Classifier Evaluation in Data Mining: ROC Analysis

José Hernández-Orallo

Dpto. de Sistemas Informáticos y Computación,
Universidad Politécnica de Valencia,

jorallo@dsic.upv.es

Albacete (Spain), January 13rd, 2004.

Outline

- Introduction. The Classification Task and its Evaluation
- Skew-sensitive Evaluation
- ROC Analysis for Crisp Classifiers
- ROC Analysis for Soft Classifiers
- The AUC Metric: the Area Under the ROC Curve
- Relation between AUC and Error. Threshold choice.
- Applications
- Extension to More than Two Classes
- Conclusions

Introduction. The Classification Task and its Evaluation

- Classification.
  - One of the most important tasks in data mining.
  - Goal: to obtain a model, pattern or function that tells between two or more exclusive classes.

- Classification Evaluation.
  - Traditional Measure for Evaluating Classifiers:
    - Error (also inversely *accuracy*): percentage of badly classified instances (wrt. the test set or using cross-validation / bootstrapping).

Introduction. The Classification Task and its Evaluation

- A classifier provides support in decision making (between different actions).

Are we still making decisions in an unscientific way?

- This question is raised by:
  - Swets, J.A., Dawes, R.M., & Monahan, J. (2000). "Better decisions through science" *Scientific American*, 283, 82-87.

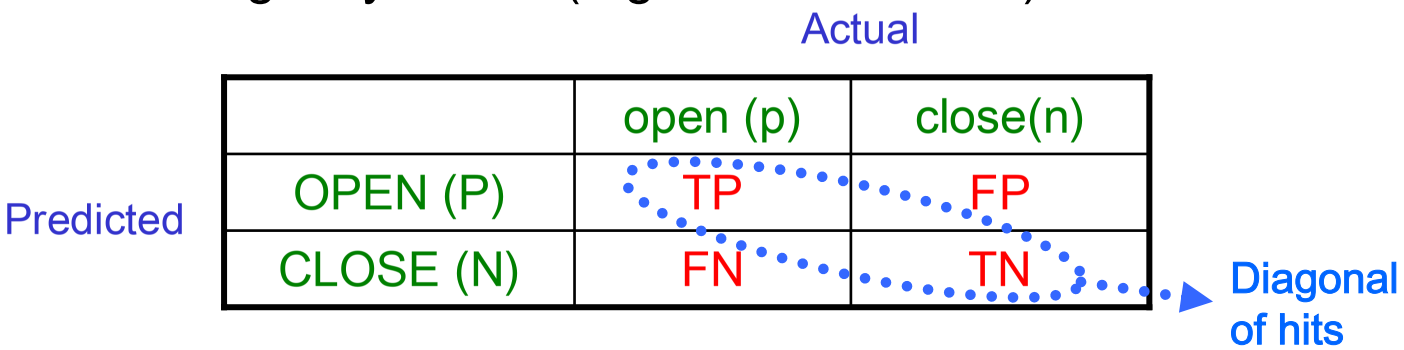
Skew-sensitive Evaluation

- Distribution-sensitive Evaluation:
  - The classes of the problem do not have the same distribution (they are not well balanced, e.g. 50% each of them)
- Example:
  - We have several classifiers c_1, c_2, c_3 that predict whether a refrigeration valve has to be opened or closed in a nuclear plant.
  - In order to evaluate the classifiers we use a dataset obtained during last month, where an operator has been deciding to open or to close the valve.
    - 100,000 examples, from which 99,500 are of class “Close” and 500 are of class “Open”.
  - Let’s say that classifier c_2 always predict “Close” (trivial classifier).
    - Error of c_2: 0.5%.

Is this a good classifier?

Skew-sensitive Evaluation

- Confusion/contingency Matrix (e.g. for the test set):



- From here, several **metrics** have been defined:
  - Pr(P|p) ≈ True Positive Rate: TPR = TP / (TP + FN). ("*recall*" ot "*sensitivity*" ot *"positive accuracy"). *
  - Pr(N|p) ≈ False Negative Rate: FNR = FN / (TP + FN). ("*positive error*")
  - Pr(N|n) ≈ True Negative Rate: TNR = TN / (TN + FP). ("*specificity*" ot "*negative accuracy*").
  - Pr(P|n) ≈ False Positive Rate: FPR = FP / (TN + FP). ("*negative error*")
  - Pr(p|P) ≈ Positive Predictive Value: PPV = TP / (TP + FP). ("*precision*").
  - Pr(n|N) ≈ Negative Predictive Value: NPV = TN / (TN + FN).
  - Macro-average = AVG(TPR, TNR). (The mean can be arithmetic, geometric or other)
  - BREAK-EVEN= (Precision + Recall) / 2 = (PPV + TPR) / 2
  - F-MEASURE= (Precision * Recall) / BREAK-EVEN = 2*PPV*TPR / (PPV + TPR)

Skew-sensitive Evaluation

- Example: (test dataset with 100,000 instances)

Actual

c_1	open	close
OPEN	300	500
CLOSE	200	99000

Pred.

Actual

c_2	open	close
OPEN	0	0
CLOSE	500	99500

Actual

c_3	open	close
OPEN	400	5400
CLOSE	100	94100

ERROR: 0,7%

TPR= 300 / 500 = 60%
FNR= 200 / 500 = 40%
TNR= 99000 / 99500 = 99,5%
FPR= 500 / 99500 = 0.5%
PPV= 300 / 800 = 37.5%
NPV= 99000 / 99200 = 99.8%

Macroavg= (60 + 99.5) / 2 = 79.75%

ERROR: 0,5%

TPR= 0 / 500 = 0%
FNR= 500 / 500 = 100%
TNR= 99500 / 99500 = 100%
FPR= 0 / 99500 = 0%
PPV= 0 / 0 = INDEFINIDO
NPV= 99500 / 10000 = 99.5%

Macroavg= (0 + 100) / 2 = 50%

ERROR: 5,5%

TPR= 400 / 500 = 80%
FNR= 100 / 500 = 20%
TNR= 94100 / 99500 = 94.6%
FPR= 5400 / 99500 = 5.4%
PPV= 400 / 5800 = 6.9%
NPV= 94100 / 94200 = 99.9%

Macroavg= (80 + 94.6) / 2 = 87.3%

Sensitivity

Specificity

Which classifier is best?

Skew-sensitive Evaluation

- Cost-sensitive Evaluation:
  - In many circumstances, not all the errors produced by a predictive model have the same consequences:
    - Example: keeping a valve closed in a nuclear plant when it should be open, can provoke an explosion, while opening a valve when it should be closed, can provoke a stop.
  - Cost matrix:

		Actual	
		open	close
Predicted	OPEN	0	100€
	CLOSE	2000€	0

  - The important thing is not to obtain the “classifier” with fewer errors but the one with lowest cost.
  - From this matrix, we can compute the cost of a classifier.
    - Classifiers are evaluated with these costs.
    - The classifier with lowest cost is chosen.

Skew-sensitive Evaluation

- Examples:

Confusion Matrices

Actual

Pred

c_1	open	close
OPEN	300	500
CLOSE	200	99000

Actual

c_2	open	close
OPEN	0	0
CLOSE	500	99500

Actual

c_3	open	close
OPEN	400	5400
CLOSE	100	94100

Actual

Predicted

	open	close
OPEN	0	100€
CLOSE	2000€	0

Cost Matrix

Resulting Matrices

c_1	open	close
OPEN	0€	50,000€
CLOSE	400,000€	0€

TOTAL COST: 450,000€

c_2	open	close
OPEN	0€	0€
CLOSE	1,000,000€	0€

TOTAL COST: 1,000,000€

c_3	open	close
OPEN	0€	540,000€
CLOSE	200,000€	0€

TOTAL COST: 740,000€

Skew-sensitive Evaluation

- What affects the final cost?
  - For two classes. It depends on a **context** or **skew**:
    - The cost of false positives and false negatives: FPcost and FNcost
    - The percentage of examples of the negative class wrt. the examples of the positive class. (*Neg / Pos*).
    - We compute: (for the previous examples)

$$\frac{FPcost}{FNcost} = \frac{100}{2000} = \frac{1}{20} \qquad \frac{Neg}{Pos} = \frac{99500}{500} = 199 \qquad slope = \frac{1}{20} \times 199 = 9.95$$

  - For two classes, the value "slope" is sufficient to determine which classifier is best.

Classif. 1: FNR= 40%, FPR= 0.5%
Cost per unit =
1 x 0.40 + 9.95 x 0.005 = 0.45

Classif. 2: FNR= 100%, FPR= 0%
Cost per unit =
1 x 1 + 9.95 x 0 = 1

Classif. 3: FNR= 20%, FPR= 5.4%
Cost per unit =
1 x 0.20 + 9.95 x 0.054 = 0.74

ROC Analysis of Crisp Classifiers

- The classifier with lowest error rate is frequently not the best classifier.
- The context or *skew* (the **class distribution** and the **costs** of each error) determine the goodness of classifiers.

- PROBLEM:
  - In many circumstances, *until the application time,* we do not know the class distribution and/or it is difficult to estimate the cost matrix. E.g. a spam filter.
  - But models are learned *before*.

- **ROC** (*Receiver Operating Characteristic*) **Analysis**.
  - First used in the WW2 to evaluate radars, later was used to study the response of transistors, and further developed for medical diagnosis applications from the seventies. It begins to be known in data mining in the late nineties.

ROC Analysis of Crisp Classifiers

- The ROC Space
  - Each column of the confusion matrix is normalised: TPR, FNR TNR, FPR.

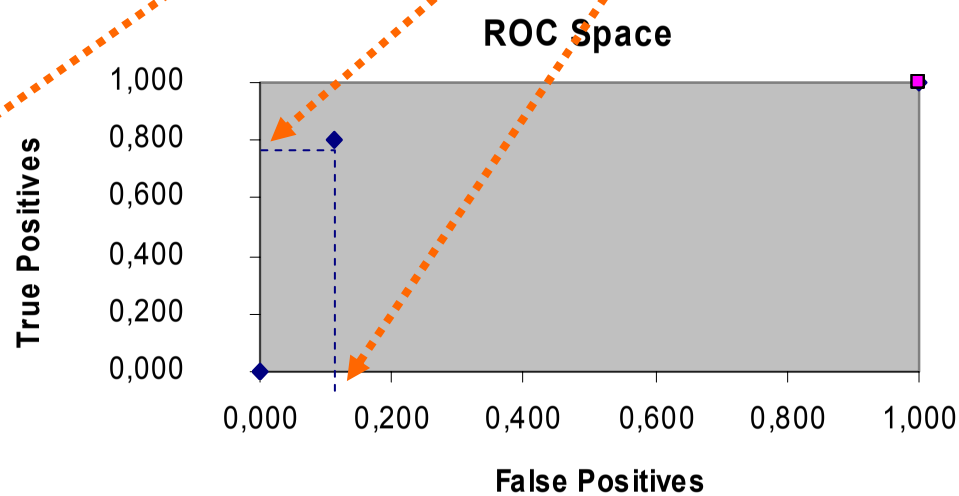
Actual

Pred	open	close
OPEN	400	12000
CLOSE	100	87500

Actual

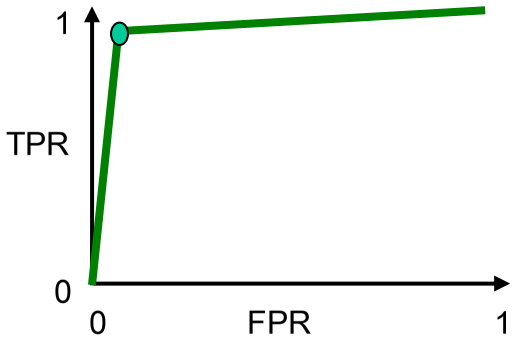
Pred	open	close
OPEN	0.8	0.121
CLOSE	0.2	0.879

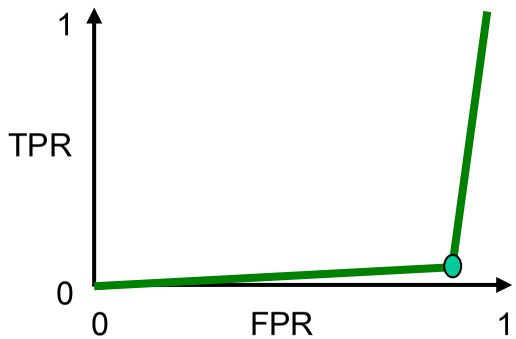
TPR= 400 / 500 = 80%
FNR= 100 / 500 = 20%
TNR= 87500 / 99500 = 87.9%
FPR= 12000 / 99500 = 12.1%

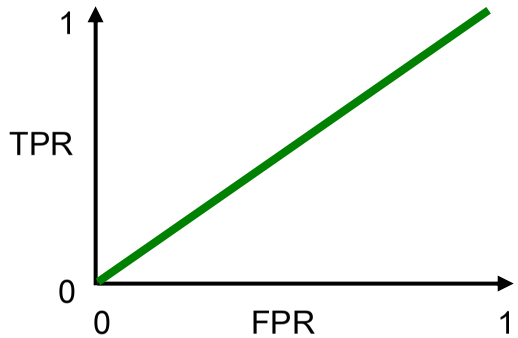


ROC Analysis of Crisp Classifiers

- ROC space: good and bad classifiers.



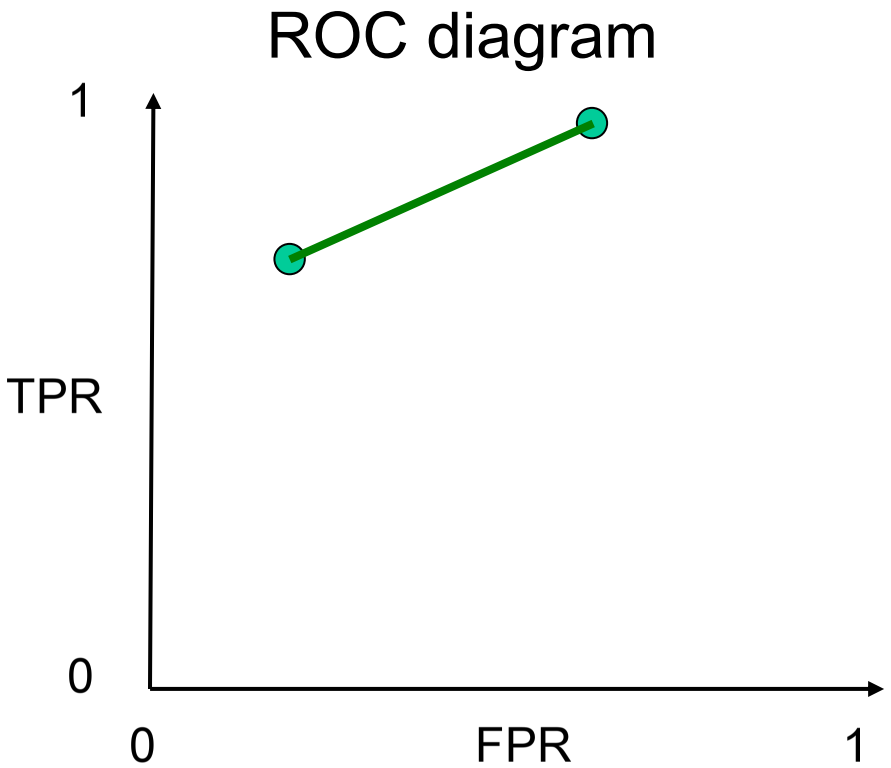




- Good classifier.
  - High TPR.
  - Low FPR.
- Bad classifier.
  - Low TPR.
  - High FPR.
- Bad classifier (real picture).

ROC Analysis of Crisp Classifiers

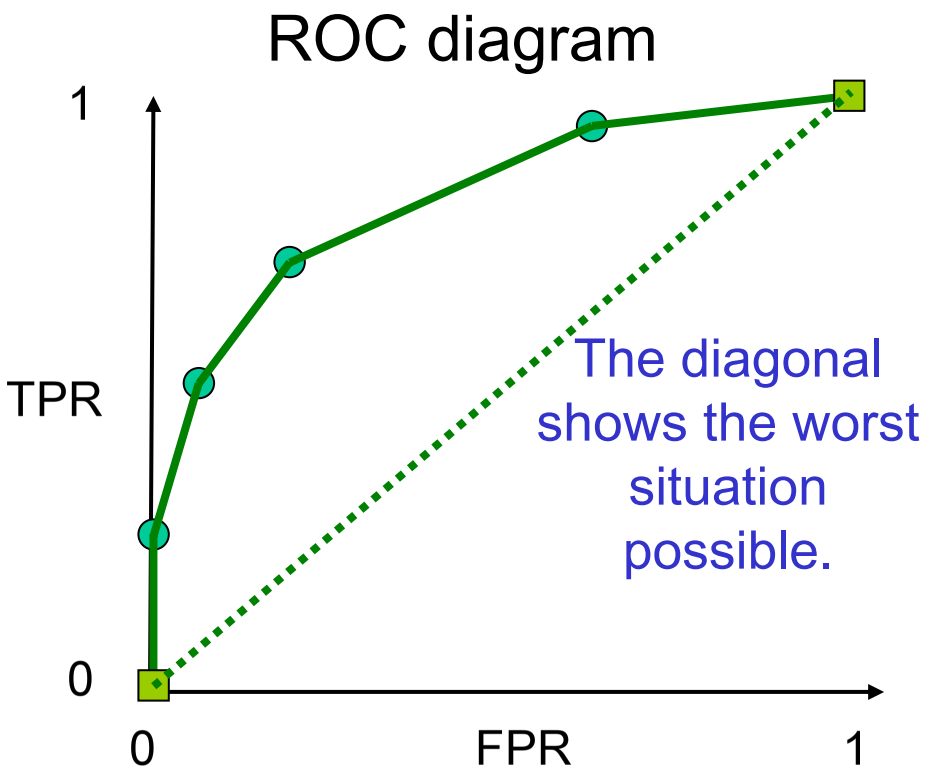
- The ROC Curve ROC. “Continuity”.



- Given two classifiers:
  - We can construct any “intermediate” classifier just randomly weighting both classifiers (giving more or less weight to one or the other).
  - This creates a “continuum” of classifiers between any two classifiers.

ROC Analysis of Crisp Classifiers

- ROC Curve. Construction.

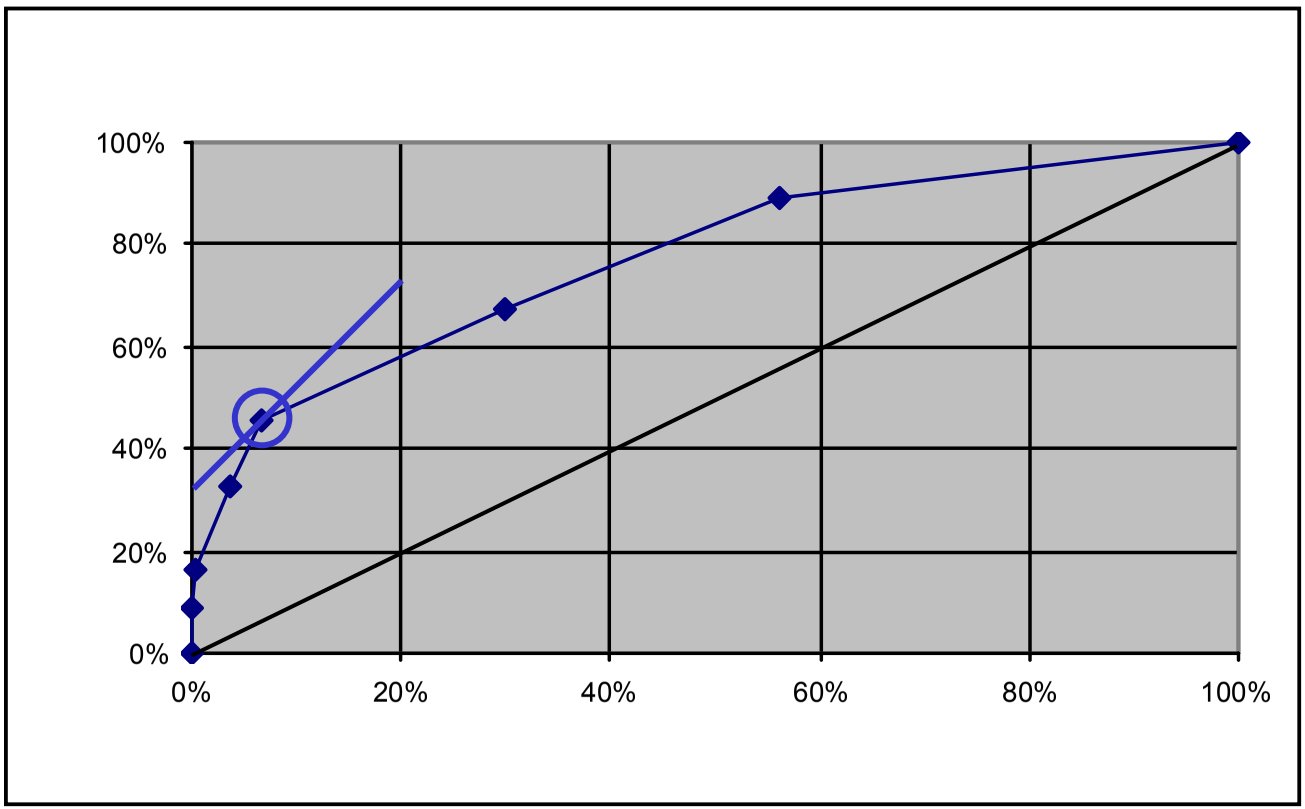


- Given several classifiers:
  - We construct the convex hull of their points (FPR,TPR) as well as the two trivial classifiers (0,0) and (1,1).
  - The classifiers below the ROC curve are discarded.
  - The best classifier (from those remaining) will be selected in application time…

We can discard those which are below because there is no combination of class distribution / cost matrix for which they could be optimal.

ROC Analysis of Crisp Classifiers

- In the **context of application**, we choose the optimal classifier from those kept. Example 1:



Context (skew):

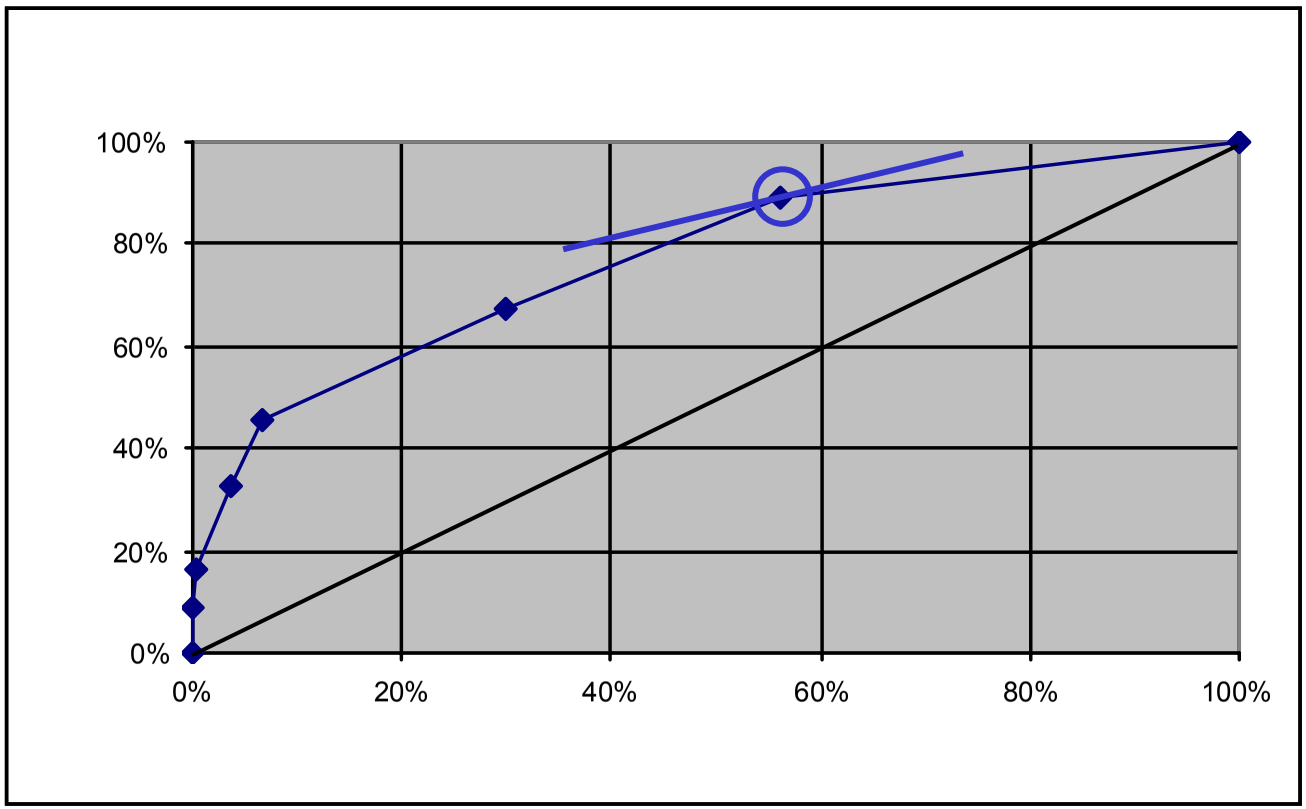
$$\frac{FPcost}{FNcost} = 1/2$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/2 = 2$$

ROC Analysis of Crisp Classifiers

- In the **context of application**, we choose the optimal classifier from those kept. Example 2:



Context (skew):

$$\frac{FPcost}{FNcost} = 1/8$$

$$\frac{Neg}{Pos} = 4$$

$$slope = 4/8 = .5$$

ROC Analysis of Crisp Classifiers

- What have we learned from this?
  - The optimality of a classifier depends on the class distribution and the error costs.
  - From this **context** / **skew** we can obtain the "slope", which characterises this context.
    - If we know this context, we can select the best classifier, multiplying the confusion matrix and the cost matrix.
    - If we don't know this context in the learning stage, by using ROC analysis we can choose a subset of classifiers, from which the optimal classifier will be selected when the context is known.

Can we go further than this?

ROC Analysis of Soft Classifiers

- Crisp and Soft Classifiers:
  - A “crisp” classifier predicts a class between a set of possible classes.
  - A “soft” classifier (probabilistic) predicts a class, but accompanies each prediction with an estimation of the reliability (confidence) of each prediction.
    - Most learning methods can be adapted to generate this confidence value.
- A special kind of soft classifier is a class probability estimator.
  - Instead of predicting “*a*”, “*b*” or “*c*”, it gives a probability estimation for “*a*”, “*b*” or “*c*”, i.e., “p_a”, “p_b” and “p_c”. Example:
    - Classifier 1: p_a= 0.2, p_b= 0.5 and p_c= 0.3.
    - Classifier 2: p_a= 0.3, p_b= 0.4 and p_c= 0.3.
  - Both predict *b*, but classifier 1 is “surer”.

ROC Analysis of Soft Classifiers

- "Rankers":
  - Whenever we have a probability estimator for a two-class problem:
    - $p_a = x$, then $p_b = 1 - x$.
  - It is only necessary to specify the probability of one class.
  - Let's call one class 0 (neg) and the other class 1 (pos).
  - A *ranker* is a soft classifier that gives a value between 0 and 1 of the probability of class 1. This value is also called "*score*" and determines whether the prediction is closer to class 0 or class 1.
  - Examples:
    - Probability of a customer buying a product.
    - Probability of a message being spam.
    - ...

ROC Analysis of Soft Classifiers

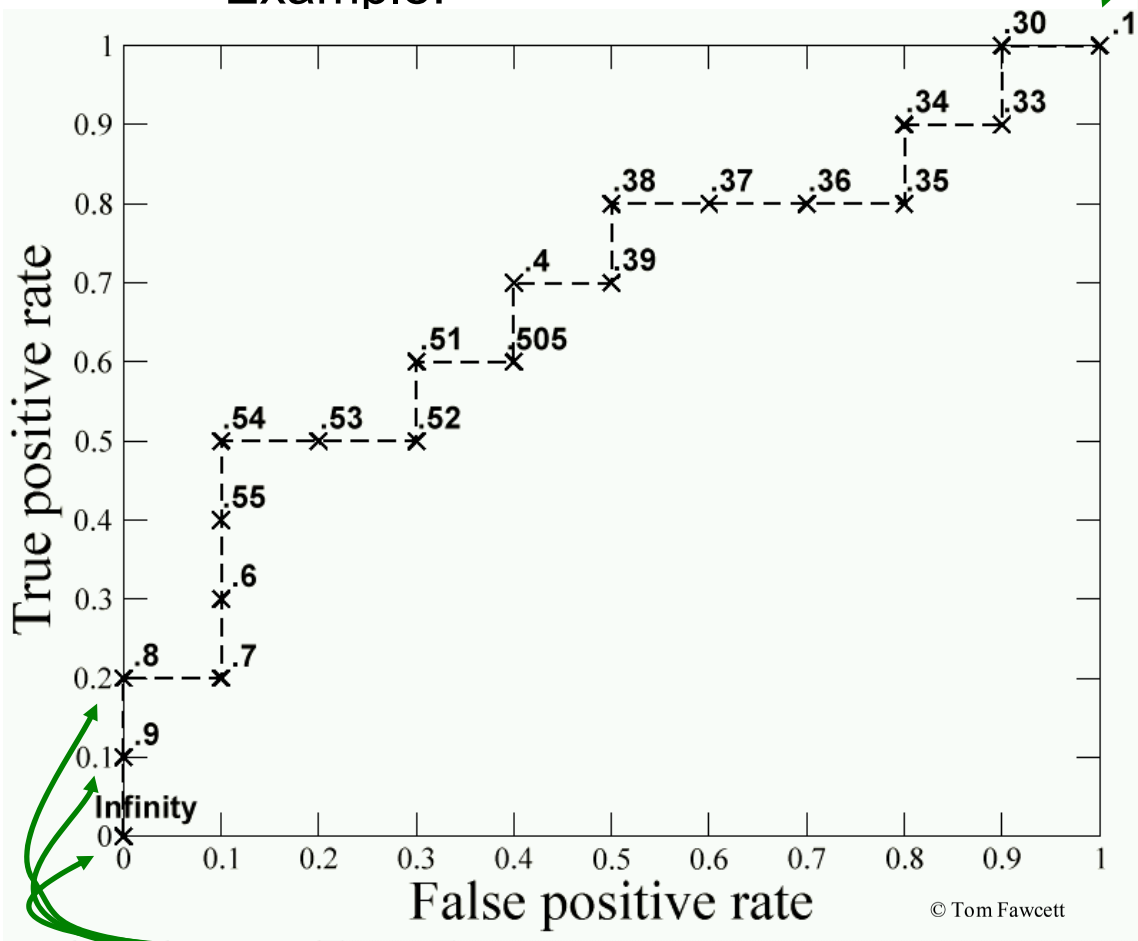
- ROC Curve of a Soft Classifier:
  - A soft classifier can be converted into a crisp classifier using a threshold.
    - Example: “if score > 0.7 then class A, otherwise class B”.
    - With different thresholds, we have different classifiers, giving more or less relevance to each of the classes (without need of oversampling and undersampling).
  - We can consider each threshold as a different classifier and draw them in the ROC space. This generates a curve…

We have a “curve” for just one soft classifier

- This curve is like a stair (the convex hull is not done generally).

ROC Analysis of Soft Classifiers

- ROC Curve of a Soft Classifier:
  - Example:



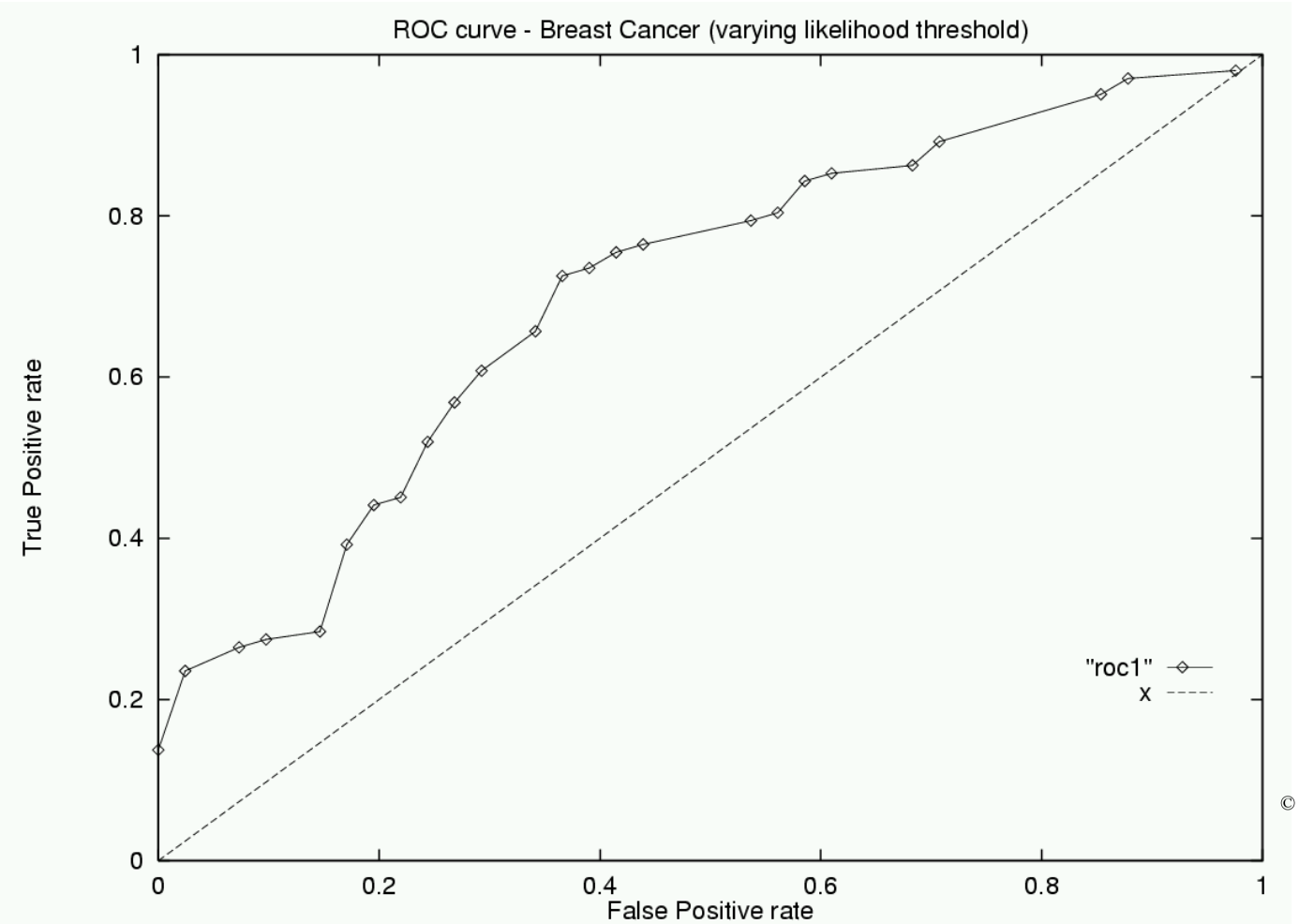
Actual Class

Predicted Class

Inst#	Class	Score					
1	p	.9	n	p	p		p
2	p	.8	n	n	p		p
3	n	.7	n	n	n		p
4	p	.6	n	n	n		p
5	p	.55	n	n	n		p
6	p	.54	n	n	n		p
7	n	.53	n	n	n		p
8	n	.52	n	n	n		p
9	p	.51	n	n	n		p
10	n	.505	n	n	n	...	p
11	p	.4	n	n	n		p
12	n	.39	n	n	n		p
13	p	.38	n	n	n		p
14	n	.37	n	n	n		p
15	n	.36	n	n	n		p
16	n	.35	n	n	n		p
17	p	.34	n	n	n		p
18	n	.33	n	n	n		p
19	p	.30	n	n	n		p
20	n	.1	n	n	n		p

ROC Analysis of Soft Classifiers

- ROC Curve of a Soft Classifier:

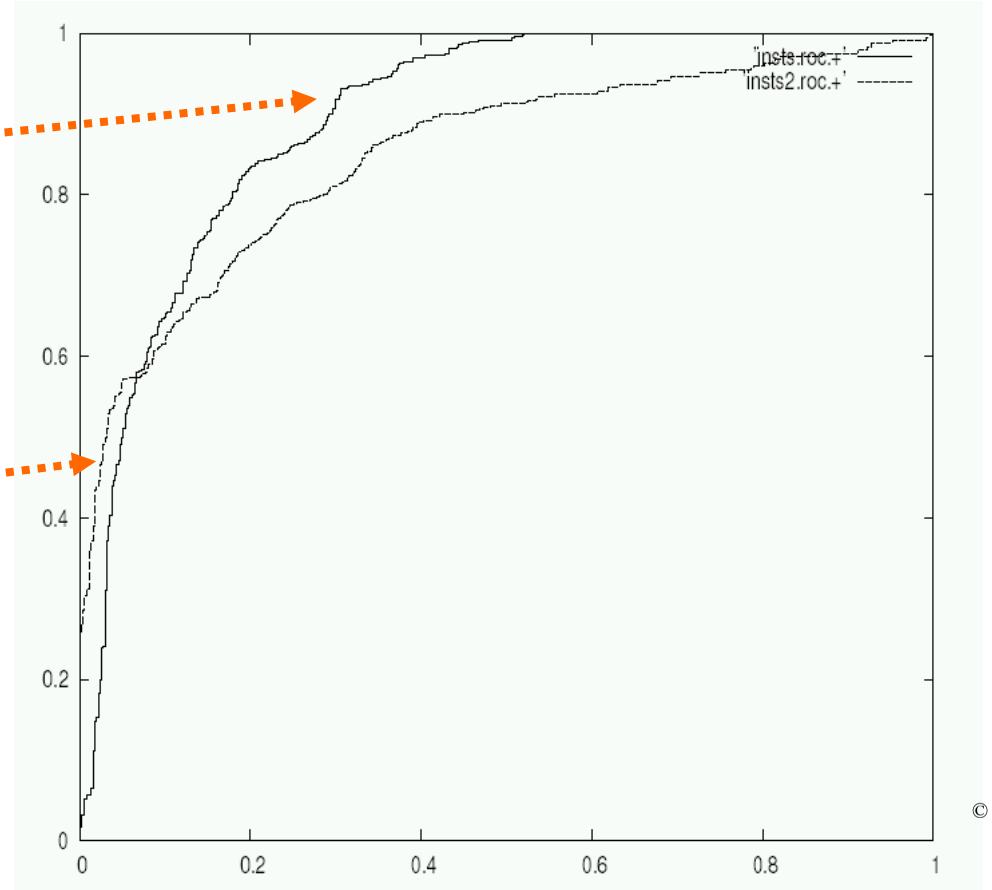


ROC Analysis of Soft Classifiers

- ROC Analysis of several “soft” classifiers:

In this zone the best classifier is “insts”

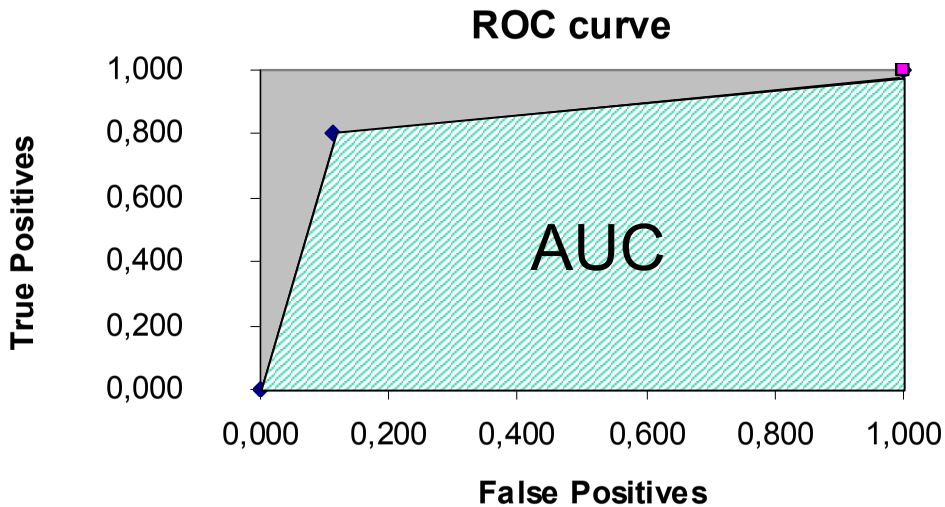
In this zone the best classifier is“insts2”



- We must preserve the classifiers that have at least one “best zone” and then behave in the same way as we did for crisp classifiers.

The “AUC” Metric: the Area Under the ROC Curve

- What if we want to select just one classifier?
  - The classifier with greatest *Area Under the ROC Curve* (AUC) is chosen.



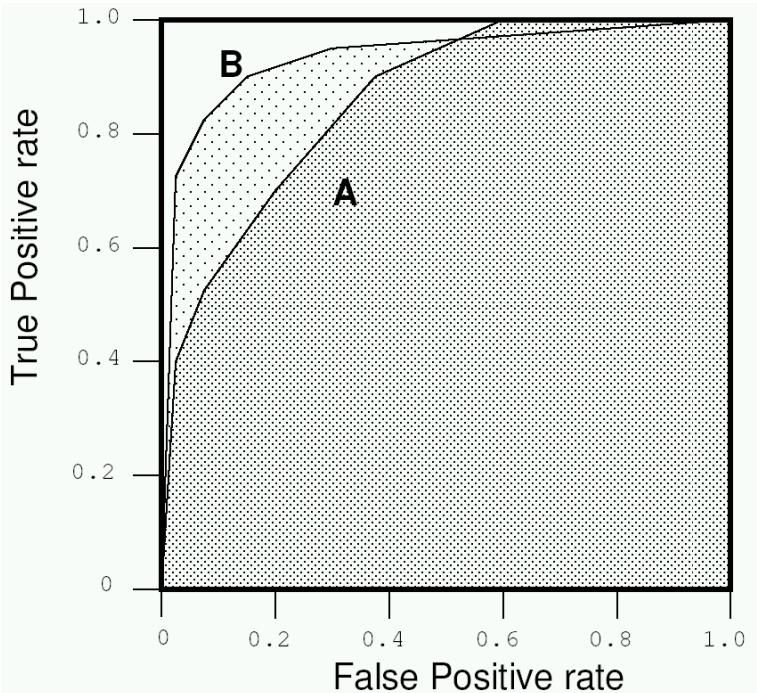
  - For crisp classifiers it is equivalent to the macroaverage.

Alternative to error for evaluating classifiers

- A data mining / learning method will be better if it generates classifiers with high AUC.

The “AUC” Metric: the Area Under the ROC Curve

- What if we want to select just one *soft* classifier?
  - The classifier with greatest *Area Under the ROC Curve* (AUC) is chosen.



In this case we select B.

But for the “soft” case we have surprises…

The "AUC" Metric: the Area Under the ROC Curve

- The AUC and the Wilcoxon-Mann-Whitney (WMW) (Wilcoxon 1945) (Mann & Whitney 1947) statistic are equivalent.
  - The WMW test is useful to determine whether one of two random variables is stochastically greater than the other.

$$P[X > Y]$$

  - If we choose X as the examples of one class and Y as the examples of the other class, and the values X and Y as the estimated *score* by the classifier, we have that the AUC is equivalent to the WMW statistic.

Should I be happy of this?

The AUC really estimates the probability that, if we choose an example of class 1 and an example of class 0, the classifier will give more *score* to the first one than to the second one.

(Care! This does not mean that it classifies well both examples). But:

- There does exist a threshold from which it can classify both correctly.
- It cannot classify both wrongly, for any threshold whatsoever.

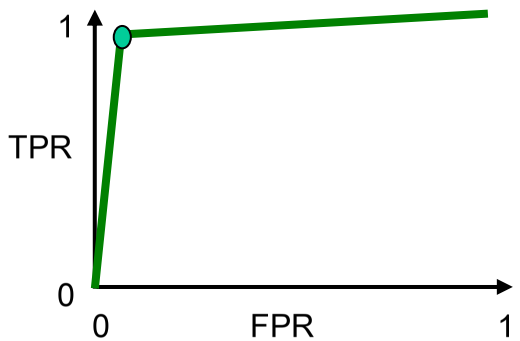
The "AUC" Metric: the Area Under the ROC Curve

- The AUC metric for soft classifers or *rankers*.
  - Evaluates how well a classifier performs a ranking of its predictions.

    or, in other words,
  - Evaluates how well a classifier is able to sort its predictions according to the confidence it assigns to them.
  - The "rankings" of predictions are fundamental in many applications:
    - Fraud detection.
    - Mailing campaign design.
    - Spam filtering.
    - Failure detection, medical diagnosis.
    - …
  - And many other data mining methods:
    - Combination of classifiers.
    - Collaborative Methods. *Recommender systems*…

Relation between AUC and error. Threshold choice

- We have seen that AUC is a better evaluation measure than error rate. But, which relation is there between both?
  - Logically, an AUC closer to 1 will give an error rate close to 0.

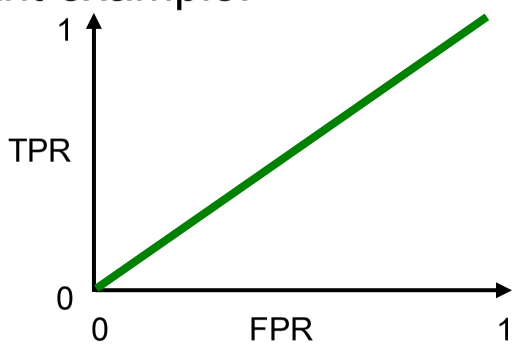


- Whatever skew we have, a low error is expected.

  - But an error rate closer to 0 does not ensure an AUC close to 1.
    - Let's recall the nuclear plant example:

c_2	open	Close
OPEN	0	0
CLOSE	500	99500

ERROR: 0.005



AUC = 0.5
(Minimum possible value)
Macroavg= AUC =
(0 + 100) / 2 = 50%

Relation between AUC and error. Threshold choice

- Many learning methods are soft by definition and are converted into crisp for performing the classification.
  - For example, a Naïve Bayes classifier, for a two-class problem (*a* and *b*), estimates two probabilities:

    $$P(a|x) \text{ and } P(b|x).$$

  - The classification rule is the following one:

    If $P(a|x) > P(b|x)$ then **class a**
    Otherwise **clase b**

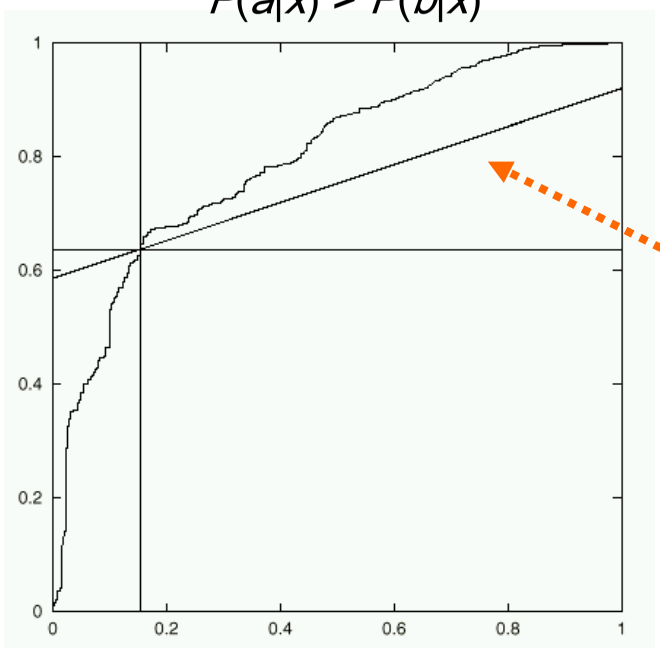
  - This rule "wastes" a Bayesian classifier. But, which other rule could be used? (Lachiche & Flach 2003)
    - First, we convert the probabilities into *scores*, as follows.

      $$s(x) = P(a|x) / P(b|x)$$

Relation between AUC and error. Threshold choice

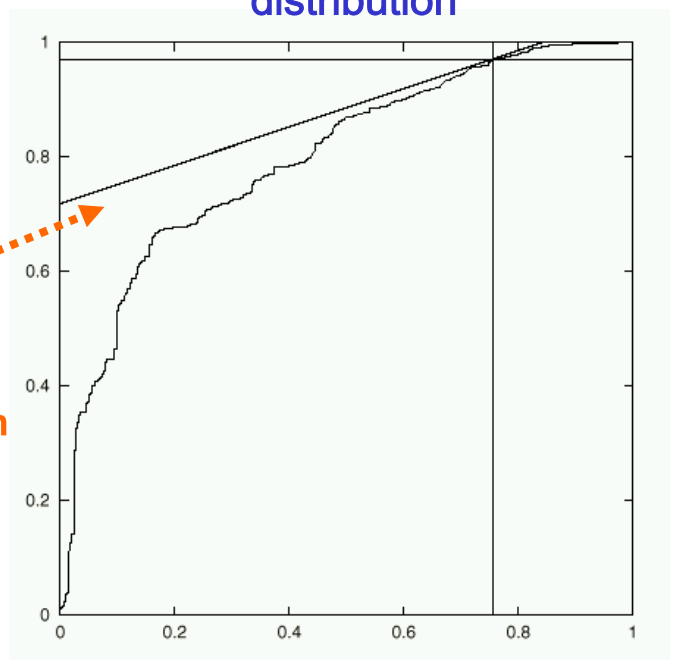
- Now, simply, let's do ROC analysis.
  - We draw the ROC curve using the score.
  - We compute the slope from the a priori class distribution (or the distribution inapplication time, and costs if we have them).
  - We choose the threshold between both classes for $s(x)$.

Choice using rule
$P(a|x) > P(b|x)$



Slope given by the original distribution (train or test).

Choice using ROC analysis and original distribution



Relation between AUC and error. Threshold choice

- Example of results (*accuracy*) for 25 datasets:

  (Lachiche & Flach 2003)

  - Results are improved by just changing the decision rule of the Naïve Bayes Classifier.

Settings	cl.	1BC	1BC opt.
Audiology	24	**67.5%**	**78.5%**
Bridges 2 (t-or-d)	2	85.3%	88.2%
Bridges 2 (material)	3	86.8%	84.9%
Bridges 2 (span)	3	67.4%	67.4%
Bridges 2 (rel-l)	3	68.0%	68.9%
Bridges 2 (type)	7	58.5%	59.4%
Car	4	**85.3%**	**88.8%**
Credit	2	86.5%	**85.5%**
Dermatology	6	97.5%	97.3%
Ecoli	8	83.6%	**82.1%**
Flag (religion)	8	64.9%	62.4%
Flare 2 (common)	9	**76.1%**	**82.8%**
Flare 2 (moderate)	9	**91.5%**	**96.3%**
Flare 2 (severe)	9	**97.5%**	**99.4%**
Glass	7	67.3%	**65.4%**
Horse-colic (surgical)	2	79.6%	79.6%
Horse-colic (site)	12	**40.2%**	**45.7%**
Horse-colic (type)	5	55.1%	56.8%
Horse-colic (subtype)	4	**56.0%**	**63.0%**
Horse-colic (code)	11	37.8%	38.3%
Image segmentation	7	88.9%	**88.4%**
Mushroom	2	**95.5%**	**98.1%**
Nursery	5	**90.3%**	**91.5%**
Post-operative	3	70.0%	71.1%
Vote	2	**90.1%**	**88.0%**

Relation between AUC and error. Threshold choice

- In the other way round, it would be interesting to convert data mining methods that obtain crisp classifiers into soft classifiers.
  - This would allow its use as rankers, for combination, ...
  - This would make it possible to choose a better threshold and improve results.
- Many classical methods have been redesigned and re-thought. For instance, for decision trees:
  - Decision trees are learned using AUC as splitting criterion (Ferri et al. 2002).
  - The leaf probabilities are smoothed (using Laplace correction or other more sophisticated smoothing methods) and pruning is shown to be counterproductive to obtain good AUC measures (Provost & Domingos 2003) (Ling & Yan 2003) (Ferri et al. 2003a).

Applications

- An example: “mail campaign design”:
  - A company wants to make a mailed offer to increase the sales of one of its products. In case of positive reply customers buy products with a mean value of 100€. If 55% are production costs (fix and variable), we have that for each positive reply there is a mean gain of 45€.
  - Each post sent has a cost of 1€ (mail, leaflet) and the whole of the campaign (indep. of the number of postings) would have a base cost of 20,000€.
  - With 1,000,000 customers, for which, with a first sample of 1,000 customers, we have estimated that 1% replies (buys)...

How should we act?

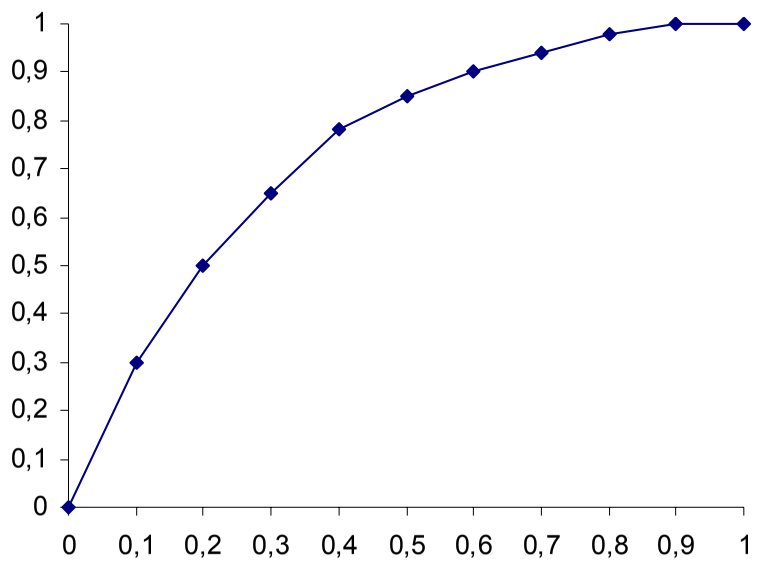
Applications

- "mail campaign design":
  - Let's calculate the costs / benefits.
    - COSTS (assuming that we send the campaign to all customers):
      - 20,000€ design of the campaign.
      - 1,000,000 x 1€ = 1,000,000€.
      - TOTAL: 1,020,000€
    - BENEFITS
      - 1% of replies from 1,000,000 are 10,000 replies, 45€ each one.
      - TOTAL: 450,000€
  - More costs than benefits → Cancelled campaign.

Have we done well?

Applications

- "mail campaign design":
  - Let us train a soft classifier with the sample and let us draw its ROC curve (with a previously removed part for test set).



Cost Matrix

Buys

Sent

	no	sí
NO	0€	-45€
SÍ	1€	-44€

But the confusion matrix has an *impossible* cell: "Not sent and bought"

- Additionally, the classifier is not very good (closer to the diagonal) to know which customers will buy and which customers won't.
- Even though, we can do many things...

Applications

- "mail campaign design":
  - We can use the classifier to determine who to send the offer.

DECILE	GAINS	CUM	LIFT	SIZE	SIZE(YES)	SIZE(NO)	PROFIT
0%	0.0%	0%	0.000	0	0	0	-($20,000)
10%	30.0%	30%	3.000	100,000	3,000	97,000	+ $15,000
20%	20.0%	50%	2.500	200,000	5,000	195,000	+ $5,000
30%	15.0%	65%	2.167	300,000	6,500	293,500	-($27,500)
40%	13.0%	78%	1.950	400,000	7,800	392,200	-($69,000)
50%	7.0%	85%	1.700	500,000	8,500	491,500	-($137,500)
60%	5.0%	90%	1.500	600,000	9,000	591,000	-($215,000)
70%	4.0%	94%	1.343	700,000	9,400	690,600	-($297,000)
80%	4.0%	98%	1.225	800,000	9,800	790,200	-($379,000)
90%	2.0%	100%	1.111	900,000	10,000	890,000	-($470,000)
100%	0.0%	100%	1.000	1,000,000	10,000	990,000	-($570,000)

Cost of Campaign
20,000 --> 20,000
100,000 x 1 --> 100,000
Total: 120,000

Benef. Campaign
3,000 x 45 --> 135,000

Net Benef.: 15,000

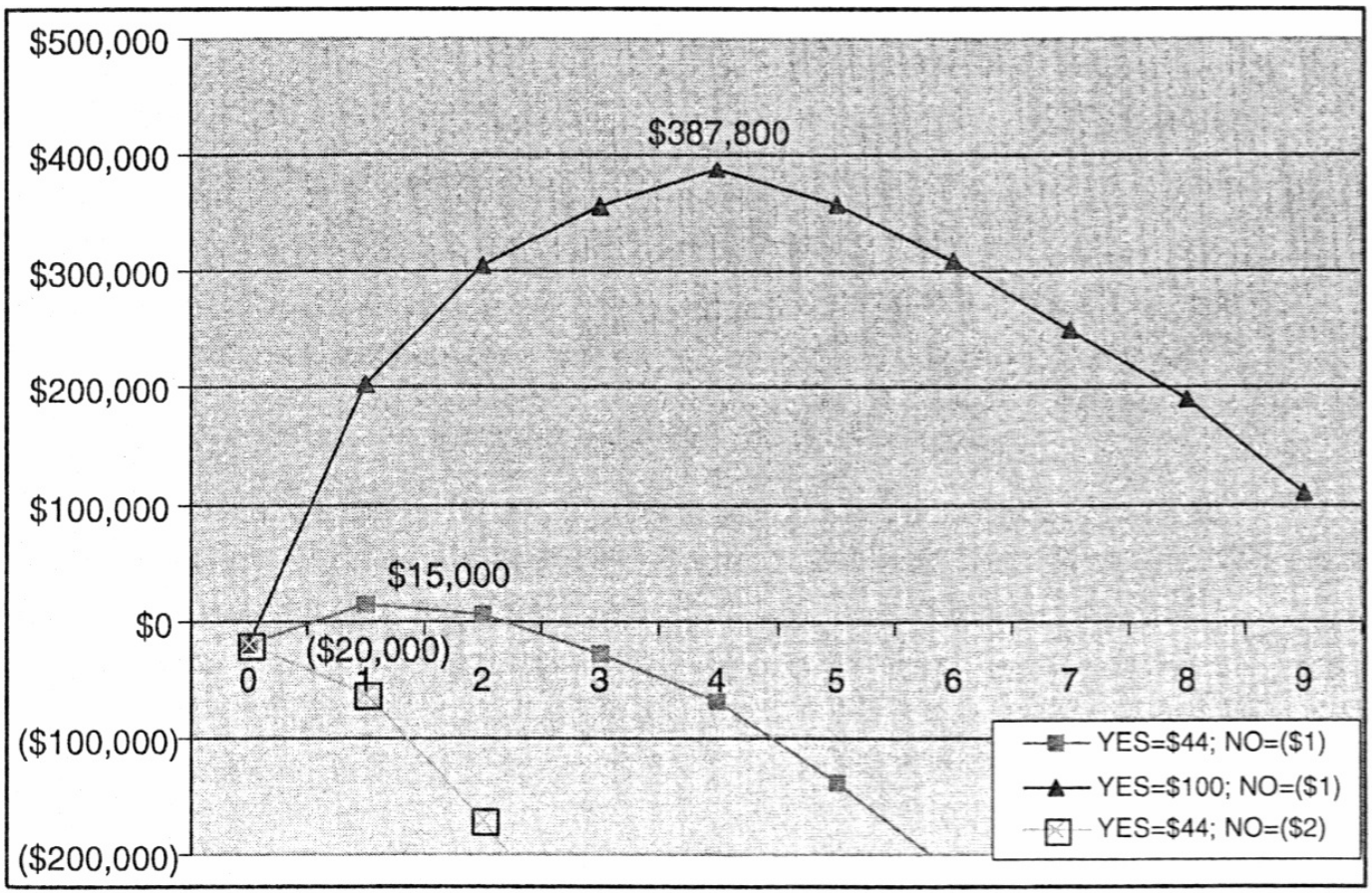
Cost of Campaign
20,000 --> 20,000
200,000 x 1 --> 100,000
Total: 220,000

Benef. Campaign
5,000 x 45 --> 225,000

Net Benef.: 5,000

Applications

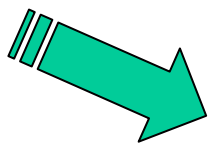
- "mail campaign design":
  - Graph showing the benefit for three different campaigns...



Extension to More than Two Classes

- The techniques presented here have been illustrated for two classes:
  - The evaluation of *multiclass* classifiers based on costs can be performed equally.
    - Example:

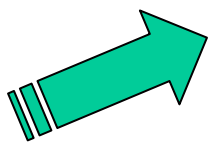
COST		*actual*		
		low	medium	high
predicted	low	0€	5€	2€
	medium	200€	-2000€	10€
	high	10€	1€	-15€



Total cost:

-29787€

ERROR		*actual*		
		low	medium	high
predicted	low	20	0	13
	medium	5	15	4
	high	4	7	60



Extension to More than Two Classes

- ROC analysis, though, is not easily extensible:
  - Given *n* classes, there is a *n* x (*n*–1) dimensional space.
    - Calculating the convex hull impractical.
    - The "context" now is not determined by a value (*slope*), but *n* x (*n*–1) –1 values.
  - There have been approximations (for three classes, Mossman 1999) or efforts to tackle the general problem (Srinivasan 1999) (Ferri et al. 2003b).
- The AUC measure has been extended.
  - All-pair extension (Hand & Till 2001).

$$AUC_{HT} = \frac{1}{c(c-1)} \sum_{i=1}^{c} \sum_{j=1, j<i}^{c} AUC(i, j)$$

  - "one vs. rest" extension(Fawcett)
  - Other extensions (Yan et al. 2003) (AUC*, Ting 2002).

Conclusions

- ROC Analysis:
  - Highlights the fact that the evaluation of classifiers goes much beyond the estimation of the error rate.
  - Makes it possible to work *with* costs and distributions, or *without* this information, improving the generation, selection and application of classifiers.
  - Provides a set of varied metrics and techniques to evaluate classifiers depending on the task: minimise error, minimise cost, improve a ranking, etc.
- It is a hot research area of wide applicability in data mining and machine learning.

Some References

- Bradley, A.P. (1997) “The use of the area under the ROC curve in the evaluation of machine learning algorithms” *Pattern Recognition*, 30(7), 1145-1159.
- Egan, J.P. (1975). *Signal Detection Theory and ROC Analysis*. Series in Cognition and Perception. Academic Press, New York.
- Fawcett, T.(2001). “Using rule sets to maximize ROC performance”.In Proceedings of the IEEE International Conference on Data Mining (ICDM-2001), pp.131-138.
- Fawcett, T., & Provost, F. (1997). “Adaptive fraud detection”. *Data Mining and Knowledge Discovery*, 1(3),291-316.
- Fawcett,T. (2003). “ROC graphs: Notes and practical considerations for data mining researchers” Tech report HPL-2003-4. HP Laboratories, PaloAlto, CA, USA. Available: http://www.purl.org/net/tfawcett/papers/HPL-2003-4.pdf.
- Ferri, C., Flach, P.; Hernández-Orallo, J. (2002). “Learning Decision Trees using the Area Under the ROC Curve”, in C. Sammut; A. Hoffman (eds.) “The 2002 International Conference on Machine Learning” (ICML2002), IOS Press, Morgan Kaufmann Publishers, pp. 139-146.
- Ferri, C.; Flach, P.A.; Hernández-Orallo, J. (2003a) "Improving the AUC of Probabilistic Estimation Trees". European Conference on Machine Learning, ECML 2003: 121-132
- Ferri, C.; Hernández-Orallo, J.; Salido, M.A. (2003b) "Volume under the ROC Surface for Multi-class Problems". European Conference on Machine Learning, ECML 2003: 108-120
- Flach, P.; Blockeel, H.; Ferri, C.; Hernández-Orallo, J.; Struyf, J. (2003) “Decision Support for Data Mining: Introduction to ROC analysis and its applications” in *Data Mining and Decision Support: Integration and Collaboration*, Kluwer Academic Publishers, Boston, 2003.
- Flach, P.A. "The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics". (2003) International Conference on Machine Learning, ICML 2003: 194-201
- Fürnkranz, J.; Flach, P.A.: An Analysis of Rule Evaluation Metrics. (2003) International Conference on Machine Learning, ICML 2003: 202-209
- Hand, D.J., & Till, R.J. (2001). “A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems”, *Machine Learning*, 45, pp. 171-186.
- Hanley, J.A. , & McNeil, B.J. (1982). “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. *Radiology*,143,29-36.

Some References

- Lachiche, N. & Flach, P.A. (2003). "Improving Accuracy and Cost of Two-class and Multi-class Probabilistic Classifiers Using ROC Curves". International Conference on Machine Learning, ICML 2003: 416-423
- Lane, T. (2000). "Extensions of ROC analysis to multi-class domains". In Dietterich, T., Margineantu, D., Provost, F., & Turney, P. (Eds.), ICML-2000 Workshop on Cost-Sensitive Learning
- Ling, C.X.; Yan, R.J. (2003) "Decision Tree with Better Ranking" The 2003 International Conference on Machine Learning (ICML2003), IOS Press, Morgan Kaufmann Publishers, to appear.
- Mann, H. B. & Whitney, D. R. (1947). "On a test whether one of two random variables is stochastically larger than the other". Ann. Math. Statist., 18, pp. 50-60.
- Mossman,D.(1999). "Three-way ROCs". *Medical Decision Making*,19,78-89.
- Provost, F., & Domingos, P. (2003). "Tree Induction for Probability-based Ranking", *Machine Learning* 52:3 (in press), 2003.
- Srinivasan, A. (1999) "Note on the Location of Optimal Classifiers in N-dimensional ROC Space" Technical Report PRG-TR-2-99, Oxford University Computing Laboratory, Oxford.
- Swets, J.A. (1988). "Measuring the accuracy of diagnostic systems". *Science*, 240,1285-1293.
- Swets, J.A., Dawes, R.M., & Monahan, J. (2000). "Better decisions through science". Scientific American, 283, 82-87. http://www.psychologicalscience.org/pdf/pspi/sciam.pdf.
- Ting, Kai Ming (2002). "Issues in Classifier Evaluation using Optimal Cost Curves" The Proceedings of the International Conference on Machine Learning" International Conference on Machine Learning, ICML 2002, pp. 642-649.
- Turney, P. (2000) "Types of Cost in Inductive Concept Learning" Proceedings Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning (WCSL at ICML-2000), 15-21.
- Weiss, G. and Provost, F. "The Effect of Class Distribution on Classifier Learning: An Empirical Study" Technical Report ML-TR-44, Department of Computer Science, Rutgers University, 2001.
- Wilcoxon, F. (1945). "Individual comparisons by ranking methods". Biometrics, 1, pp. 80-83.
- Yan, L., Dodier, R., Mozer, M. C., & Wolniewicz, R. (2003). "Optimizing classifier performance via the Wilcoxon-Mann-Whitney statistic. In The Proceedings of the International Conference on Machine Learning" International Conference on Machine Learning, ICML (pp. 848-855). http://www.cs.colorado.edu/~mozer/papers/
- Zweig, M.H.; Campbell, G. (1993) "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine", Clin. Chem, 1993; 39: 561-77.

Some Websites to Know More

- Tom Fawcett's page on ROC Analysis:
  http://www.hpl.hp.com/personal/Tom_Fawcett/ROCCH/

- ROC Analysis Software
  http://epiweb.massey.ac.nz/ROC_analysis_software.htm
  http://cs.bris.ac.uk/~farrand/rocon/

- ROC Analysis Extensive Bibliography:
  http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html

- 1st Workshop on "ROC Analysis in AI", Valencia, 22/23 August 2004 (within ECAI'2004)
  http://www.dsic.upv.es/~flip/ROCAI2004/