INTRODUCTION

Pedro Larrañaga, Concha Bielza

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

Machine Learning

Master in Data Science + Master HMDA

Outline

Outline

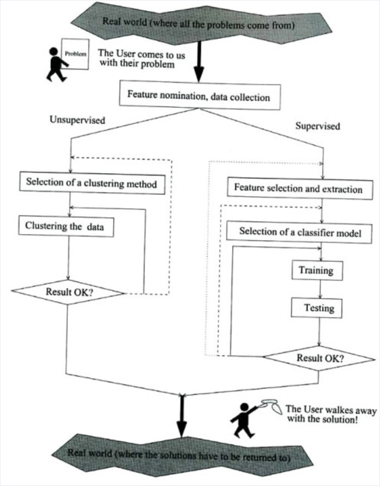
Basic ideas

Supervised vs Unsupervised

- Supervised classification
  - Given N instances (objects, examples, ...) characterized by some predictor variables (attributes) and one label (class) variable
  - The objective is to transform these data into a classification model able to predict with high accuracy the class of a new instance only characterized by the predictor variables
- Unsupervised classification
  - Given N instances (objects, examples, ...) characterized by some predictor variables (attributes)
  - The objective is to obtain some groups (clusters) of instances with a high variability between the clusters and low variability within a given cluster

Pattern recognition

The cycle of a pattern recognition system (Kuncheva, 2004)



Outline

From labelled data sets to classification models

A data file containing instances characterized by n predictor variables and one class variable

	X_1	$\dots$	X_n	C
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\dots$	$x_n^{(1)}$	$c^{(1)}$
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\dots$	$x_n^{(2)}$	$c^{(2)}$
$\dots$		$\dots$		$\dots$
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\dots$	$x_n^{(N)}$	$c^{(N)}$

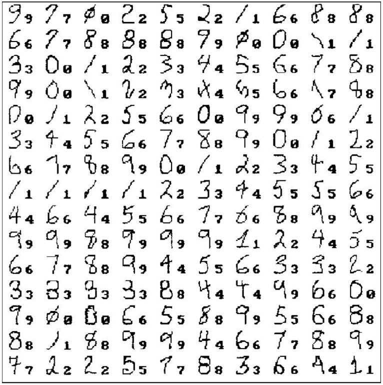
Machine learning algorithms to automatically transforming labelled data sets into classification models

Supervised classification

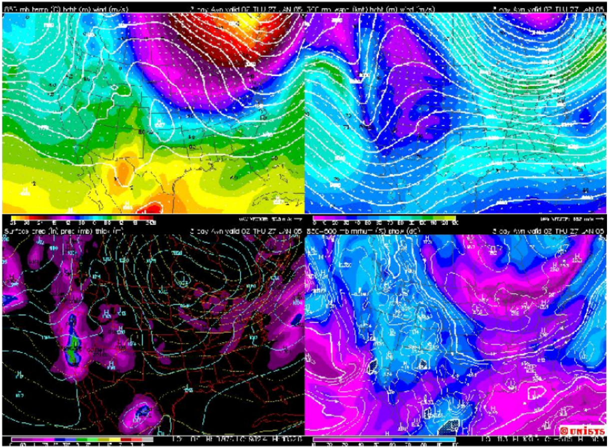
Application domains

- Decision support systems for diagnosis and prognosis
- Loan decision
- Spam detection
- Prediction of sport results
- Handwriting character recognition
- Weather forecast
- Prediction of the secondary structure of proteins
- ...

Optical character recognition



Weather forecast



Prediction of the secondary structure of proteins

Paradigms for supervised classification

Non-probabilistic and probabilistic classifiers

- Non-probabilistic classifiers
  - k–NN classifiers (Fix and Hodges, 1951)
  - Rule induction (Michalski, 1969)
  - Artificial neural networks (McCulloch and Pitts, 1943)
  - Support vector machines (Vapnik, 1988)
  - Classification trees (Quinlan, 1986; Breiman et al. 1984)
- Probabilistic classifiers
  - Discriminant analysis (Fisher, 1936)
  - Logistic regression (Hosmer and Lemeshow, 2000)
  - Bayesian classifiers (Friedman et al., 1997; Bielza and Larrañaga, 2004a)

Metaclassifiers

Metaclassifiers (Kuncheva, 2004)

Feature subset selection

Feature subset selection (Lewis, 1962)

Multi-dimensional supervised classification (Zhang and Zhou 2014)

	X_1	$\ldots$	X_m	C_1	$\ldots$	C_d
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_m^{(1)}$	$c_1^{(1)}$	$\ldots$	$c_d^{(1)}$
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_m^{(2)}$	$c_1^{(2)}$	$\ldots$	$c_d^{(2)}$
$\ldots$		$\ldots$			$\ldots$	
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_m^{(N)}$	$c_1^{(N)}$	$\ldots$	$c_d^{(N)}$
$\boldsymbol{x}^{(N+1)}$	$x_1^{(N+1)}$	$\ldots$	$x_m^{(N+1)}$	???	$\ldots$	???

- Multi-label classification
- Multiple faut diagnosis

Semisupervised classification

Labelled and unlabelled instances

	X_1	$\ldots$	X_n	C
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	1
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	0
$(\boldsymbol{x}^{(3)}, c^{(3)})$	$x_1^{(3)}$	$\ldots$	$x_n^{(3)}$	?
$(\boldsymbol{x}^{(4)}, c^{(4)})$	$x_1^{(4)}$	$\ldots$	$x_n^{(4)}$	?
$(\boldsymbol{x}^{(5)}, c^{(5)})$	$x_1^{(5)}$	$\ldots$	$x_n^{(5)}$	1
$(\boldsymbol{x}^{(6)}, c^{(6)})$	$x_1^{(6)}$	$\ldots$	$x_n^{(6)}$	?
$\ldots$		$\ldots$		$\ldots$
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	?

Partially supervised classification

Positive and unlabelled instances

- Discovering genes (instances) associated with a given disease (class variable)
- We know that some of the genes are associated with the disease (positive instances)
- For the rest of the genes it is not possible to say that they are not associated with the disease (we don't have negative instances)

Partially supervised classification

Positive and unlabelled instances

	X_1	$\ldots$	X_n	C
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\ldots$	$x_n^{(1)}$	1
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\ldots$	$x_n^{(2)}$	1
$(\boldsymbol{x}^{(3)}, c^{(3)})$	$x_1^{(3)}$	$\ldots$	$x_n^{(3)}$	1
$(\boldsymbol{x}^{(4)}, c^{(4)})$	$x_1^{(4)}$	$\ldots$	$x_n^{(4)}$	?
$(\boldsymbol{x}^{(5)}, c^{(5)})$	$x_1^{(5)}$	$\ldots$	$x_n^{(5)}$	?
$(\boldsymbol{x}^{(6)}, c^{(6)})$	$x_1^{(6)}$	$\ldots$	$x_n^{(6)}$	?
$\ldots$		$\ldots$		$\ldots$
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\ldots$	$x_n^{(N)}$	?

Statistics and Machine Learning

The two cultures (Breiman, 2001)

Statistics	Machine learning
Stochastic data model	Algorithm modeling
Model selection	Structure and parameter learning
Fitting	Learning
Likelihood ratio	Predictive accuracy
Forward, backward, stepwise	Metaheuristic
Collinearity	Feature subset selection
Bayesian approaches	Ensembles
Probabilistic output	Deterministic output

Statistics and Machine Learning

Statistics and machine learning methods in this course

Statistics	Machine learning
Feature selection (filter)	Feature selection (wrapper)
k-nearest neighbors	k-nearest neighbors
Classification trees	Classification trees
Logistic regression	Rule induction
Bayesian network classifiers	Artificial neural networks
Multi-dimensional classification	Support vector machines
Hierarchical clustering	Metaclassifiers
Partitional clustering	Multi-label classification
Probabilistic clustering	

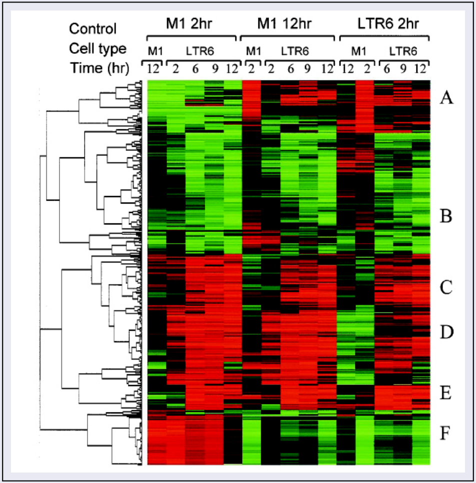
Outline

Clustering of a data file

	X_1	$\ldots$	X_i	$\ldots$	X_n
O_1	x_1^1	$\ldots$	x_i^1	$\ldots$	x_n^1
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
O_j	x_1^j	$\ldots$	x_i^j	$\ldots$	x_n^j
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
O_N	x_1^N	$\ldots$	x_i^N	$\ldots$	x_n^N

Clustering of microarray data

Hierarchical clustering of microarray data



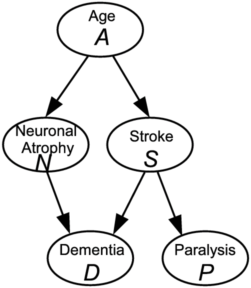
Paradigms for unsupervised classification

Clustering methods

- Non-probabilistic clustering
  - Hierarchical clustering (Sorensen, 1948)
  - Partitional clustering (MacQueen, 1967)
- Probabilistic classifiers
  - Gaussian mixture models (McLachlan and Krishnan, 1997)

Outline

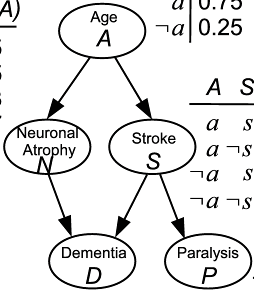
"Risk of dementia" (Bielza and Larrañaga, 2014b)



“Risk of dementia”

A	$p(A)$
a	0.75
$\neg a$	0.25

A	N	$p(N\|A)$
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97



A	S	$p(S\|A)$
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	$p(D\|N,S)$
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

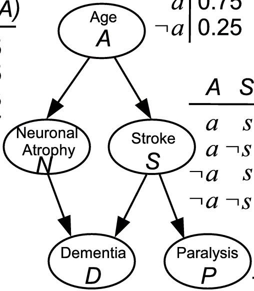
S	P	$p(P\|S)$
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

$$p(A, N, S, D, P) = p(A)p(N|A)p(S|A)p(D|N, S)p(P|S)$$

"Risk of dementia"

A	$p(A)$
a	0.75
$\neg a$	0.25

A	N	$p(N\|A)$
a	n	0.15
a	$\neg n$	0.85
$\neg a$	n	0.03
$\neg a$	$\neg n$	0.97



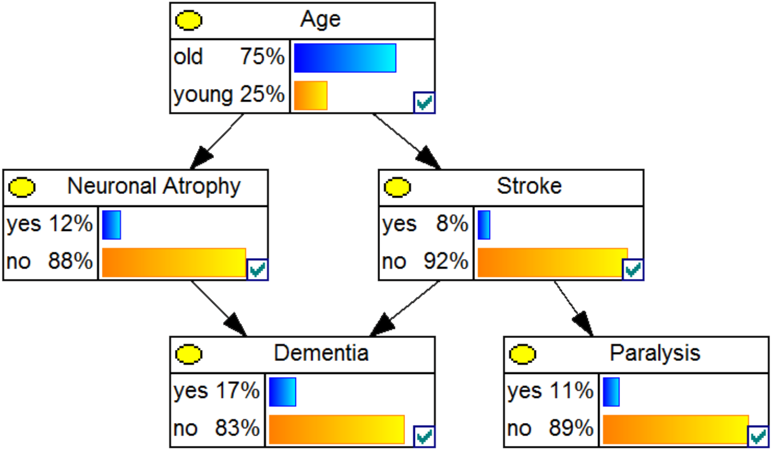
A	S	$p(S\|A)$
a	s	0.10
a	$\neg s$	0.90
$\neg a$	s	0.02
$\neg a$	$\neg s$	0.98

N	S	D	$p(D\|N,S)$
n	s	d	0.96
n	s	$\neg d$	0.04
n	$\neg s$	d	0.40
n	$\neg s$	$\neg d$	0.60
$\neg n$	s	d	0.45
$\neg n$	s	$\neg d$	0.55
$\neg n$	$\neg s$	d	0.10
$\neg n$	$\neg s$	$\neg d$	0.90

S	P	$p(P\|S)$
s	p	0.75
s	$\neg p$	0.25
$\neg s$	p	0.05
$\neg s$	$\neg p$	0.95

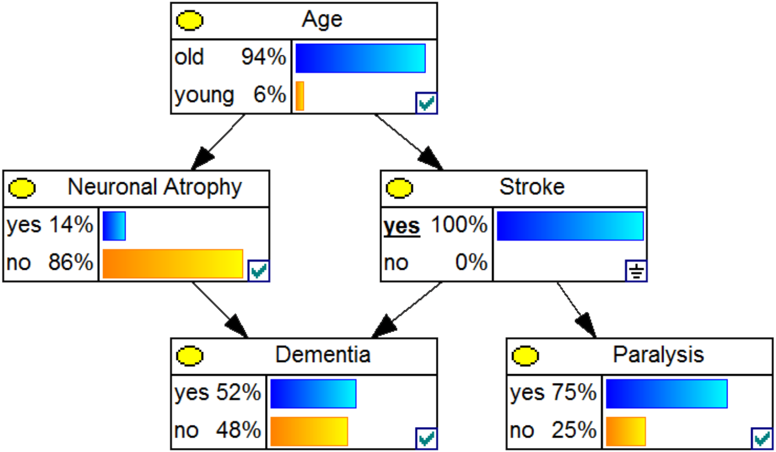
$$p(A, N, S, D, P) = p(A)p(N|A)p(S|A)p(D|N, S)p(P|S)$$

Inference (reasoning) with Bayesian networks



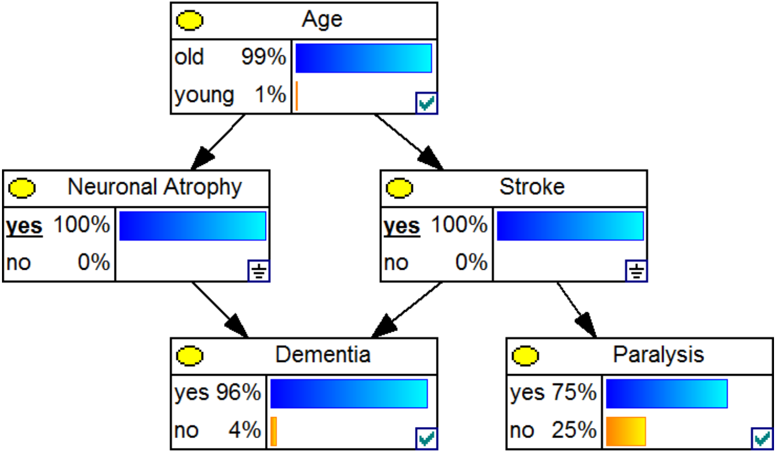
No evidence

Inference (reasoning) with Bayesian networks



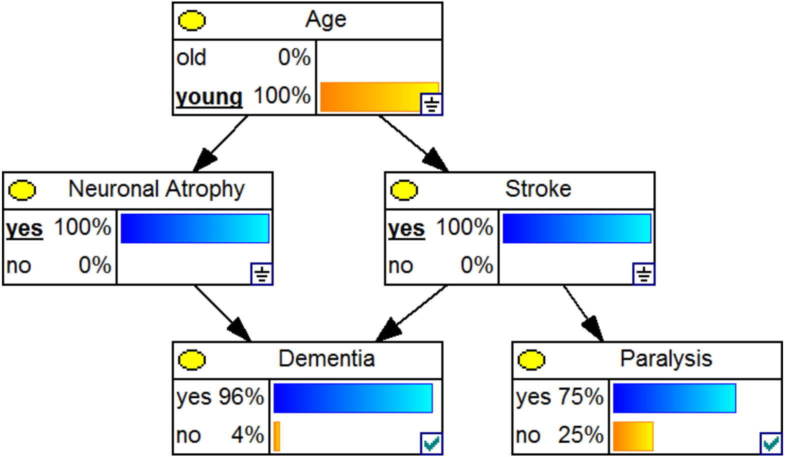
Evidence: "Stroke = yes"

Inference (reasoning) with Bayesian networks



Evidence: "'Stroke = yes, Neuronal Atrophy=yes"

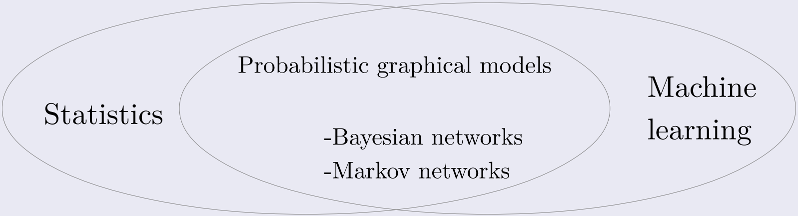
Inference (reasoning) with Bayesian networks



Evidence: "'Stroke = yes, Neuronal Atrophy=yes, Age= young"

Bayesian networks and Markov networks

PGMs and the two cultures



References

- C. Bielza, P. Larrañaga (2014a). Discrete Bayesian network classifiers: A survey. *ACM Computing Surveys*, 47(1), Article 5
- C. Bielza, P. Larrañaga (2014b). Bayesian networks in neuroscience: A survey. *Frontiers in Computational Neuroscience*, 8, Article 131
- L. Breiman (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199-231
- L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone (1984). *Classification and Regression Trees*. Wadsworth Press
- R. Fisher (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179-188
- E. Fix, J.L. Hodges (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation. Technical Report 4
- N. Friedman, D. Geiger, M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning*, 25, 131-163
- D.W. Hosmer, S. Lemeshow (2000). *Applied Logistic Regression*. Wiley Interscience
- L.I. Kuncheva (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience
- P.M. Lewis (1962). The characteristic selection problem in recognition systems. *IRE Transactions on Information Theory*, 8:171-178
- J. MacQueen (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297
- W. McCulloch and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115-133
- G.J. McLachlan, T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley
- R.S. Michalski (1969). On the quasi-minimal solution of the general covering problem. *Proceedings of the 5th International Symposium on Information Processing*, 125-128
- J.R. Quinlan (1986). Induction of decision trees. Machine Learning, 1(1), 81-106
- T. Sorensen (1948). A method for establishing groups of equal amplititude in plant sociology based on similarity of species contents and its application to analyze of the vegetation on Danish commons. *Biologiske Skrifter*, 5, 1-34
- V.N. Vapnik (1988). *Statistical Learning Theory*. Wiley
- Zhang, M.-L., Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8), 1819-1837