Fouille de données

▷ Quality assessment and validation of classifiers

Philippe Lenca
IMT Atlantiqueaa
2016-2017

Outline

Outline

Evaluation & Credidibility Issues

Evaluation is a major issue in Data Mining

KDD is a non-trivial (decision aid interactive and iterative) process where user(s) seek to identify **valid**, novel, potentially useful, and ultimately understandable **patterns in data**.

- how reliable are the predicted results?
- how much should we be confident with what was learned?
- what measure should we use?
- how should we measure performance?
- how to compare the relative performance among competing models?

↪ Metrics, framework, comparisons and tests.

Evaluation & Credidibility Issues

Basic example

Most widely-used metric: accuracy which is the proportion of true results in the data set.

Consider a 2-class problem:

- 100 examples of class C_1
- 9900 examples of class C_2
- if model M_1 predicts everything to be class C_2, accuracy is $9900/10000 = 99.0$ %
- but M_1 does not offer any value to help to predict C_1

$\hookrightarrow$ Accuracy could be high but the model unuseful.

Evaluation & Credidibility Issues

Basic example

	Predicted C_2	Predicted C_1
C_1	50	100
C_2	9,700	150

	Predicted C_2	Predicted C_1
C_1	150	0
C_2	9,850	0

$\mathtt{acc}_{M_1} = \frac{9{,}700+100}{9{,}700+50+100+150} = 98\%$

M_1 is *fair*.

$\mathtt{acc}_{M_2} = \frac{9{,}850+0}{9{,}850+150+0+0} = 98.5\%$

M_2 is trivial and better than M_1.

M_2 reduces the rate of inaccurate predictions from 2% to 1.5% (an apparent improvement of 25%!!), but it never predicts C_1.
The class variable is generally imbalanced and the most interesting class is the smaller one.

$\hookrightarrow$ The less accurate model is more useful than the more accurate model. Accuracy is not irrelevant but not enough, other metrics should be used.

Evaluation & Credidibility Issues

Is it easy to define such quantitative measures?

- validity: measures of certainty, robusteness
  - estimated prediction accuracy, confidence on new data
- utility: gain
  - in money saved because of better predictions
  - speedup in response time
- novelty, surprising: more subjective
  - if the pattern contradicts a user expectation
  - with a stochastic model
- understandability: more subjective
  - estimated by simplicity (size of the pattern)

$\hookrightarrow$ Metrics for classifiers.

Outline

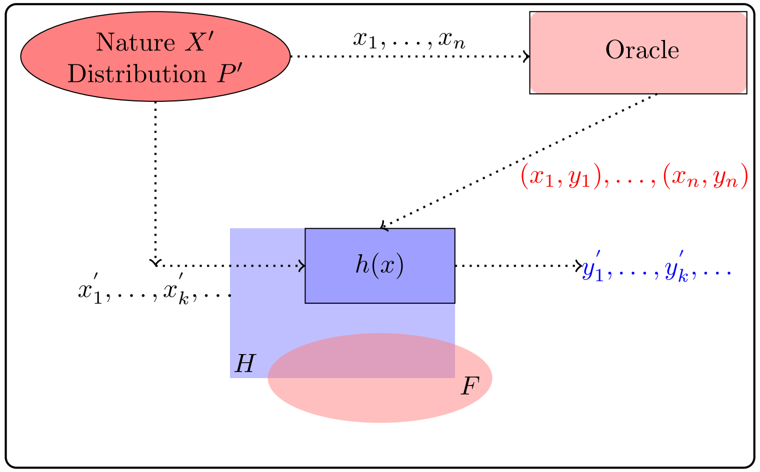
Supervised learning

Formally.

- a sample set $S = (x_1, y_1) \dots (x_n, y_n)$ coming from an (unknown) distribution P on $X \times Y$ where X is the d-dimensional space of the attributes $X_1, \dots, X_d$ and Y the target space
- goal: infer an hypothesis $h : X \to Y \in H$ (hypothesis space) such that the generalization error $Pr(x, y)_{\equiv P}[h(x) \neq y]$ is minimal (taking into account a cost function)

$\hookrightarrow$ The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered [Mit80].

Supervised learning



Supervised learning

Issues.

- but the true distribution P is unknown
- noise
- non separable labels

$\hookrightarrow$ The generalization error $Pr(x, y)_{\equiv P}[h(x) \neq y]$ of $h(x)$ can be decomposed into three terms: incompressible error, biais and variance [Bre98].

Biais & variance

Issues.

- the target function $f(x) \in F$ has the smallest generalization error
- let h^* be the best function for S and H
- ideally, $h(x) = f(x)$
- but because of non separable cases there is an incompressible error $\epsilon = Pr(x, y)_{\equiv P}[f(x) \neq y]$

$\hookrightarrow$ So $Pr(x, y)_{\equiv P}[h(x) \neq y] = \epsilon + E[h(x) - f(x)]$
Goal: min $E[h(x) - f(x)]$

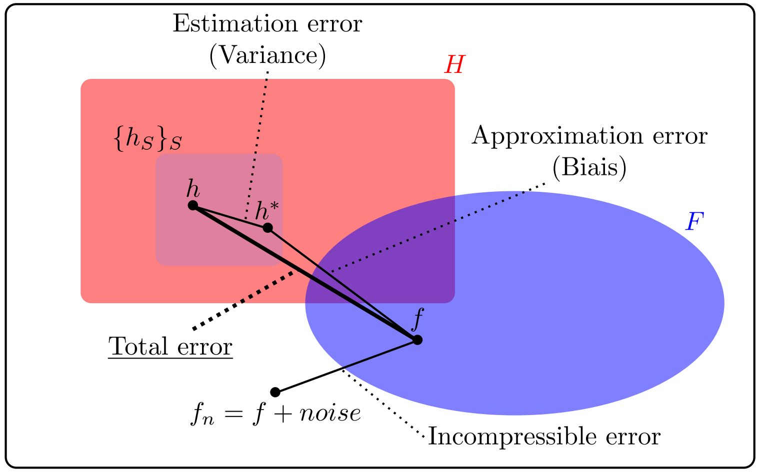
Biais & variance

$E[h(x) - f(x)]$.

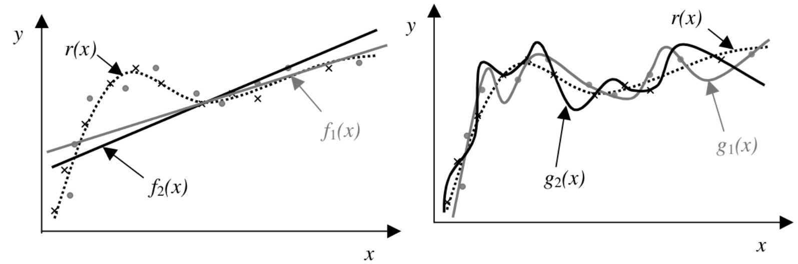
- $h^* - f$: biais (approximation error, H vs. F))
- $h - h^*$: variance (estimation error, sample dependency)
- $h - f$: total error

$\hookrightarrow$ Need for compromise.

Biais & variance



Biais & variance: illustration [Mon99]



- large biais and low variance: affine functions f have a large discrepancy with r, but this gap depends little on the learning base
- low biais and large variance: complex functions g are able to adjust as closely as possible to the observed points of r, but their forms vary greatly according to the learning base

First conclusions

Learning is an ill-posed problem:

- data is not sufficient and there is a need for inductive bias, assumptions about H
- error on learning set is not a good indicator of the quality of the model
- there is a dilemma between biais and variance
- take into account model complexity: H too much complex/overfitting, H not enough complex/underfitting

$\hookrightarrow$ Trade-off between complexity of H, size of learning set, generalization error (which first decreases and then increases with complexity –stopping criteria–).

Occam's razor: *simple solutions generalize well; when you have two competing theories that make exactly the same predictions, the simpler one is the better; among competing hypotheses, the one with the fewest assumptions should be selected; the simplest explanation for some phenomenon is more likely to be accurate than more complicated explanations*. . . Keep things simple!

- simpler to compute and to use (lower computational complexity)
- easier to train/tune (lower space complexity)
- easier to explain (more interpretable)
- better generalization ability (lower variance)

↪ Use the less complex model as possible.

Confusion matrix

Confusion matrix summarizes the performance of a classifier on a dataset: a cross-table of predicted labels (columns) and actual labels (rows).

	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	n

TP true positive
FN false negative
FP false positive
TN true negative

↪ Basis material of evaluation.

Accuracy and error rate

Proportion of true (or false) results in the data set: focus on the predictive capability of a model.

	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	**n**

accuracy: $acc = \frac{a+d}{n} = 1 - e$

error rate: $e = \frac{b+c}{n} = 1 - acc$

↪ Accuracy/error rate are not satisfying. They can be irrelevant in case of imbalanced class (which is the most common case) and where the most interesting class is the smaller one.

Basis material

Cost matrix

Errors do not have the same cost. Cost matrix summarizes the cost of misclassifying (non symmetric matrix).

	Predicted	
	C_1	C_2
C_1	$c(C_1\|C_1)$	$c(C_2\|C_1)$
C_2	$c(C_1\|C_2)$	$c(C_2\|C_2)$

$c(i,j)$ cost of misclassifying class j example as class i

$\hookrightarrow$ Basis material of evaluation. But cost matrix are not easy to obtain.

Cost-sensitive measures

Precision (pattern recognition and information retrieval)

Precision is the fraction of retrieved instances that are relevant.

	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	n

$$p = \frac{TP}{TP+FP} = \frac{a}{a+c}$$

Information retrieval

$p = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$

$p = 1.0$: every result retrieved by a search was relevant (but says nothing about whether all relevant documents were retrieved).

$\hookrightarrow$ The higher p, the lower the FPs, biased towards $c(C_1|C_1)$ and $c(C_1|C_2)$.

Cost-sensitive measures

Recall (pattern recognition and information retrieval)

Recall is the fraction of relevant instances that are retrieved.

	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	n

$$r = \frac{TP}{TP+FN} = \frac{a}{a+b} \text{ (True Positive Rate)}$$

Information retrieval

$r = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$

$r = 1.0$: all relevant documents were retrieved (but says nothing about how many irrelevant documents retrieved). It is trivial to achieve recall of 1.0 by returning all documents...

$\hookrightarrow$ The higher r, the lower the FNs, biased towards $c(C_1|C_1)$ and $c(C_2|C_1)$.

Cost-sensitive measures

F_β-measure (for non-negative real values of β)

F_β-measure combines precision and recall. It "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision" [van Rijsbergen (1979)].

$$F_\beta = (1+\beta^2) \cdot \frac{p \cdot r}{\beta^2 \cdot p + r} = \frac{(1+\beta^2) \cdot TP}{((1+\beta^2) \cdot TP + \beta^2 \cdot FN + FP)}$$

	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	n

Harmonic mean of precision and recall: $F_1 = 2\frac{p \cdot r}{p+r} = 2\frac{TP}{(2TP+FN+FP)}$

Weights recall higher than precision: $F_2 = 5\frac{p \cdot r}{4p+r}$ – Weights precision higher than recall: $F_{0.5} = 1.25\frac{p \cdot r}{0.25p+r}$

Cost-sensitive measures

F_1-measure

Harmonic mean of precision and recall:

$$F_1 = 2\frac{\mathrm{p}\cdot\mathrm{r}}{\mathrm{p}+\mathrm{r}} = 2\frac{\mathrm{TP}}{(2\mathrm{TP}+\mathrm{FN}+\mathrm{FP})}$$

F1-measure is a compromise between p and r, it is high when both precision and recall are reasonably high.

$\hookrightarrow$ The higher F_1, the lower the FPs and FNs, biased towards all except $c(C_2|C_2)$.

Cost-sensitive measures

Sensitivity/specificity, PPV and NPV (medicine)

Sensitivity is the fraction of total positives cases which are correctly predicted: ability to identify positive results: $se = \frac{TP}{TP+FN} = \frac{a}{a+b}$ $= r =$TP-rate

Specificity is the fraction of total negative cases which are correctly predicted as negatives: ability of to identify negative results. $sp = \frac{TN}{TN+FP} = \frac{d}{c+d} =$ 1- FP-rate

	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	n

FP-rate (or false alarm rate) is the proportion of negative cases predicted as positives: FP-rate $= \frac{FP}{FP+TN}$

Positive predictive value: $PPV = \frac{TP}{TP+FP} = p$

Negative predictive value: $NPV = \frac{TN}{FN+TN}$

$\hookrightarrow$ The higher r, the lower the FNs, biased towards $c(C_1|C_1)$ and $c(C_2|C_1)$.

Cost-sensitive measures

Notes

- Sensitivity = Recall = TP-rate : proportion of total positives cases which are correctly predicted

  If recall = sensitivity = 0.90, 90% from the sick group are predicted as sick
- Fp-rate (or false alarm rate) is the proportion of total negative cases which are erroneously predicted as positives
- Specificity is the proportion of of total negative cases which are correctly predicted as negatives

  If Specificity = 0.80, 80% from the healthy group are predicted as healthy, while 20% are predicted as sick
- Precision is the proportion of predicted positives cases which are truly positive

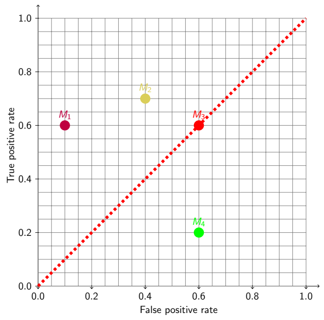
  If Precision = 0,60, 60% of the predicted positives cases are truly positive

Outline

ROC

Receiver Operating Characteristic

- signal detection (World War II, 1950s)
- ROC graph: True positive rate against False positive rate
- performance of classifiers (different algorithms, parameter settings, training scenarios, cost matrix, etc.) is represented as a point on the ROC space
- ROC graph depicts relative trade-offs between benefits (TP) and costs (FP)



↪ First application of ROC in machine learning in [Spa89].

ROC space & Aera Under ROC Curve

ROC curve

ROC curves are insensitive to class imbalance.

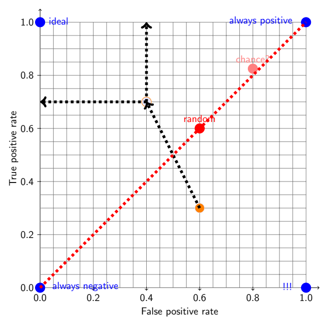
- to choose the best trade-off between tp-rate and fp-rate (idem for precision-recall curve)
- to compare the performance of two classifiers

AUC

- the greater the AUC, the better the performance

↪ Readings: [Bra97, Faw06, Fla10].

ROC space



	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	n

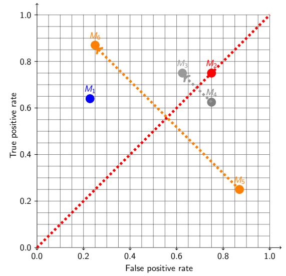
$TPR = \frac{TP}{TP+FN} = \frac{a}{a+b}$ $FPP = \frac{FP}{FP+TN} = \frac{c}{c+d}$

(0, 0): no false positive errors but also gains no true positives
(1, 1): always decide positive
(x, x): decide randomly
(0, 1): perfect classifier

↪ Go northwest (TPR is higher, FPR is lower, or both).

ROC

	Predicted		
	C_1	C_2	Total
C_1	a (TP)	b (FN)	$a+b$
C_2	c (FP)	d (TN)	$c+d$
Total	$a+c$	$b+d$	n



	M_1 **Predicted**		
	C_1	C_2	Total
C_1	640	360	1000
C_2	280	720	1000
Total	920	1080	2000

$TPR = 0.64$ $FPP = 0.28$

$Acc = 0.68$

	M_2 **Predicted**		
	C_1	C_2	Total
C_1	750	250	1000
C_2	750	250	1000
Total	1500	500	2000

$TPR = 0.75$ $FPR = 0.75$

$Acc = 0.50$

	M_3 **Predicted**		
	C_1	C_2	Total
C_1	300	100	400
C_2	1000	600	1600
Total	1300	700	2000

$TPR = 0.75$ $FPR = 0.62$

$Acc = 0.45$

	M_4 **Predicted**		
	C_1	C_2	Total
C_1	1000	600	1600
C_2	300	100	400
Total	1300	700	2000

$TPR = 0.62$ $FPR = 0.75$

$Acc = 0.55$

	M_5 **Predicted**		
	C_1	C_2	Total
C_1	250	750	1000
C_2	870	130	1000
Total	1120	880	2000

$TPR = 0.25$ $FPR = 0.87$

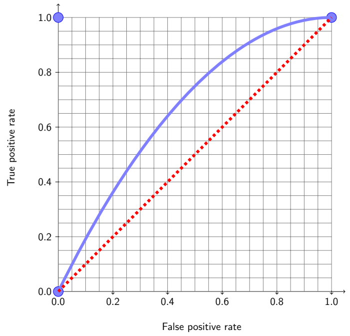
$Acc = 0.19$

	M_6 **Predicted**		
	C_1	C_2	Total
C_1	870	130	1000
C_2	250	750	1000
Total	1180	880	2000

$TPR = 0.87$ $FPR = 0.25$

$Acc = 0.81$

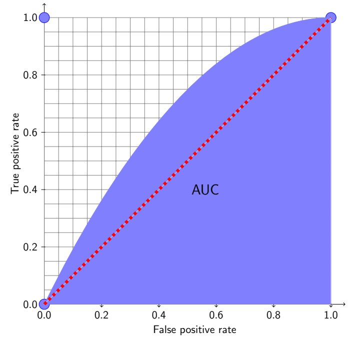
AUC



Full ROC curve [Bra97])

To generate a full ROC curve from a classifier instead of just a single point, it is necessary to generate scores from the considered classifier rather than just a class label.
A precision-recall curve can also be generated.

AUC

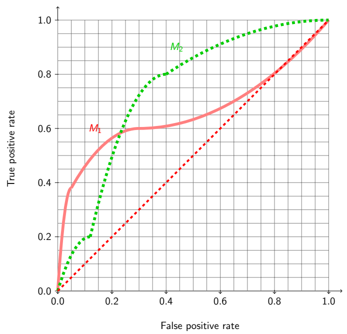


Area Under Curve (area under ROC Curve [Bra97])

The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

↪ The greater the AUC, the better the performance.

ROC

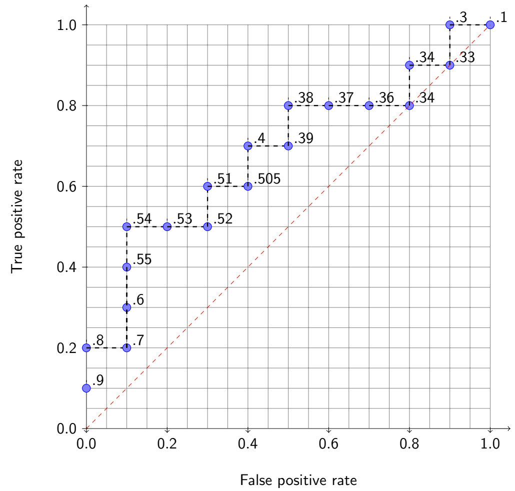


ROC curve for Model Comparison

- no model consistently outperform the other
- M_1 is better for small FPR
- M_2 is better for large FPR

ROC example [Faw06]

class	score
p	.9
p	.8
n	.7
p	.6
p	.55
p	.54
n	.53
n	.52
p	.51
n	.505
p	.4
n	.39
p	.38
n	.37
n	.36
n	.35
p	.34
n	.33
p	.3
n	.1



Outline

References I

[Bra97] Andrew P. Bradley.
The use of the area under the ROC curve in the evaluation of machine learning algorithms.
Pattern Recognition, 30(7):1145–1159, 1997.

[Bre98] Leo Breiman.
Arcing classiifiers.
The Annals of Statistics, 26(3):801–849, 1998.

[Faw06] Tom Fawcett.
An introduction to ROC analysis.
Pattern Recognition Letters, 27(8):861–874, 2006.

[Fla10] Peter A. Flach.
ROC analysis.
In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 869–875. Springer, 2010.

[Mit80] T.M. Mitchell.
The need for biases in learning generalizations.
Technical Report CBM-TR-117, Rutgers Computer Science Department, 1980.

References II

[Mon99] Gaétan Monari.
Sélection de modèles non linéaires par leave-one-out. Etude théorique et application des réseaux de neurones au procédé de soudage par points.
PhD thesis, Université Paris 6, 1999.

[Spa89] Kent A. Spackman.
Signal detection theory: valuable tools for evaluating inductive learning.
In *Proceedings of the sixth international workshop on Machine learning*, pages 160–163, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.