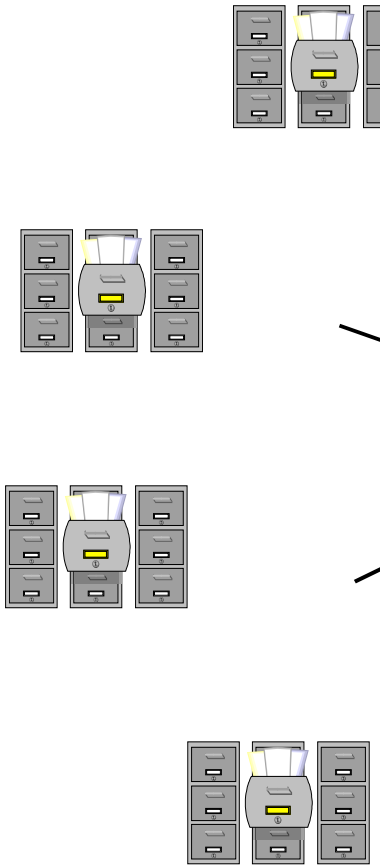
Introduction to Supervised Learning

Ricco Rakotomalala
Ricco.Rakotomalala@univ-lyon2.fr

Dataset



Variables, attributes, …

Success	Wages	Job	Refunding
Y	0	Unemployed	Slow
N	2000	Skilled Worker	Slow
N	1400	Worker	Slow
N	1573	Retired	Slow
Y	2776	Skilled Worker	Slow
N	2439	Retired	Fast
N	862	Office employee	Slow
Y	1400	Salesman	Slow
N	1700	Skilled Worker	Slow
Y	785	Employee	Fast
Y	1274	Worker	Slow
N	960	Employee	Fast
N	1656	Worker	Fast
N	0	Unemployed	Slow

Instances, cases, individuals,

Target and input variables

Success	Wages	Job	Refunding
Y	0	Unemployed	Slow
N	2000	Skilled Worker	Slow
N	1400	Worker	Slow
N	1573	Retired	Slow
Y	2776	Skilled Worker	Slow
N	2439	Retired	Fast
N	862	Office employee	Slow
Y	1400	Salesman	Slow
N	1700	Skilled Worker	Slow
Y	785	Employee	Fast
Y	1274	Worker	Slow
N	960	Employee	Fast
N	1656	Worker	Fast
N	0	Unemployed	Slow

Target variable
Class attribute
Endogenous variable

Categorical variable (qualitative)

Input variables
Predictive variables
Descriptors
Exogenous variables

Any type (categorical, quantitative)

Definition of the classification problem

Population Ω

$$\begin{cases} Y & \text{target variable} \\ X & \text{input variables} \end{cases}$$

A set of variables
$X=(x_1|\ldots|x_p)$

We want to construct a classification function [model, classifier] **f(.)** such as : $Y = f(X, \alpha)$

Goal of the induction algorithm (basic version)

Using a sample (training set) Ωa (portion of the population chosen randomly) to select the function f(.) and to calculate its parameters **α** such as to minimize the probability of error (ET)

$$ET = \frac{1}{card(\Omega)} \sum_{\Omega} \Delta[Y, \hat{f}(X, \hat{\alpha})]$$

$$où\ \Delta[.] = \begin{cases} 1 si\ Y \neq \hat{f}(X, \hat{\alpha}) \\ 0 si\ Y = \hat{f}(X, \hat{\alpha}) \end{cases}$$

TO DO :
- ☞ choosing the function f(.)
- ☞ estimating its parameters α
- ☞ we use a sample, but we want to obtain an efficient classifier in the population

Bayes rule

(Special case : binary problem – Positive class value vs. negative)

Learning process from a dataset:

- estimating the posterior probabilities P(Y / X)
- assigning [Y = +] if P(Y = + / X) > P(Y = - / X)

Note:

- P(Y = + / X) is the "positive posterior probability"
- This approach minimizes the expected probability of error

Bayes rule

(Y has more than 2 categories, known as a multiclass classification problem)

2 steps :

- estimating the posterior probabilities $P(Y = y_k / X)$
- assigning $y_{k*} = \arg \max_k P(Y = y_k / X)$

Note: When all the descriptors X are categorical, we have logical classification rules.

If X1 = ? and X2 = ? and X3 = ? … then Y = ?

premise conclusion

Bayes rule – An example on a small dataset

Direct estimation of P(Y/X) by counting the frequencies

Y X

Maladie	Poids	Taille	Marié	Etud.Sup
Présent	45	Trapu	Non	Oui
Présent	57	Elancé	Non	Oui
Absent	59	Elancé	Non	Non
Absent	61	Trapu	Oui	Oui
Présent	65	Elancé	Non	Oui
Absent	68	Elancé	Non	Non
Absent	70	Trapu	Oui	Non
Présent	72	Trapu	Non	Oui
Absent	78	Trapu	Oui	Non
Présent	80	Elancé	Oui	Non

- IF taille = ? THEN Maladie = ?
- IF taille = ? AND etud.sup = ? THEN Maladie = ?

Advantages and drawbacks of the direct approach

\+ Very simple but efficient

\- No direct solution for continuous descriptors

(discretization or assumptions about conditional distributions)

\- No checking about the relevance of the descriptors

(no variable selection process)

\- When the number of descriptors increases

- Calculations are infeasible

E.g. 100 binary descriptors ➔ 2^100 rules

- Data fragmentation problem

Rules with very low support (≈ 0)

This approach is not usable in practice

Evaluation of classifiers

Understandability

The model (classifier) expresses a "knowledge" about a domain

- Explanation : understanding the causality
- Validation : domain expert can evaluate the relevance of the model
- Improvement : domain expert can improve the model (e.g. by choosing more relevant discretization cut point, etc.)

Quickness

- During the learning step : this enables to test more solutions, to process large databases (scalability), etc.
- In the generalization step
- Ability to updating an existing model with a set of labeled instances

! Accuracy

- Checking the prediction performance of the model in the deployment phase

Confusion matrix

Comparing the observed labels and the predicted labels (binary problem)

		Predicted		
		+	-	Total
Observed	+	a	b	a+b
	-	c	d	c+d
	Total	a+c	b+d	n

Definitions :

- True positive TP = a
- False positive FP = c
- Error rate = (c+b)/n
- Sensitivity = Recall = True positive rate (TPR) = a/(a+b)
- Precision = a/(a+c)
- False Positive Rate (FPR) = c/(c+d)
- Specificity = d/(c+d) = 1 – FPR

Misclassification cost matrix

Comparison of two classifiers

		Prédite		
		+	-	Total
Observée	+	40	10	50
	-	20	30	50
	Total	60	40	100

		Prédite		
		+	-	Total
Observée	+	20	30	50
	-	0	50	50
	Total	20	80	100

☞ Calculate and compare the performance measurements

An additional information

The misclassification costs

		Prédite	
		+	-
Observée	+	0	5
	-	1	0

☞ New measure : average cost (error rate is a specific case of this measure)

Hold-out scheme for the evaluation of classifiers

Problem : We must not use the same sample to estimate the parameters of the classifier and to evaluate its generalization performance. We have biased (optimistic) estimation of the performance measurements in this case.

Success	Wages	Job	Refunding
Y	0	Unemployed	Slow
N	2000	Skilled Worker	Slow
N	1400	Worker	Slow
N	1573	Retired	Slow
Y	2776	Skilled Worker	Slow
N	2439	Retired	Fast
N	862	Office employee	Slow
Y	1400	Salesman	Slow
N	1700	Skilled Worker	Slow
Y	785	Employee	Fast
Y	1274	Worker	Slow
N	960	Employee	Fast
N	1656	Worker	Fast
N	0	Unemployed	Slow

Random partition

Training (learning) set
Used for the learning of the model
~ 70%

Test set
Used for the evaluation of the model
~ 30%
☞ Recall, precision, error rate…

(exercise : LOAN data file – Success vs. Housing & Refunding)…

References

« Machine learning », T. Mitchell, McGraw Hill, 1997.
http://www.cs.cmu.edu/~tom/mlbook.html

« The elements of statistical learning - Data Mining, Inference and Prediction », T. Hastie, R. Tibshirani, J. Friedman, Springer 2009.
http://statweb.stanford.edu/~tibs/ElemStatLearn/