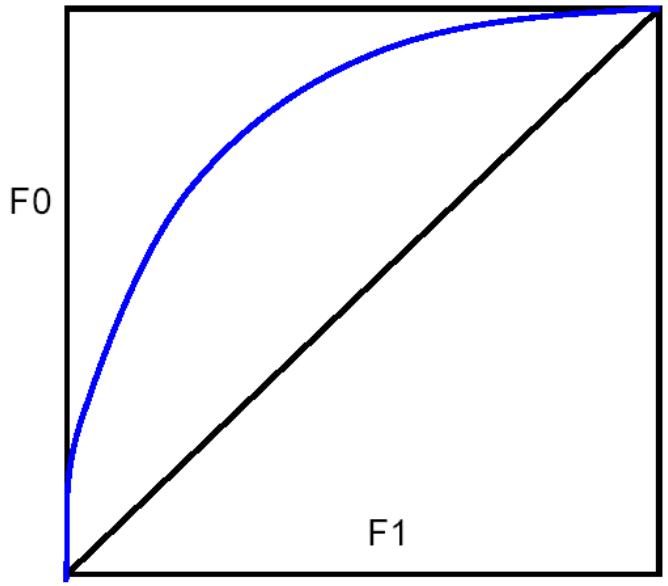
A particular choice of $w(c)$

$$w^*(c) = \pi_0 f_0\left(T(c)\right)\frac{dT(c)}{dc} + \pi_1 f_1\left(T(c)\right)\frac{dT(c)}{dc}$$

leads to

$$L = 1 - 2\pi_0\pi_1 \int_{-\infty}^{\infty} F_0(s)\, f_1(s)\, ds$$

$$= 1 - 2\pi_0\pi_1 AUC$$



To summarise:

$$L = \int_0^1 \left\{ Loss(c) \right\} w(c)\, dc$$

And using $w(c) = w^*(c)$, where

$$w^*(c) = \pi_0 f_0\left(T(c)\right)\frac{dT(c)}{dc} + \pi_1 f_1\left(T(c)\right)\frac{dT(c)}{dc}$$

means *L* is equivalent to the AUC

That is, AUC is equivalent to averaging the loss over a cost weight function $w^*(c)$ which depends on the observed score distributions

$$w^*(c) = \pi_0 f_0\left(T(c)\right)\frac{dT(c)}{dc} + \pi_1 f_1\left(T(c)\right)\frac{dT(c)}{dc}$$

BUT

Knowledge of c, or equivalently knowledge of the costs c_0 and c_1 , must come from information other than the score distributions !

Using the AUC is equivalent to saying your belief in the relative severity which will be assigned to the different types of misclassifications depends on choice of classifier

This is absurd

Like measuring my height in millimetres, and yours in feet, and saying I am taller because my number is larger than yours

But AUC has a several nice interpretations

e.g. AUC is the average value of $F_0(s)$ if $v = F_1(s)$ is chosen from a uniform $x \sim [0,1]$ distribution

$$AUC = \int_0^1 F_0\left(F_1^{-1}(v)\right) dv$$

F1 and F0 are *operating characteristics* of the classifier

I can *choose* F1 (or F0) according to whim

In particular, I might choose a different 1-F1 for different rules

	t		**t'**	
	1-F1(t)	**F0(t)**	**1-F1(t')**	**F0(t')**
Rule 1	95	89	80	90
Rule 2	95	10	80	90

- In contrast, the distribution $w(c)$ is a matter of *belief*, not choice
- It measures the beliefs about the relative severity of the different kinds of misclassification
- $w(c)$ cannot vary from classifier to classifier
- **I cannot say**: *if I use logistic regression then misclassifying a cancer sufferer as healthy is ten times as serious as the reverse,*
  but if I use a tree classifier then it is a hundred times as serious
- c ***is a property of the problem, not the classifier***

Relationship between costs and sensitivity 1-F1 depends on empirical score distributions f_0 *and* f_1

$\Rightarrow$ fundamental ***complementarity***

$\varphi(x)$ ~ choice of *x* distribution for the averaging $1-F_1$

$w(c)$ ~ choice of *c* distribution for the averaging costs

One cannot simultaneously choose φ and *w* independently of the empirical distributions

If one is distribution is independent of the empirical distributions, the other distribution necessarily depends on the classification rule used

Which should we use?

x is a matter of ***choice***, *c* is a ***property of the problem***

AUC uses $\varphi(x) \sim U[0,1]$

$\Rightarrow$ ***AUC averages over cost distributions which vary from classifier to classifier***

[Same applies to PAUC]

*It is OK for **different** researchers to choose **different** w*

- because they are interested in different aspects of performance

*It is wrong for a **single** researcher to choose **different** w for **different** classifiers*

- doing so implies using different measures for different classifiers

Need to choose a weight function which is the same for all classifiers

Choice (1): choose *w* to reflect your personal beliefs in the likely values of *c* (for each problem)

Choice (2): choose a universal standard *w*

- I suggest $beta(2,2)$

Recommendation: *use both*

Summary:

Different application domains tend to favour different performance measures:

- medicine: *Se* and *Sp*
- machine learning: *error rate*
- information retrieval: *F measure*
- statistics: *likelihood*
- consumer credit: *KS* in US, *Gini* in UK
- .....

Conclusions:

- *Don't use measures of model fit if classification is the aim*
- *Choose a measure which best matches objectives*
- *Take special aspects of problem into account*

- *Don't use the Area Under the ROC curve*

Choosing an inappropriate measure can mean

mismatched models,

wrong results,

and dreadful decisions

I have used supervised classification as an example

but the same applies in modelling in general ....

thank you !

A paper describing the fundamental incoherence of the AUC in detail is due to appear in Machine Learning

that paper also describes a coherent alternative