Evaluating supervised classification methods

---

error rate, ROC curves, and beyond

David J. Hand
Imperial College
and
Winton Capital Management

20 August 2014

Structure of talk:

Background

Problem-based performance measures

Classification accuracy measures

Some recent results

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

Example applications:

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

Basic approach (for the two class case):

Calculate score, s, for each object
Compare score with threshold t

Assign to class 1 if $s > t$
Assign to class 0 if $s \leq t$

Many different methods:

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, Parzen kernel methods, quantile classification, …

Poor choice of method

➔ poor conclusions, poor decisions, poor actions

➔ mistakes

in medical diagnosis

in bank loans

in speech recognition

in spam filtering

in ……..

Need a *performance criterion*

- to evaluate models
  (e.g. is it good enough?)
- to choose between models
  (e.g. which one should we use?)
- 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,..)

Some side issues

Overfitting

‘Apparent performance’

- Design / training set
- Test / evaluation set

Cross-validation, leave-one-out, bootstrap, jackknife, ...

Two types of measure

Problem-based performance measures

Classification accuracy measures

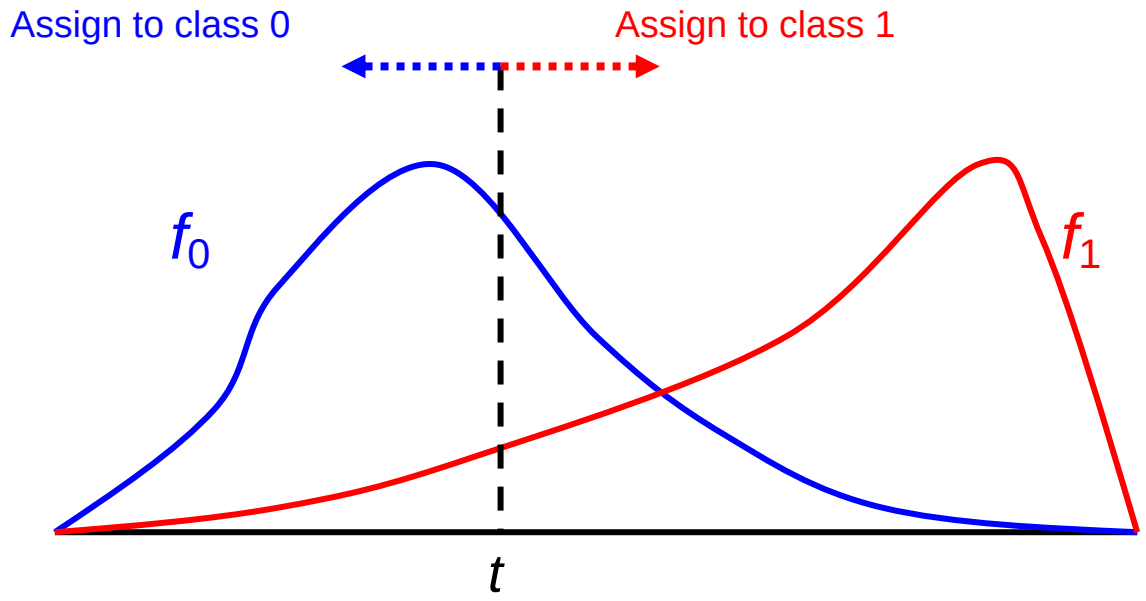
Problem-based performance measures

Problem-specific features

- speed of construction and updating
  (e.g. internet classification problems; spam detection)
- speed of classification
  (e.g. credit card fraud detection)
- handle large data sets (sampling not always possible)
- small-*n*-large-*p*
- incomplete data
- interpretability
- identifying important characteristics
- unbalanced classes
- accuracy of estimates of class membership probs

$f_0(s)$ = distribution of scores for class 0, cdf $F_0(s)$

$f_1(s)$ = distribution of scores for class 1, cdf $F_1(s)$

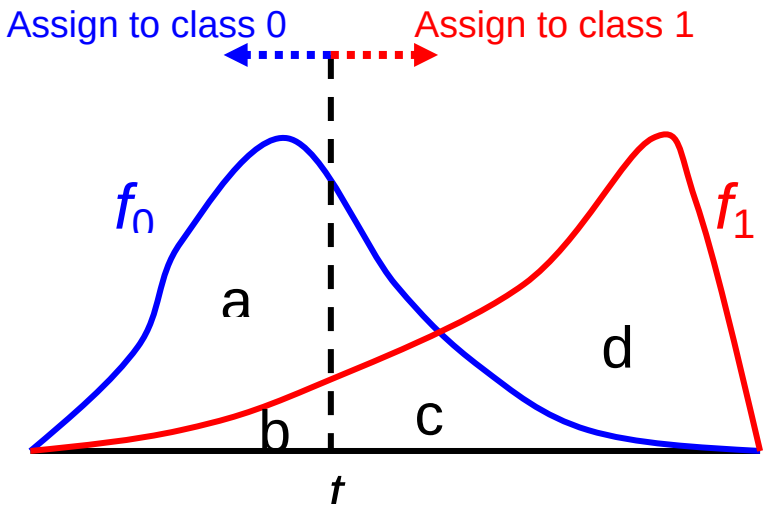


For simplicity in what follows

1) I shall assume the distributions are accurately estimated
2) I shall assume $F_0(s) > F_1(s)$ for all s

A given *t* yields a misclassification table

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



which yields various measures of performance (with different disciplines using different names for the same concept)

Note:

1) need simple numerical summary so can automatically search and compare models

2) given *t*, distance (*s* - *t*) is irrelevant, only sign matters
(could build separate model for severity of error)

This means that the measures must be invariant to monotonic transformations of the score

Calibration

Here I define a classifier as *calibrated* if $\pi_1 f_1(s) / f(s) = s$

That is:

- of objects with score *s*, a proportion *s* belong to class 1;
- probability is *s* that an object with score *s* belongs to class 1;

Note:

calibration is not classification accuracy

e.g. assign everyone score $s = \pi_1$,

so that the classifier is perfectly calibrated

but useless for decision making

To calibrate a classifier, estimate the probability of class 1 membership at each score and apply a monotonic transformation of estimated score so that $\pi_1 f_1(s) / f(s) = s$

[I'm not saying it's easy …]

Since $\pi_0 f_0(s) + \pi_1 f_1(s) = f(s)$

We have that, for calibrated classifiers,

$f_0(s) = f(s) \times (1-s)/\pi_0$ $\qquad$ $f_1(s) = f(s) \times s/\pi_1$

[and hence, for calibrated classifiers: $f_0(s) = f_1(s) \times (1-s)\pi_1 / s\pi_0$]

Henceforth suppose we have calibrated the classifier

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

$c/(c+d)$ proportion of predicted class 1 which are wrong;

$(c+d)/(a+b+c+d)$ proportion assigned to class 1;

$(a+d)/(a+b+c+d)$ proportion correctly classified, p_c;

[$p_E = 1 - p_c$, error rate]

$d/(b+d)$ proportion class 1 correctly class'd.

. . . . and other ratios in various contexts

Kappa statistic: chance adjusted proportion correct

$$K = \frac{p_c - p_{Ch}}{1 - p_{Ch}} = \frac{2(ad - bc)}{(a+b)(a+c) + (c+d)(b+d)}$$

F measure

$$\frac{2d}{(b+c)+2d} = \frac{2}{(True1 \mid Pred1)^{-1} + (Pred1 \mid True1)^{-1}}$$

Matthews coefficient (= Pearson correlation)

$$\frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$$

Ideally the threshold should be chosen on the basis of knowledge

- *what is the desired (estimated) proportion to be assigned to class 1 ?*
- *what is the desired (estimated) class 1 rate amongst those assigned to class 1 ?*
- *etc*

Or the choice can be based on optimising some performance measure

But often (usually) the threshold, *t*, which will be used when the classifier is applied *in the future* is unknown

Because

- we don't have precise details of the future population
- economic circumstances change
- patient populations change

But to evaluate a classifier we must pick a t

Two strategies

STRATEGY 1: Choose *t* to optimise some criterion

STRATEGY 2: Average over a distribution of possible *t*

STRATEGY 1: Choose *t* to optimise some criterion

e.g. 1) Proportion correctly classified

$$\max_t \left(\pi_0 F_0(t) + \pi_1 \left(1 - F_1(t)\right)\right)$$
$$= \pi_1 + \max_t \left(\pi_0 F_0(t) - \pi_1 F_1(t)\right)$$

For fixed class proportions the initial π_1 is irrelevant

so this measure is equivalent to $\max_t \left(\pi_0 F_0(t) - \pi_1 F_1(t)\right)$

e.g. 2) Kolmogorov-Smirnov test statistic

Max proportion correctly classified is equivalent to

$$\max_t \left(\pi_0 F_0(t) - \pi_1 F_1(t) \right)$$

If the two classes had equal sizes, this is

$$\max_t \left(F_0(t) - F_1(t) \right) \quad = \quad KS \text{ measure}$$

1) So the *KS* measure is equivalent to the proportion which would be correctly classified if the two classes were of equal size

2) It's also equivalent to minimising the misclassification loss if you assume that misclassifying a class 1 case is π_0/π_1 times as serious as misclassifying a class 0 case (i.e. $c_1 = c_0\pi_0/\pi_1$)

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

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

$$= c_0\pi_0 \left\{ 1 + \max_t \left(F_0(t) - F_1(t) \right) \right\}$$

3) It's also equivalent to choosing the relative misclassification costs so that misclassifying *all* class 0 points costs the same as misclassifying *all* class 1 points: $c_0\pi_0 = c_1\pi_1$

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

$$= c_0\pi_0 \left\{ 1 + \max_t \left(F_0(t) - F_1(t) \right) \right\}$$

*This KS default choice could be **good** or **bad***

It depends on what you are trying to do

4) Yet Another way of looking at KS

(a) $\max_t \left(F_0(t) - F_1(t) \right)$ is given by solving $f_0(T) = f_1(T)$

(b) If the classifier is calibrated

$$t = P(1|t) = \pi_1 f_1(t) \Big/ \left(\pi_0 f_0(t) + \pi_1 f_1(t) \right)$$

so that $$f_1(T) = \left[\pi_0 T / \pi_1 (1-T) \right] f_0(T)$$

(a) + (b) $\Rightarrow$ so that $T = \pi_1$

This is the case for all calibrated classifiers

Is this the threshold you want?

So maximising the proportion correct, and using the *KS* statistic both make ***default*** assumptions

Are these defaults reasonable for your problem?

In general the relative ***severities*** of the two kinds of misclassification will depend on the particular application

Default choices are unwise

STRATEGY 2: Average over a distribution of possible *t*

Case 1: Choose a distribution for *t* directly

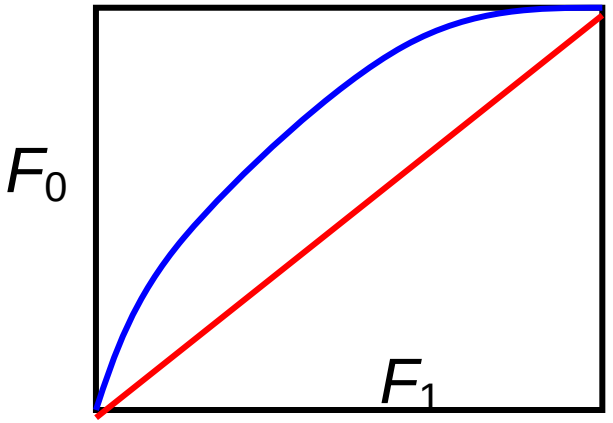
Case 2: Choose a distribution for $W = P(s < t) = F(t)$

Or $W = F_1(t)$, or $W = 1 - F_0(t)$, etc

Mathematically, these are all equivalent, since there is a 1-1 mapping between *t* and *W*

But conceptually...

Case 1: Choose a distribution for *t*



AUC (and *Gini* = 2×*AUC* - *1*) does this

AUC tells us what proportion of class 0 points we get right, on average, when we choose *t* from the distribution $f_1(t)$

$$AUC = \int F_0(t)\, f_1(t)\, dt \quad = 1 - \int F_1(t)\, f_0(t)\, dt = \int \left(1 - F_1(t)\right) f_0(t)\, dt$$

equivalent to the proportion of class 1 points we get right on average when we choose *t* from the distribution $f_0(t)$

The standard interpretation of the *AUC*, that

the probability that a randomly chosen class 0 object will have a score lower than a randomly chosen class 1 object

is irrelevant to most classification problems (≡ *MW* statistic)

Alternative interpretations:

The *AUC* is equivalent to (i.e. 'is a linear transformation of')

The average proportion correctly classified (averaging over a threshold following the class 0 score distribution, the class 1 distribution, or the population distribution)

$$\int P_C \; f_1(t)\,dt = \int \left[\pi_0 F_0(t) + \pi_1 \left(1 - F_1(t)\right) \right] f_1(t)\,dt = \pi_0 AUC + \pi_1/2$$

The AUC is equivalent to (i.e. 'is a linear transformation of')

The average loss due to misclassification, averaging over a threshold following the class 0 score distribution, the class 1 distribution, or the population distribution) and where threshold is chosen to minimise loss for given misclassification costs

$$\int Q(t)\times\left(\pi_0 f_0(t)+\pi_1 f_1(t)\right)dt$$

$$=\int\left[c\pi_0\left(1-F_0(t)\right)+\left(1-c\right)\pi_1 F_1(t)\right]\left(\pi_0 f_0(t)+\pi_1 f_1(t)\right)dt$$

$$c=\pi_1 f_1(t)\Big/\left[\pi_0 f_0(t)+\pi_1 f_1(t)\right]$$

$$=2\pi_0\pi_1\left(1-AUC\right)$$

But I have a ***calibrated*** classifier

So a threshold is a *probability of belonging to class 1*

Why would we consider different distributions of the threshold *probability* for different classifiers?

Why would we think it more likely that we'd want to assign to class 1 those objects which had a probability of being in class 1 greater than 0.7 if we used logistic regression

but greater than 0.5 if we used naive Bayes?

Taking different threshold distributions is equivalent to saying that,

if classifier A were to be used,

then we would be very likely to choose probability 0.9 as our classification threshold

whereas if classifier B were to be used

then we would be very *unlikely* to choose probability threshold 0.9

The threshold probability distribution is a property of the *problem*, not the classifier, and so should be the same for all classifiers applied to the same problem.

But *AUC* (and *Gini*) are averages, over a distribution of the threshold

but where this distribution is different for different classifiers

⇒ ***This is foolish***

It means the choice of measuring instrument depends on the thing being measured

Case 2: Choose a distribution for $P(s < T)$

That is, we will want to assign a proportion
$W = P(s < T) = F(T)$ to class 1
where we aren't sure of what *W* will be,
so we take a distribution over *W*

$$
\begin{aligned}
AUC &= \int F_0(t)\, f_1(t)\, dt \\
&= \pi_1^{-1} \int F_0(t) \left[\pi_0 f_0(t) + \pi_1 f_1(t) \right] dt - \pi_0 / 2\pi_1 \\
&= \pi_1^{-1} \int F_0(t)\, dF(t) - \pi_0 / 2\pi_1 \\
&= \pi_1^{-1} \int F_0\left(F^{-1}(W)\right) dW - \pi_0 / 2\pi_1
\end{aligned}
$$

$$AUC = \pi_1^{-1} \int F_0\left(F^{-1}(W)\right) dW - \pi_0 / 2\pi_1$$

So using the *AUC* is equivalent to saying you think it *equally likely* that you will want to assign to class 1 *any* proportion of the population of objects

You think it equally likely that you will want to assign 99% of the objects to class 1, or just 1%

This is unrealistic

So the AUC is inappropriate measures in this case also

Distinction between Case 1 and Case 2

In case 1, the only relevant information for each object is their score and the threshold

- Any alternative would mean that classifiers which agree that an object had estimated probability *p* of belonging to class 1 could assign it to different classes

In case 2, we're also concerned about the scores of the *other* objects

Case 1 was foolish, and case 2 was unrealistic

⇒ ***AUC is not measuring anything of interest***

So what should we do ?

For Case 1: Use the *H-measure*

The (case 1) problem with the *AUC* was that it averaged over cost or (calibrated) threshold distributions *which depended on the classifier*

Overcome this by using ***the same distribution*** for all classifiers (for a given problem) for the (calibrated) threshold

In an ideal world:

Use a distribution based on your understanding of the problem

- What do you think is the most likely probability threshold to be used in the future?
- What do you think are the extremes of possible thresholds to be used in the future?
- Express in terms of the relative misclassification costs
- Different researchers will (may) have different distributions (*this is as it should be*)

But if it is difficult to decide what this distribution should be

(and it ***will be*** difficult)

Then use a universal standard

(also useful for comparative statements from different researchers)

Suggested form

$$w(t) = beta(1+\pi_1, 1+\pi_0)$$

This particular form has attractive properties

1) Its mode is at π_1, and we saw that the *KS* statistic puts the threshold at $T = \pi_1$

   The choice of $T = \pi_1$ is equivalent to

   making the loss if [all class 0 cases and no class 1 are misclassified]
   equal to the loss if [all class 1 cases and no class 0 are misclassified]

2) Most problems involve *different* costs for the two kinds of misclassification

   - this is especially true when the classes are unbalanced
   - this form puts the threshold at ½ if $\pi_0 = \pi_1$

For Case 2: Use (e.g.)

Average proportion of class 0's amongst those assigned to class 1

averaged over a predetermined distribution, g, *of class 0 proportions (not a function of* f_0 *or* f_1*)*

$$\int \left[\frac{\pi_0 \left(1 - F_0\left(P\right)\right)}{\pi_0 \left(1 - F_0\left(P\right)\right) + \pi_1 \left(1 - F_1\left(P\right)\right)} \right] g\left(P\right) dP$$

But not a uniform distribution

Again a beta distribution might be a good choice

Conclusions

- Use several measures
- Use measures which match the problem and your aims
- Dangers of poor decisions if use poor measures

- *AUC*/*Gini* and *KS* are generally inappropriate

thanks!

Description and code for the *H-measure* can be found on

http://www.hmeasure.net/