

MACHINE LEARNING IN AVIATION

PEDRO LARRAÑAGA

Computational Intelligence Group
Artificial Intelligence Department
Technical University of Madrid, Spain



Toulouse, July 11, 2013

Outline

1 Introduction

2 Supervised Classification

3 Clustering

4 Conclusions

Outline

1 Introduction

2 Supervised Classification

3 Clustering

4 Conclusions

Introduction

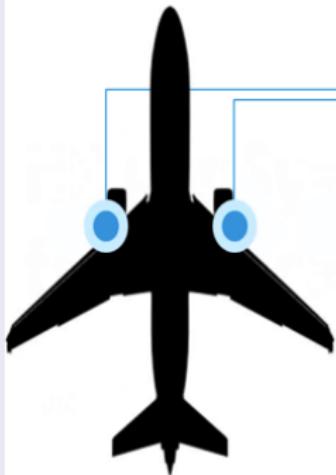
Data in aviation

- Data in our world: smartphones, social networks, atmospheric readings, financial data, medical records, bioinformatics, airplane instrumentation, etc.
- Estimation of 2.5 exabytes of data generated on the planet every day
- Data in aviation:
 - Flight tracking
 - Weather conditions
 - Airport information
 - Airline information
 - Market information
 - Passenger information
 - Air safety reports
 - Aircraft data

Introduction

Data in USA cross-country commercial flights

Sensor data from a cross-country flight



$$20 \text{ TB} \times 2 \times 6 \times 28,537 \times 365$$

20 terabytes of
information per
engine every hour

twin-engine
Boeing 737

six-hour, cross-
country flight from
New York to Los
Angeles

of commercial
flights in the sky in
the United States on
any given day.

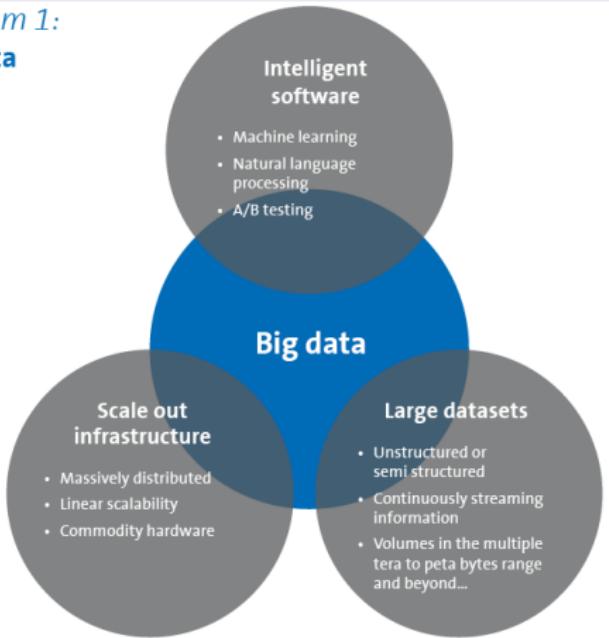
days in a year

$$= 2,499,841,200 \text{ TB}$$

Introduction

Big Data. The five V's: volume, variety, velocity, viability and value

Diagram 1:
Big data



DAVENPORT, TH (2013). *At the Big Data Crossroads: Turning Towards a Smarter Travel Experience*. Amadeus IT Group

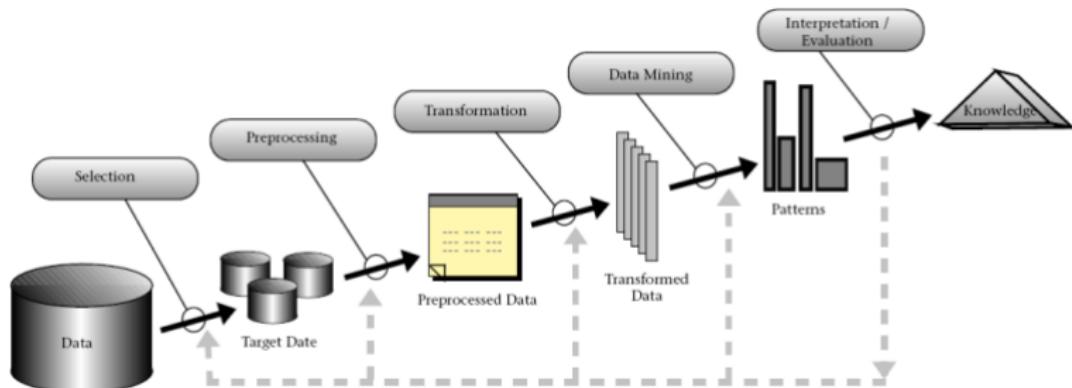
Introduction

Aviation industry

- **Characteristics:** Large scale and unstructured mixture of data in a variety of data formats
 - Radar
 - Weather (spatial-temporal)
 - Terrain (spatial)
 - Infrastructure
 - Text
- **Benefits:**
 - Help airlines increase sales
 - Customer loyalty
 - Improve fuel management and efficiency
 - Manage fleets more effectively
 - Improve customer service

Pattern recognition (PR) a step in knowledge discovery in databases (KDD)

PR a part of KDD



Outline

1 Introduction

2 Supervised Classification

3 Clustering

4 Conclusions

Supervised classification

Supervised: From labelled data to classification models

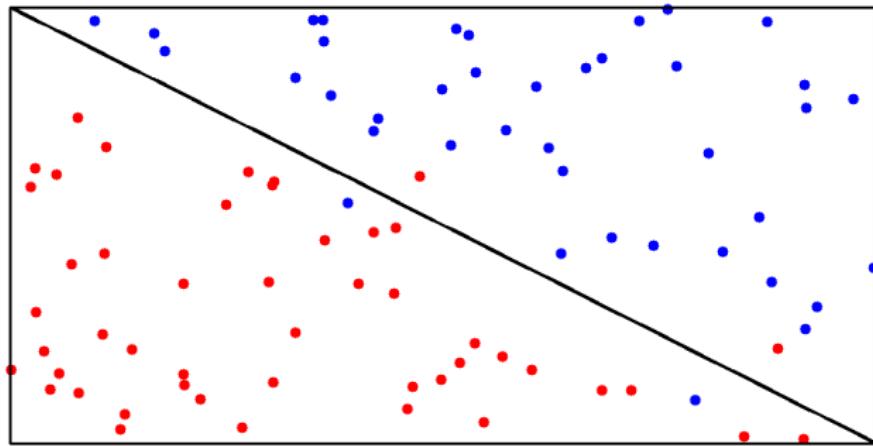
Predictor variables (attributes) and one **labelled** (class) variable:

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

Construct a **classification model** that predicts with high accuracy the class of a new instance **only characterized** by the predictor variables

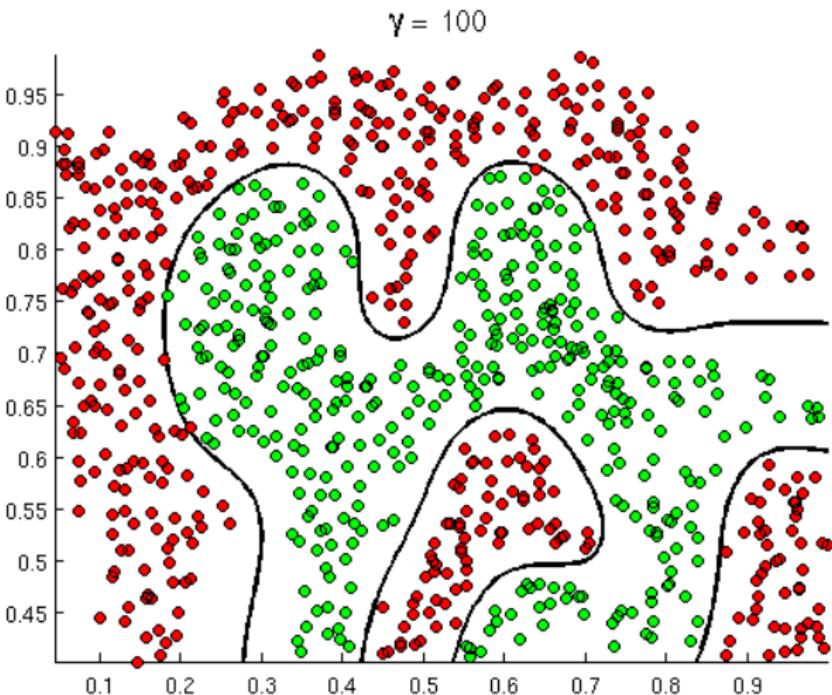
Supervised classification

Linear decision boundary



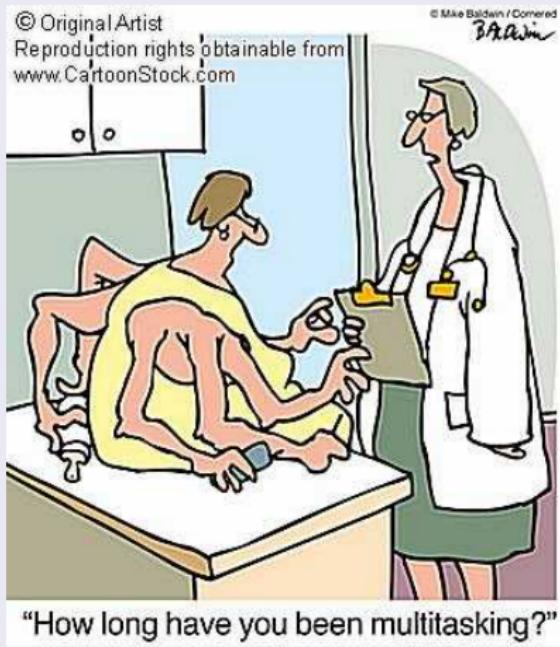
Supervised classification

Non linear decision boundary



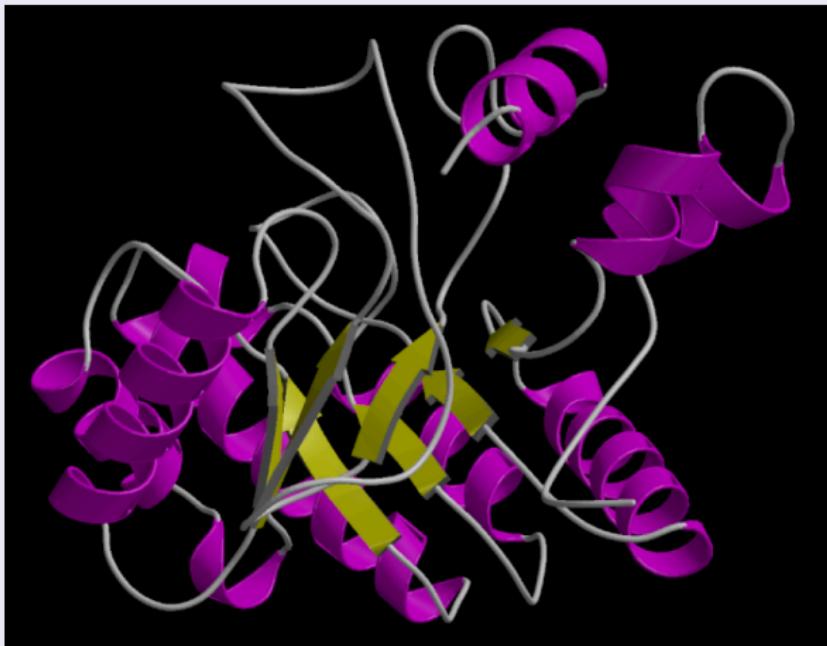
Supervised classification

Disease diagnosis



Supervised classification

Prediction of the secondary structure of proteins



Supervised classification

Fraudulent use of credit cards



Supervised classification

Spam email



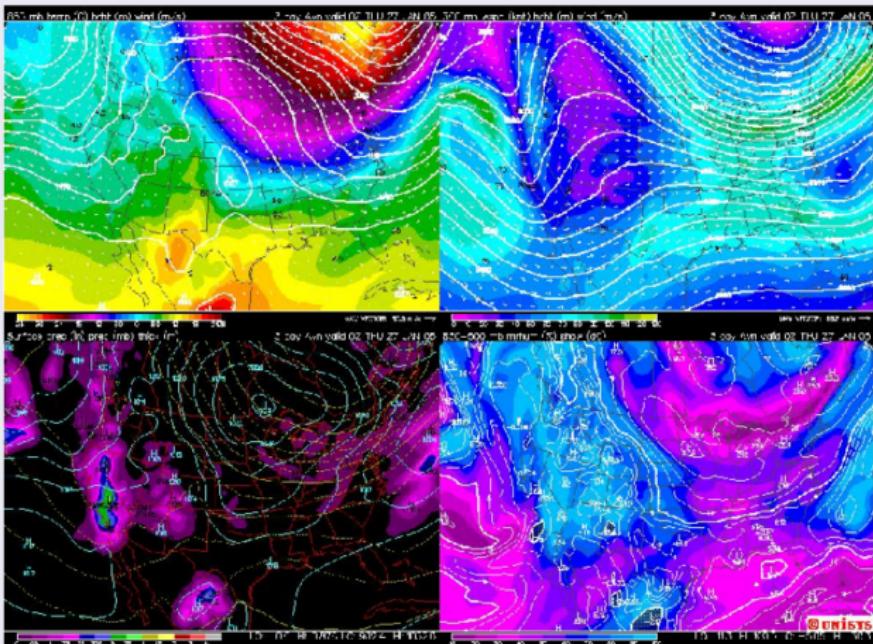
Supervised classification

Handwritten character recognition

9, 7, 0, 2, 5, 2, 1, 6, 8, 8	6, 7, 8, 8, 8, 9, 0, 0, 1, 1
3, 0, 1, 2, 3, 4, 5, 6, 7, 8	3, 0, 1, 2, 3, 4, 5, 6, 7, 8
9, 0, 1, 2, 3, 4, 5, 6, 7, 8	9, 0, 1, 2, 3, 4, 5, 6, 7, 8
0, 1, 2, 5, 6, 0, 9, 9, 0, 1	0, 1, 2, 5, 6, 0, 9, 9, 0, 1
3, 4, 5, 6, 7, 8, 9, 0, 1, 2	3, 4, 5, 6, 7, 8, 9, 0, 1, 2
6, 7, 8, 9, 0, 1, 2, 3, 4, 5	6, 7, 8, 9, 0, 1, 2, 3, 4, 5
1, 1, 1, 1, 2, 3, 4, 5, 5, 6	1, 1, 1, 1, 2, 3, 4, 5, 5, 6
4, 6, 4, 5, 6, 7, 6, 8, 9, 9	4, 6, 4, 5, 6, 7, 6, 8, 9, 9
9, 9, 8, 9, 9, 9, 1, 2, 4, 5	9, 9, 8, 9, 9, 9, 1, 2, 4, 5
6, 7, 8, 9, 4, 5, 6, 3, 3, 2	6, 7, 8, 9, 4, 5, 6, 3, 3, 2
3, 3, 3, 3, 8, 4, 4, 4, 6, 0	3, 3, 3, 3, 8, 4, 4, 4, 6, 0
9, 0, 0, 6, 5, 8, 9, 5, 6, 8	9, 0, 0, 6, 5, 8, 9, 5, 6, 8
8, 1, 8, 9, 9, 4, 6, 7, 8, 9,	8, 1, 8, 9, 9, 4, 6, 7, 8, 9,
7, 2, 2, 5, 7, 8, 3, 6, 4, 1	7, 2, 2, 5, 7, 8, 3, 6, 4, 1

Supervised classification

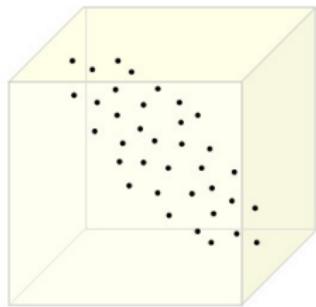
Weather forecasting



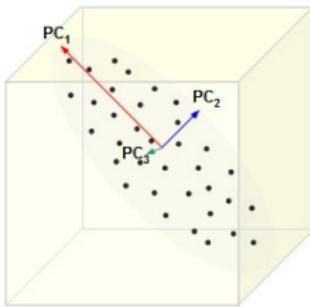
Supervised classification

Less is more. Dimensionality reduction with principal component analysis (PCA)

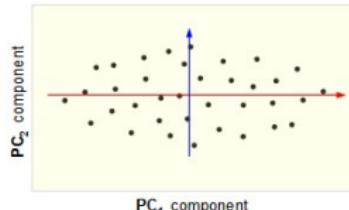
- Directions of **max variance** of the data
- 3-D data, but distributed mostly on a **2-D** surface \Rightarrow Find it



a



b

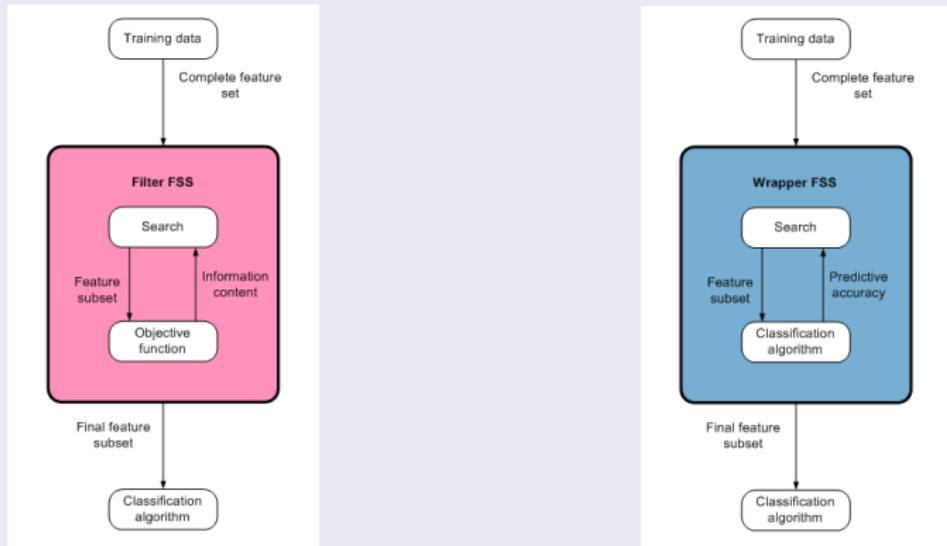


c

Supervised classification

Less is more. Feature subset selection (FSS)

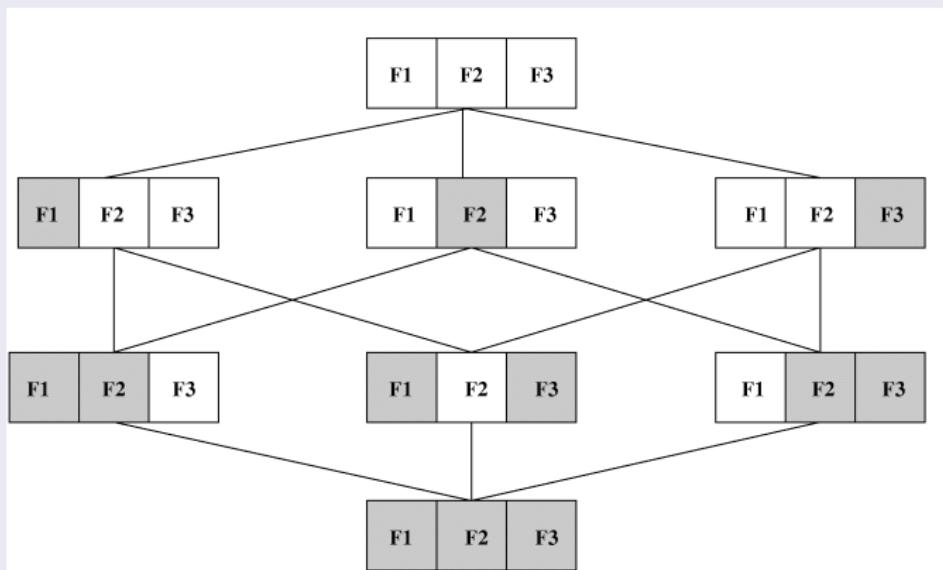
- Requirement: Scoring function to measure the quality of a subset
- Filter approach: intrinsic characteristics of the data
- Wrapper approach: knowledge about the classifier



Supervised classification

Less is more. Feature subset selection (FSS)

- FSS as a **combinatorial optimization** problem: search strategies (forward, backward, genetic algorithms,)
- Cardinality of the **search space**: 2^n



Supervised classification

Measuring the performance. Confusion matrix

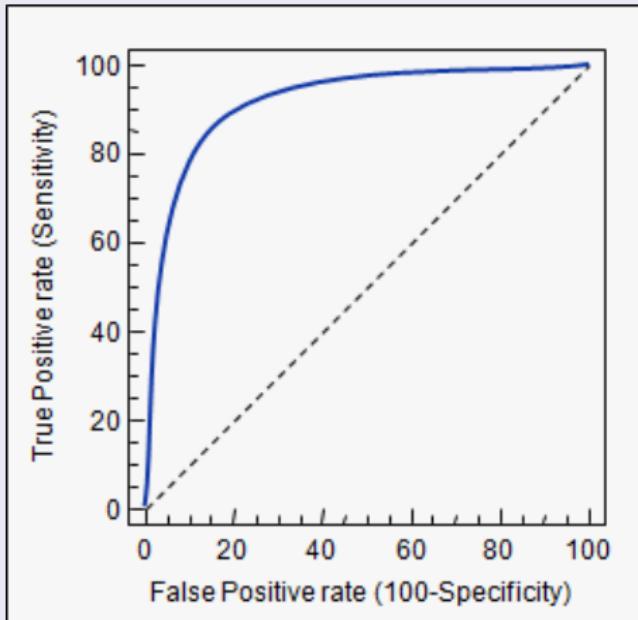
		C True class	
		+	-
C_M Predicted class	+	a	b
	-	c	d

Figures of merit

- Accuracy: $\frac{a+d}{a+b+c+d}$
- Error rate: $\frac{c+b}{a+b+c+d}$
- Rate of true positives (sensitivity): $\frac{a}{a+c}$
- Rate of true negatives (specificity): $\frac{d}{b+d}$

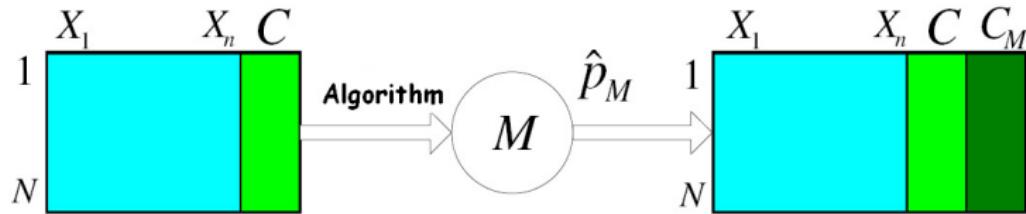
Supervised classification

ROC curve. Area under the ROC curve (AUC)



Supervised classification

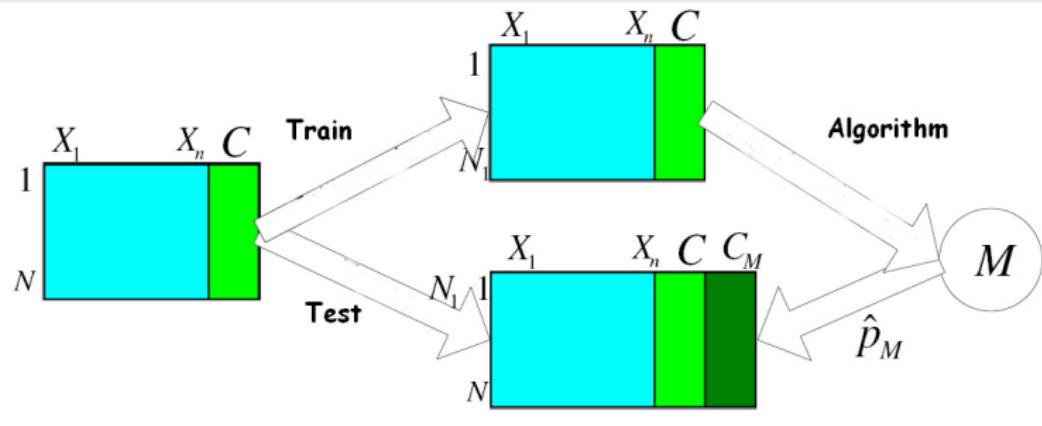
Estimation methods. No honest



$$\hat{p}_M = \frac{1}{N} \sum_{i=1}^N \delta(c^{(i)} = c_M^{(i)})$$

Supervised classification

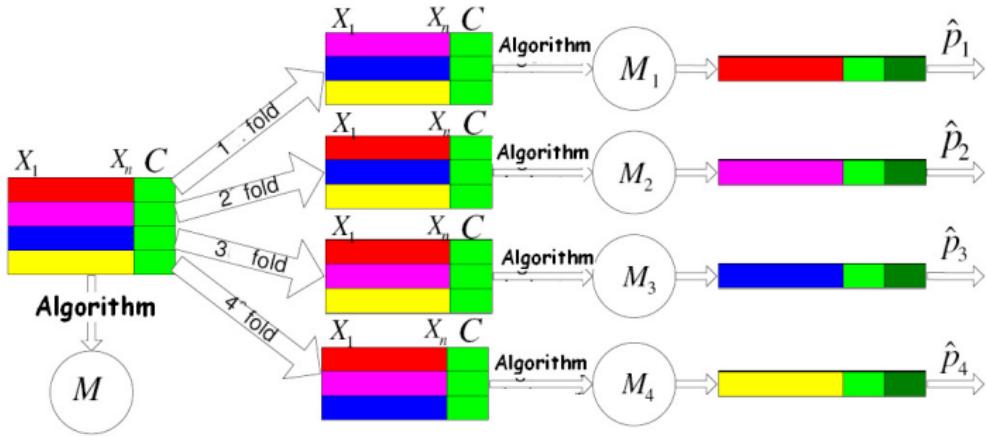
Estimation methods. Train and test



$$\hat{p}_M = \frac{1}{N - N_1} \sum_{i=1}^{N - N_1} \delta(c^{(N_1+i)} = c_M^{(N_1+i)})$$

Supervised classification

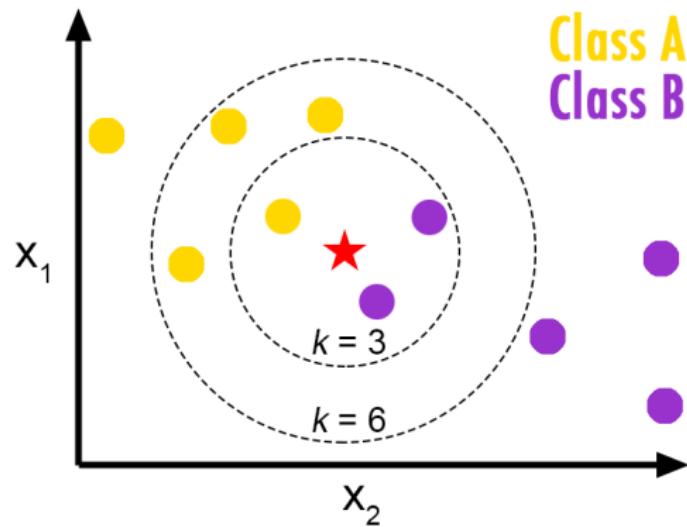
Estimation methods. k -fold cross validation



$$\hat{p}_M = \frac{1}{k} \sum_{i=1}^k \hat{p}_i$$

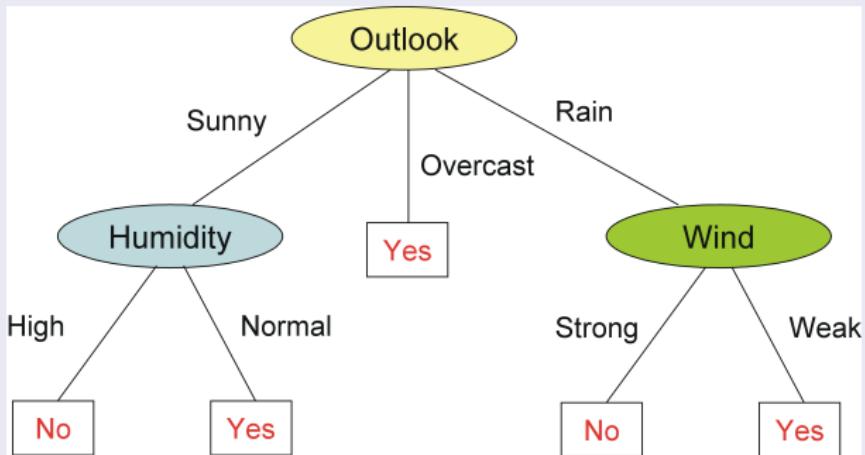
Supervised classification

Classifiers. k -NEAREST NEIGHBORS



Supervised classification

Classifiers. CLASSIFICATION TREE



- Equivalent set of rules:

- R1: If (*Outlook*=Sunny) and (*Humidity*=High) then *PLAYTENNIS*=**No**
- R2: If (*Outlook*=Sunny) and (*Humidity*=Normal) then *PLAYTENNIS*=**Yes**
- R3: If (*Outlook*=Overcast) then *PLAYTENNIS*=**Yes**
- R4: If (*Outlook*=Rain) and (*Wind*=Strong) then *PLAYTENNIS*=**No**
- R5: If (*Outlook*=Rain) and (*Wind*=Weak) then *PLAYTENNIS*=**Yes**

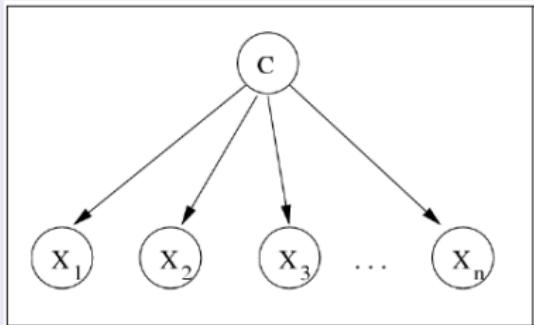
Supervised classification

Classifiers. NAIVE BAYES

Predictor variables are **conditionally independent given C**

$$P(c|x_1, \dots, x_n) \propto P(C = c) \prod_{i=1}^n P(X_i = x_i | C = c)$$

$$\Rightarrow c^* = \arg \max_c P(C = c) \prod_{i=1}^n P(X_i = x_i | C = c)$$

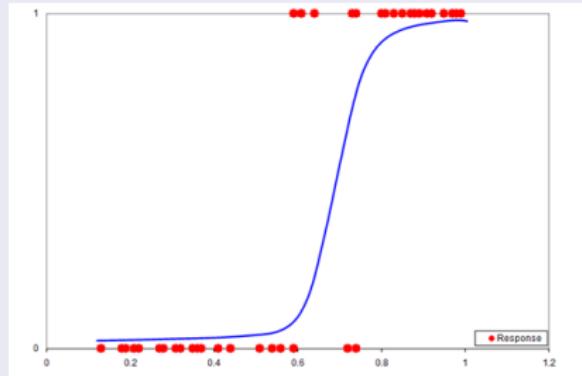


Supervised classification

Classifiers. LOGISTIC REGRESSION

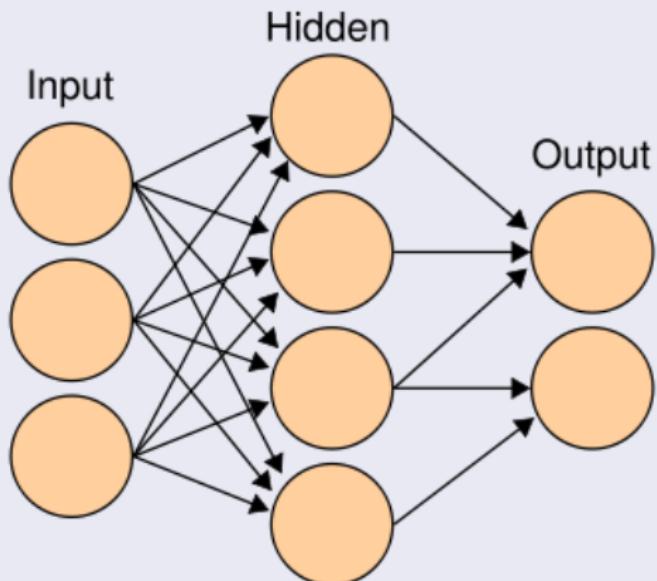
$$\pi_j = P(C = 1 | \mathbf{x}_j) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{j1} + \dots + \beta_n x_{jn})}}$$

$$\Rightarrow 1 - \pi_j = P(C = 0 | \mathbf{x}_j) = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_{j1} + \dots + \beta_n x_{jn})}}$$



Supervised classification

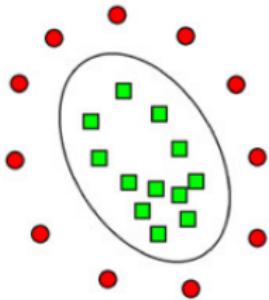
Classifiers. NEURAL NETWORK



Supervised classification

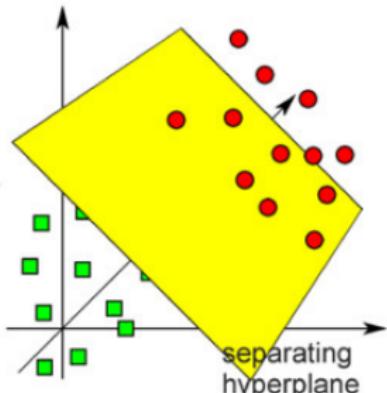
Classifiers. SUPPORT VECTOR MACHINE

Separation may be easier in higher dimensions



complex in low dimensions

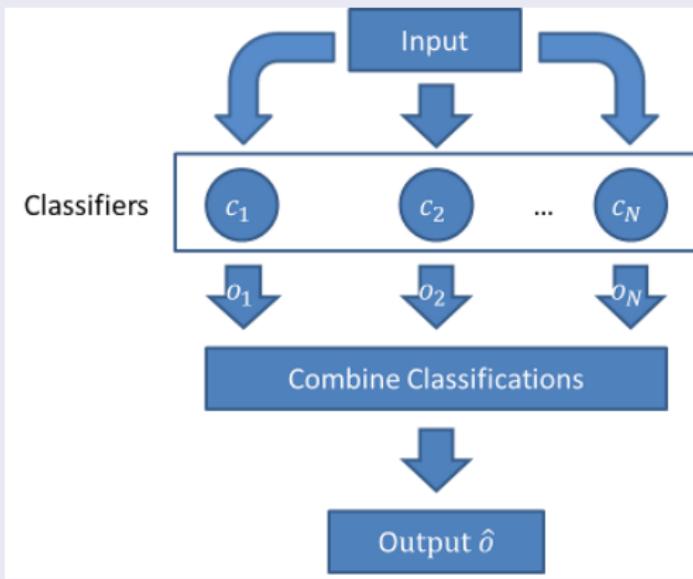
feature map



simple in higher dimensions

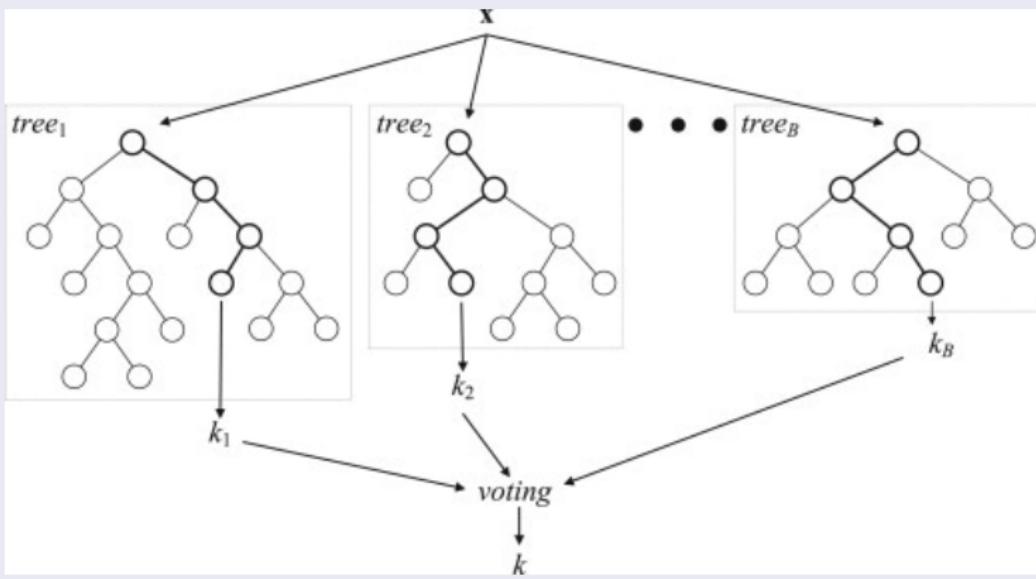
Supervised classification

Classifiers. METACLASSIFIERS



Supervised classification

Classifiers. METACLASSIFIERS. RANDOM FOREST



Supervised classification

Case study on diagnose aviation turbulence

- Predominant cause of accidents and injuries
- Costing airlines millions of euros per year in compensation: aircraft damage, and delays due to post-event inspections and repairs
- Attempts to avoid turbulent airspace cause: flight delays and en route deviations, increasing air traffic controller workload, disrupt schedules of air crews and passengers and use extra fuel



WILLIAMS JK (2013). Using random forest to diagnose aviation turbulence. *Machine Learning*, in press

Supervised classification

Case study on diagnose aviation turbulence

- Data sets (March 10 to November 4, 2010)
 - UAL from 95 United Airlines, Boeing 757: 5,623,738 instances
 - DAL from 80 Delta Air Lines, Boeing 737: 6,595,922 instances
 - Predictor variables: Numerical weather predictors, Doppler radar, Geostationary satellite images, Lightning detection network, Derived fields and features
 - Class variable: Eddy dissipation rate (EDR) an aircraft-independent atmospheric turbulence metric
- Feature subset selection: wrapper forward/backward selection guided by the AUC
- Supervised classification methods: k -nearest neighbor ($k = 100$); logistic regression; and random forest (200 trees)
- Validation: 10-fold cross-validation repeated 32 times with AUC as performance metric

Supervised classification

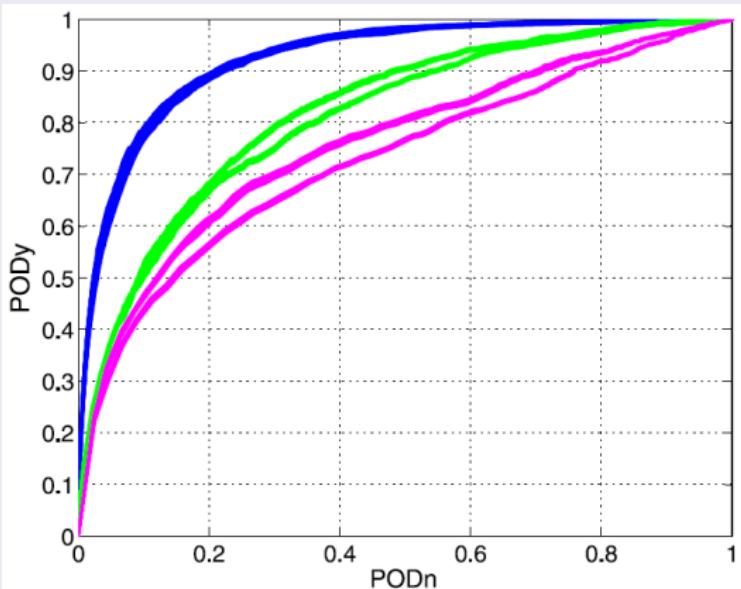
Case study on diagnose aviation turbulence

Table 2 Forward/backward selection procedure results for RF and logistic regression

Random forest (8)				Logistic regression (8)			
Rank	Mean occ.	Predictor name		Rank	Mean occ.	Predictor name	
1	115	Dist. to NSSL echo top > 10 kft		1	135	Model FRNTGTHRI	
2	114	Model FRNTGTHRI		2	134	Diff. Alt. to 80-km max NTDA sev. top	
3	104	Model RITW		3	127	Dist. to NSSL echo top > 10 kft	
4	89	Model ELLROD2		4	126	10-km max of NSSL echo top	
5	88	Diff. Alt. to 80-km max NTDA sev. top		5	121	Model ELLROD2	
6	85	Model MWT2		6	111	Model RITW	
7	79	Model ELLROD1		7	107	Model BROWN2	
8	78	160-km mean of Satellite Ch. 6		8	94	Diff. Alt. to 20-km max NTDA mod. top	
9	69	Model F2DTW		9	92	Model BROWN1	
10	68	Model MWT3		10	89	Model ELLROD1	
11	68	40-km min of Satellite Ch. 6		11	88	Model MWT3	
12	67	Model Atm. Pressure		12	87	Model EDRRI	
13	66	Model BROWN2		13	85	Model DTF3	
14	65	Satellite Ch. 4 minus Model temp.		14	83	20-km no. of good NTDA dBZ points	
15	64	Model DUTTON		15	77	10-km no. of good NTDA dBZ points	
16	63	Satellite Ch. 4 minus Satellite Ch. 3		16	76	10-km mean of NTDA composite EDR	
17	58	Model NGM2		17	74	10-km max of NTDA composite EDR	
18	56	160-km mean of Satellite Ch. 4		18	74	10-km min of Satellite Ch. 3	
19	53	Model RICH		19	73	10-km mean of NSSL echo top	
20	52	Diff. Model pres. to Mod. surf. pres.		20	69	Model IAWINDRI	

Supervised classification

Case study on diagnose aviation turbulence



PODy= TPR; PODn= FPR; Random forest (blue) with AUC=0.924; k -NN (green) with AUC=0.915; logistic regression (green) with AUC=0.915; Graphical turbulence guidance (magenta) with AUC=0.816; storm distance (magenta) with AUC=0.743

Multi-label classification

X_1	X_2	X_3	X_4	X_5	C
3.2	1.4	4.7	7.5	3.7	1
2.8	6.3	1.6	4.7	2.7	0
7.7	6.2	4.1	3.3	7.7	1
9.2	0.4	2.8	0.5	3.9	0
5.5	5.3	4.9	0.6	6.6	1

X_1	X_2	X_3	X_4	X_5	C_1	C_2	C_3	C_4
3.2	1.4	4.7	7.5	3.7	1	0	1	1
2.8	6.3	1.6	4.7	2.7	0	0	1	0
7.7	6.2	4.1	3.3	7.7	1	0	1	1
9.2	0.4	2.8	0.5	3.9	0	1	0	0
5.5	5.3	4.9	0.6	6.6	1	1	0	1

Multi-label classification

Case study on the ASRS database

- The Aviation Safety Reporting System (ASRS) database spans 30 years and contains over 700,000 aviation safety reports in free text form
- Primary concern is system health and safety: detection, diagnosis, prediction, mitigation, and prevention of ongoing and future system problems
- ASRS reports are publicly available and are written by pilots, flight controllers, technicians, flight attendants, and others including passengers
- Predictor variables, X_1 to $X_{25,729}$: 25,729 terms that occur in at least one document
- Class variables to be predicted, C_1 to C_{60} : 60 problem types that appear during flights
- A subset of 28,596 documents containing 22 classes was used

Oza N, ET AL. (2009). Classification of aeronautics system health and safety documents. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 39(6), 670-680

Multi-label classification

Case study on the ASRS database

Auto.Categorization Application - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://demo.sevensimplemachines.com/NASA/AutoCat/Default.aspx

Add new user
Log out
Stop Monitoring
User Id : admin

Auto-Classification Tool

Choose Sample

Please choose a year and a month to get started.
Click the 'Choose' button.

Year: 2003 Month: November Choose

Events

- 53 : TRAINING FOR COMPLEX ACFT
- 61 : I WAS WORKING THE MEACHUM
- 62 : I HAD TAXIED OUT TO THE R
- 64 : I HAD JUST LANDED ON RWY
- 77 : I AM FILING THIS RPT AS A
- 80 : ON IFR FLT PLAN IN SEVERE
- 81 : DFW SOLD TWO DEP IMPROPE
- 82 : ARRIVED ON CHARTER WITH P
- 1100 : I WAS FLYING ON AN IFR FL

Analysis

64 Processed

I HAD JUST LANDED ON RWY 28. SKY WAS CLR, WIND WAS FROM ABOUT 250 DEGS AT 7-9 KTS. I CLRED THE ACTIVE RWY AT TXWY A, AND ANNOUNCED ON CTAF SKYHAWK 172, CLR OF THE ACTIVE. MOST OF THE ACFT ARE HANGARDED ON THE S SIDE OF THE AIRFIELD, WHILE THE TXWYS AND FBO ARE ON THE N SIDE OF THE FIELD. THIS REQUIRES THAT AFTER LDNG, MOST OF THE ACFT MUST CROSS THE ACTIVE AT A POINT AWD ACFT IS THE ACCEPTED CUSTOM FOR ALL XING ACFT TO ANNOUNCE THEIR INTENTION TO CROSS THE ACTIVE AND THEN ANNOUNCE WHEN THEY ARE CLR. I TAXIED TOWARDS THE MIDFIELD CROSSOVER, AND ANNOUNCED SKYHAWK 172 XING THE ACTIVE RWY MIDFIELD ON THE GND.' I WAS STILL 15 SECONDS FROM THE HOLD SHORT LINE AT THIS POINT, AND STILL ON THE TXWY, LOOKING DOWN THE ACTIVE RWY. I HAD HEARD ANOTHER CESSNA ANNOUNCE ON CTAF THAT HE WAS TURNING ONTO A 3 M FINAL, AND COULD SEE HIS LDNG LIGHTS IN THE DISTANCE. AT THIS POINT, I HEARD A DIAMOND STAR 2 SEATER THAT HAD DEPARTED AFTER I HAD, ANNOUNCE STRAIGHT-IN 5 M FINAL FOR RWY 28. I STOPPED AT THE HOLD SHORT LINE, AND LOOKED DOWN THE RWY, AND NOW SAW BOTH INBOUND ACFT, EASILY VISIBLE BECAUSE OF THEIR LDNG LIGHTS. NEITHER ACFT WAS A FACTOR, SO I PROCEEDED ONTO ACTIVE RWY. JUST AS I PASSED THE HOLD SHORT

Processed

I have finished with this event. Please record my results. Done

Category	Confidence
<input checked="" type="checkbox"/> Incursions	*****
<input checked="" type="checkbox"/> Aircraft Damage Or Encounters	*****
<input checked="" type="checkbox"/> Departure Problems	*****

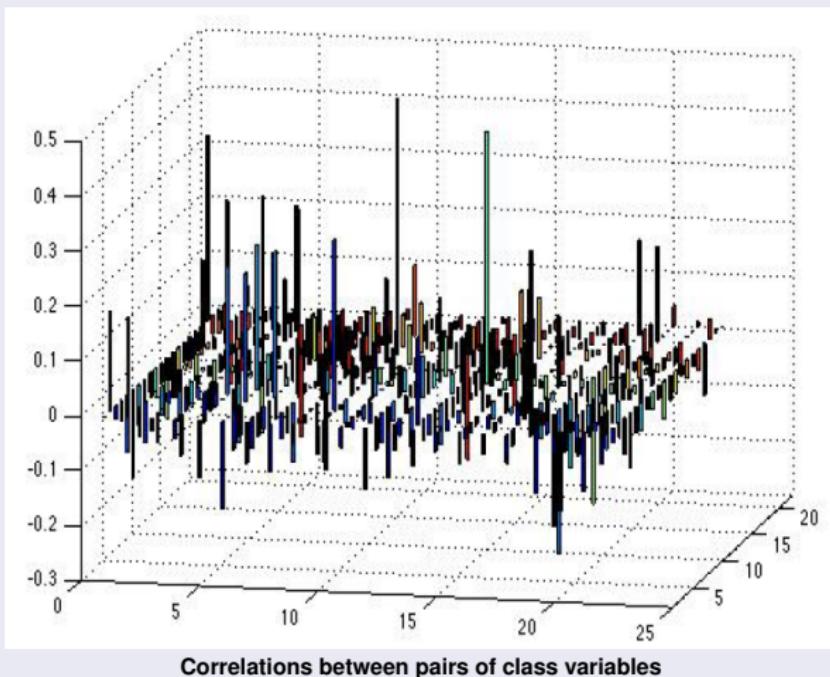
Additional Categories

- Aircraft malfunction event-Airframe
- Aircraft malfunction event-Structures
- Aircraft malfunction event-Propeller/Rotor
- Aircraft malfunction event-Power plant/Engine
- Aircraft malfunction event-Charts
- Fire Smoke or Fumes
- Illness or Injury Events
- Security Concerns
- Evacuation Event
- Safety event/concern
- Coordination/Communication Issue
- Datalink Coordination/Communication Events
- Airworthiness - Documentation
- Operation in noncompliance

Done

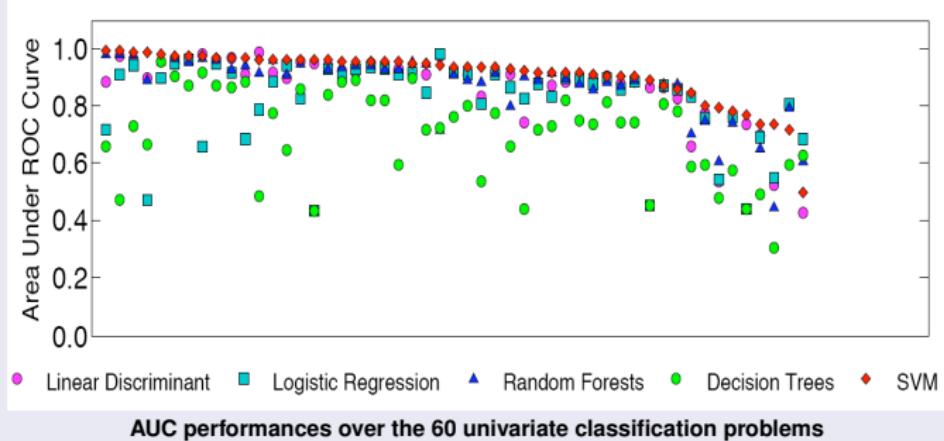
Multi-label classification

Case study on the ASRS database



Multi-label classification

Case study on the ASRS database



Outline

1 Introduction

2 Supervised Classification

3 Clustering

4 Conclusions

Clustering

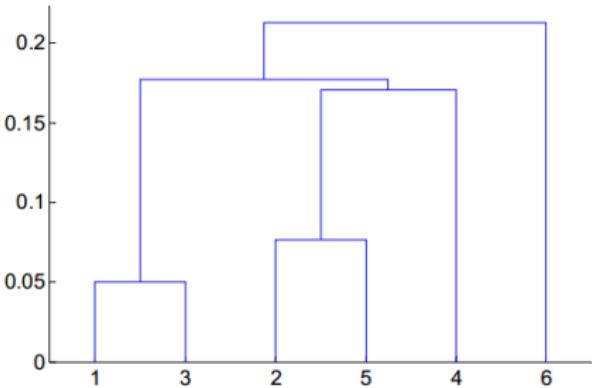
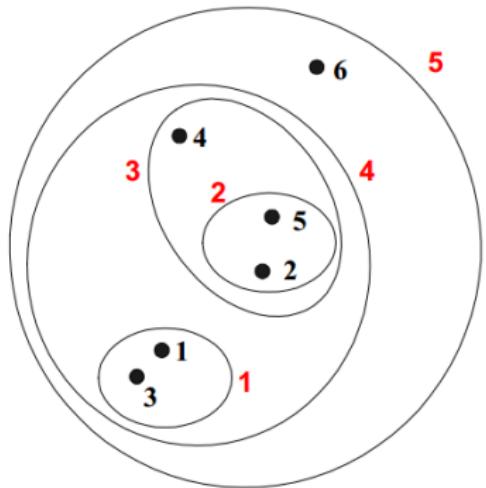
A general dataset

	X_1	\dots	X_n
$\mathbf{x}^{(1)}$	$x_1^{(1)}$	\dots	$x_n^{(1)}$
\dots		\dots	
$\mathbf{x}^{(N)}$	$x_1^{(N)}$	\dots	$x_n^{(N)}$

- Grouping objects in clusters
- High similarity within each cluster
- High dissimilarity between the different clusters
- Main methods:
 - Hierarchical clustering
 - Partitional clustering
 - Probabilistic clustering

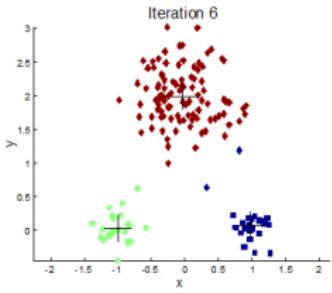
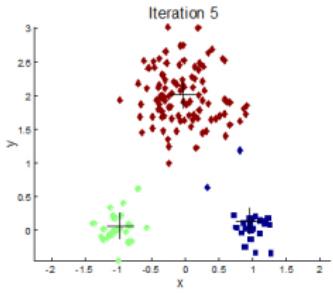
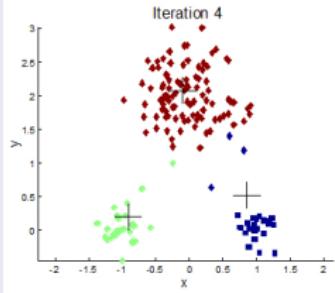
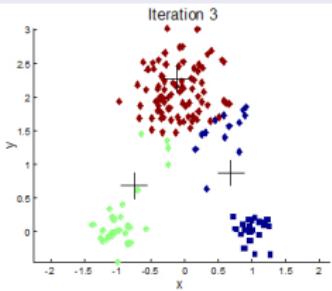
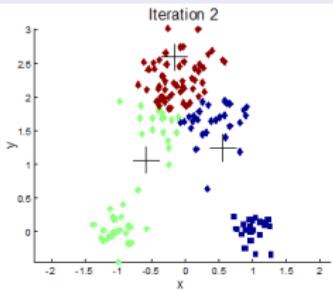
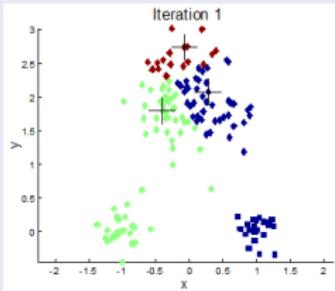
Clustering

HIERARCHICAL CLUSTERING



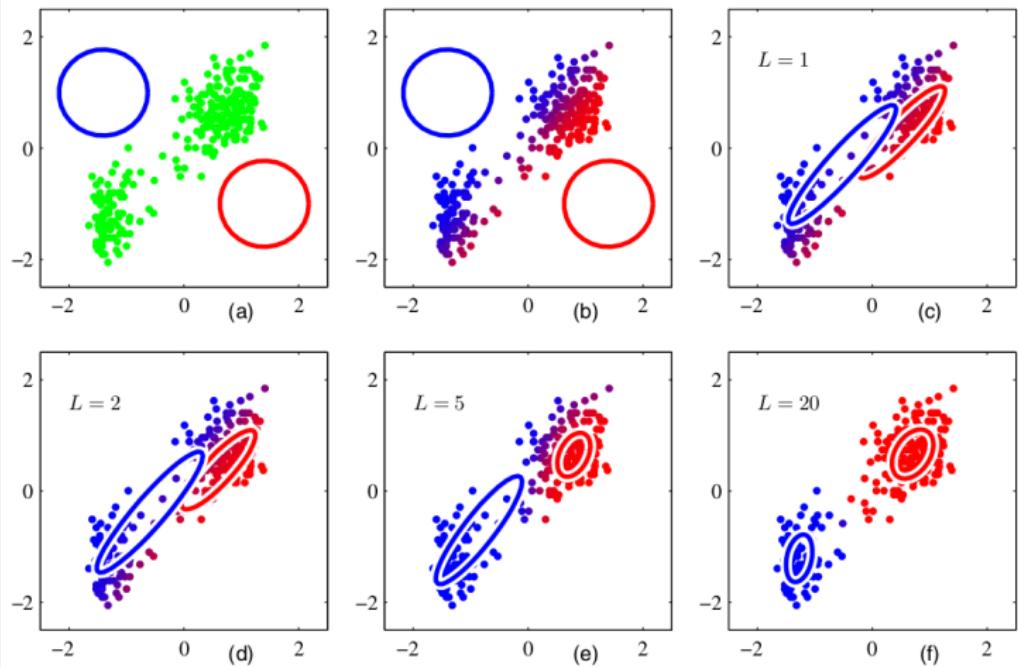
Clustering

PARTITIONAL CLUSTERING: *k*-means



Clustering

PROBABILISTIC CLUSTERING: finite mixture models with EM



Clustering

Case study on detecting anomalous landing

- A key question on the aviation safety domain
- For a specific aircraft make and model at a specific airport
- Variables: sequence of switches that a pilot flips during the course of the landing phase of the flight
- Data set: $N= 2,200$ flights, and $n= 1,500$ possible switches
- Cluster algorithm: *k-medoids*
- Anomalous landing: a sequence that is far away from the cluster centroid



BUDALAKOTI ET AL. (2009). Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 39(1), 101-113

Outline

1 Introduction

2 Supervised Classification

3 Clustering

4 Conclusions

Conclusions

Machine learning in aviation

- Aviation industry generates large scale data
- Transform these data sets into knowledge
- Machine learning methods:
 - Supervised classification
 - Clustering
- Advances in the safety, security, and efficiency of civil aviation

References on Supervised Classification

- BIELZA C, LI G, LARRAÑAGA P (2011). Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, 52, 705-727
- BISHOP CM (2006). *Pattern Recognition and Machine Learning*. Springer
- BISHOP CM (1995). *Neural Networks for Pattern Recognition*. Oxford University Press
- BREIMAN L, FRIEDMAN J, OLSHEN R, STONE C (1984). *Classification and Regression Trees*. Chapman and Hall
- COVER TM, HART PE (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27
- HASTIE H, TIBSHIRANI R, FRIEDMAN J (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer
- KUNCHEVA LI (2004). *Combining Pattern Classifiers*. John Wiley and Sons
- MURPHY KP (2012). *Machine Learning. A Probabilistic Perspective*. The MIT Press
- LIU H, MOTODA H (1998) *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers
- SAEYS Y, INZA I, LARRAÑAGA P (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517
- SHAWE-TAYLOR J, CRISTIANINI N (2000). *Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press
- ZHANG ML, ZHOU ZH (2013). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, in press

References on Clustering

- FORGY EW (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21, 768-769
- FREY BJ, DUECK D (2007). Clustering by passing messages between data points. *Science*, 315 (5814), 972-976
- HARTIGAN JA (1975). *Clustering Algorithms*. John Wiley and Sons
- JAIN AK, MURTY MN, FLYNN PJ (1999). Data clustering: A review. *ACM Computing Surveys*, 31, 3, 264-323
- JOHNSON SC (1967). Hierarchical clustering schemes. *Psychometrika*, 2, 241-254
- KAUFMAN L, ROUSSEEUW P (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley and Sons
- MACQUEEN JB (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297

MACHINE LEARNING IN AVIATION

PEDRO LARRAÑAGA

Computational Intelligence Group
Artificial Intelligence Department
Technical University of Madrid, Spain



Toulouse, July 11, 2013