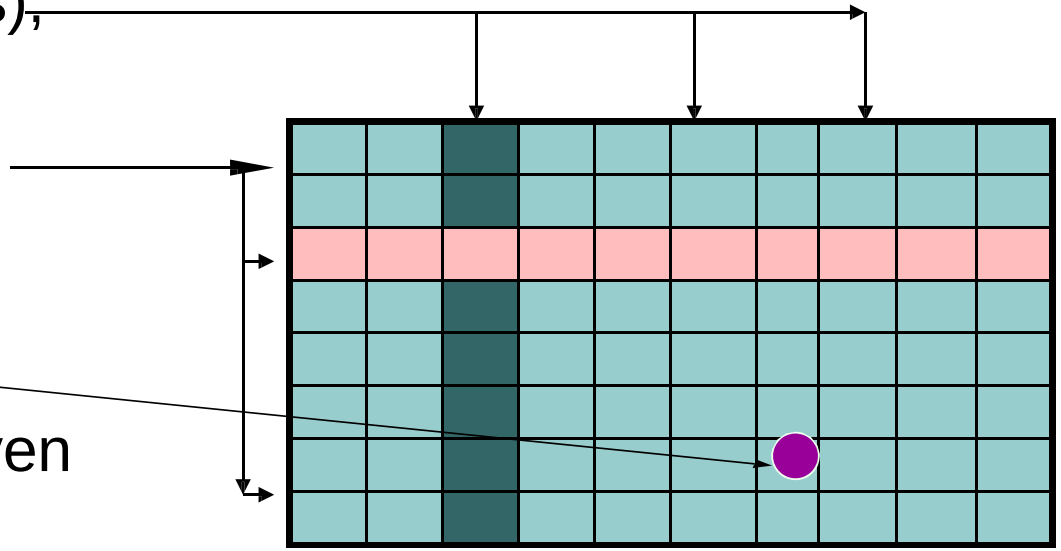
TEMA 2b

Reducció de les dades

Dimensions Reduction Of Large Data Sets

Main dimensions:

- **columns** (features),
- **rows** (cases or samples),
- **values** of the features for the given sample



Dimensions Reduction Of Large Data Sets

Why dimensionality reduction?

1. Reduced Computing Time
2. Increased predictive/descriptive accuracy
3. Better representation of the data-mining model

Why Features Reduction?

- Most data mining techniques may not be effective for high-dimensional data
  - Curse of Dimensionality
  - Accuracy and efficiency may degrade rapidly as the dimension increases.
- The intrinsic dimension may be small.
  - *For example, the number of genes responsible for a certain type of disease may be small.*

Feature Reduction

Which features to select, and how?

TRS_DT	TRS_TYP_CD	REF_DT	REF_NUM	CO_CD	GDS_CD	QTY	UT_CD	UT_PRIC
21/05/93	00001	04/05/93	25119	10002J	001M	10	CTN	22.000
21/05/93	00001	05/05/93	25124	10002J	032J	200	DOZ	1.370
21/05/93	00001	05/05/93	25124	10002J	033Q	500	DOZ	1.000
21/05/93	00001	13/05/93	25217	10002J	024K	5	CTN	21.000
21/05/93	00001	13/05/93	25216	10026H	006C	20	CTN	69.000
21/05/93	00001	13/05/93	25216	10026H	008Q	10	CTN	114.000
21/05/93	00001	14/05/93	25232	10026H	006C	10	CTN	69.000
21/05/93	00001	14/05/93	25235	10027E	003A	5	CTN	24.000
21/05/93	00001	14/05/93	25235	10027E	001M	5	CTN	24.000
21/05/93	00001	22/04/93	24974	10035E	009F	50	CTN	118.000
21/05/93	00001	27/04/93	25033	10035E	015A	375	GRS	72.000
21/05/93	00001	20/05/93	25313	10041Q	010F	10	CTN	26.000
21/05/93	00001	12/05/93	25197	10054R	002E	25	CTN	24.000

Features Reduction

Two standard approaches:

1. ***Feature selection:*** A process that chooses an optimal subset of features according to an objective function:

   - feature ranking algorithms, and
   - minimum subset algorithms.

2. ***Feature extraction:*** refers to the mapping of the original high-dimensional data onto a lower-dimensional space. Criterion for :

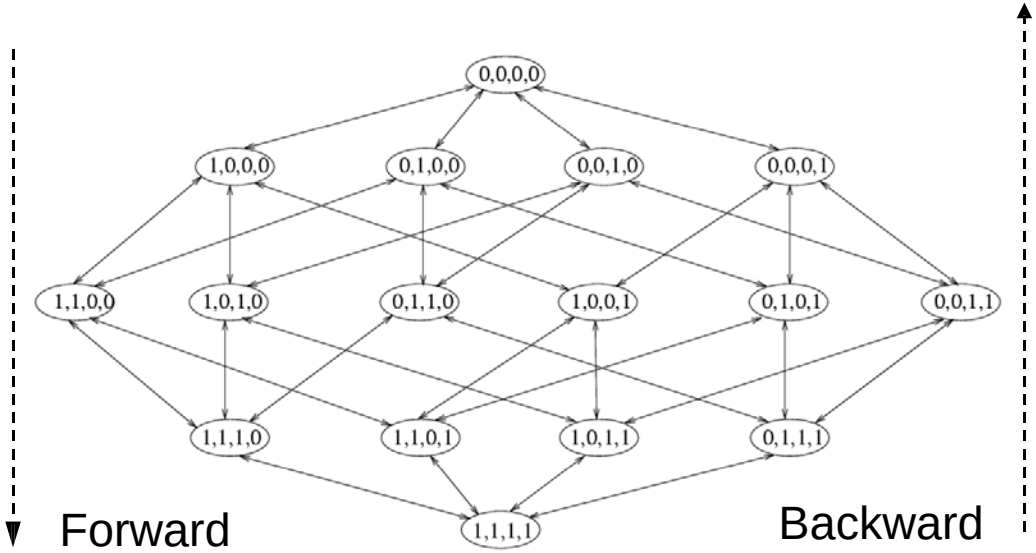
   - *Descriptive setting*: minimize the information loss
   - *Predictive setting*: maximize the class discrimination

Índex

- Tractament d'atribut
  - **Reducció d'atributs (feature selection)**
    - Estadistics
    - Filter (Reaper)
    - Wrappers
  - Extracció (o formació) de característiques
    - PCA
- Reducció de casos
- Reducció de valors

Feature Selection

- **Feature selection** in general can be viewed as a search problem (2^N).
- For practical methods, an optimal search is not feasible, and simplifications are made to produce acceptable and timely reasonable results:
  - heuristic criteria
  - bottom-up approach
  - top-down approach



Methods of Feature Selection

- Statistic methods
- Filter methods
  - Separating feature selection from classifier learning
  - Relying on general characteristics of data (*information, distance*, *dependence*, *consistency*)
  - No bias toward any learning algorithm, fast
- Wrapper methods
  - Relying on a predetermined classification algorithm.
  - Using predictive accuracy as goodness measure
  - High accuracy, computationally expensive
- Embedded methods
  - Combine Filter and Wrapper approaches

Methods of Feature Selection

- Statistic methods
- Filter methods
  - Separating feature selection from classifier learning
  - Relying on general characteristics of data (*information, distance*, *dependence*, *consistency*)
  - No bias toward any learning algorithm, fast
- Wrapper methods
  - Relying on a predetermined classification algorithm.
  - Using predictive accuracy as goodness measure
  - High accuracy, computationally expensive
- Embedded methods
  - Combine Filter and Wrapper approaches

Features Selection: Statistic methods

1. **Comparison of *means* and *variances:***

Samples of two classes (A and B) can be examined:

$$SE(A\text{-}B) = \sqrt{(var(A)/n_1 + var(B)/n_2)}$$

TEST:

$$|mean(A) - mean(B)| / SE(A\text{-}B) > \text{threshold-value}$$

where n_1 and n_2 are the corresponding number of samples for classes A and B.

Features Selection: Statistic methods

Comparison of means and variances (EXAMPLE):

X	Y	C
0.3	0.7	A
0.2	0.9	B
0.6	0.6	A
0.5	0.5	A
0.7	0.7	B
0.4	0.9	B

Threshold value is 0.5

X_A = {0.3, 0.6, 0 .5}, X_B = {0.2, 0.7, 0.4},

Y_A = {0.7, 0.6, 0.5}, and Y_B = {0.9, 0.7, 0.9}

Features Selection: Statistic methods

Comparison of means and variances (EXAMPLE):

$$SE(X_A\text{-}X_B) = \sqrt{(var(X_A)/n_1 + var(X_B)/n_2)} = \sqrt{(0.0233/3 + 0.6333/3)} = 0.4678$$

$$SE(Y_A\text{-}Y_B) = \sqrt{(var(Y_A)/n_1 + var(Y_B)/n_2)} = \sqrt{(0.01/3 + 0.0133/3)} = 0.0875$$

Tests:

$|mean(X_A) - mean(X_B)| / SE(X_A\text{-}X_B) = |0.4667 - 0.4333| / 0.4678 = 0.0735 < \mathbf{0.5}$

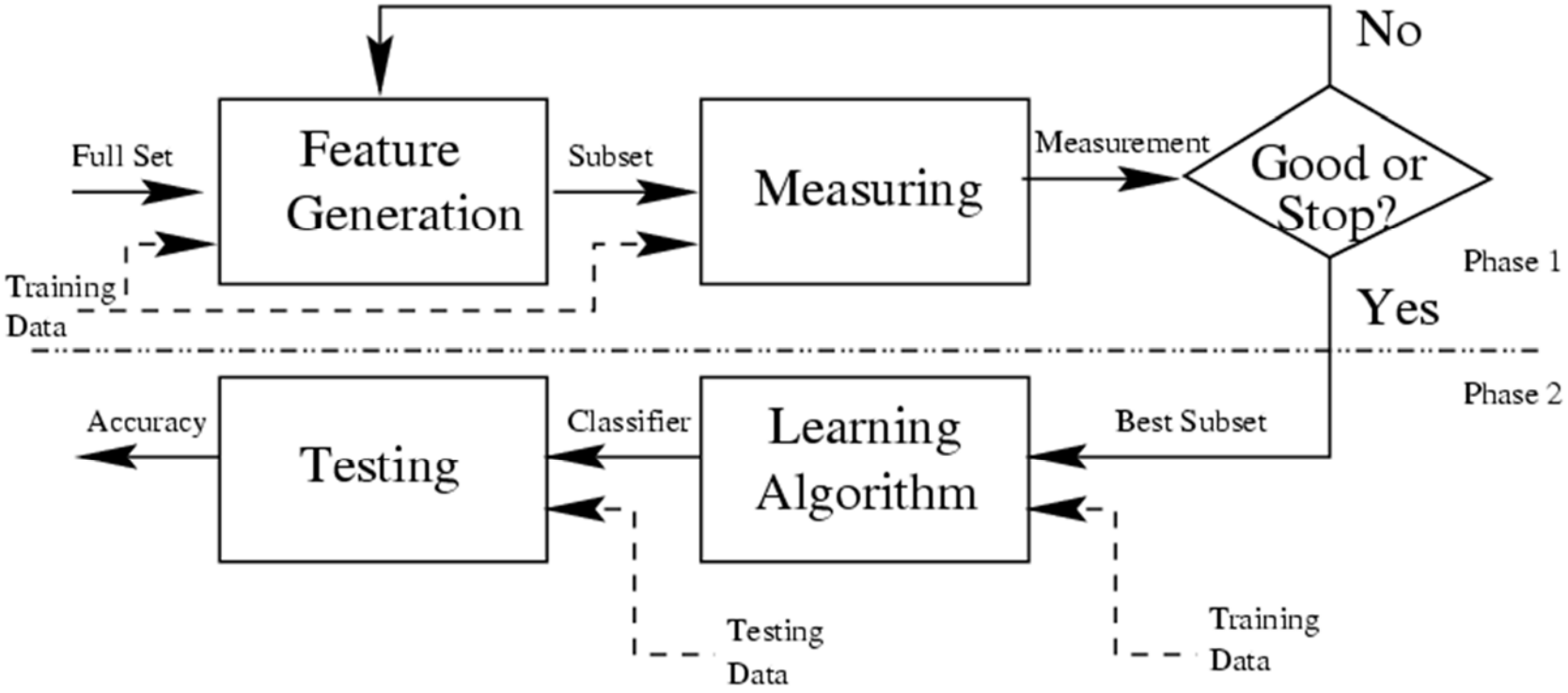
$|mean(Y_A) - mean(Y_B)| / SE(Y_A\text{-}Y_B) = |0.6 - 0.8333| / 0.0875 = 2.6667 > \mathbf{0.5}$

- X is a candidate feature for reduction because its mean values are close, and therefore the final test is below threshold value.

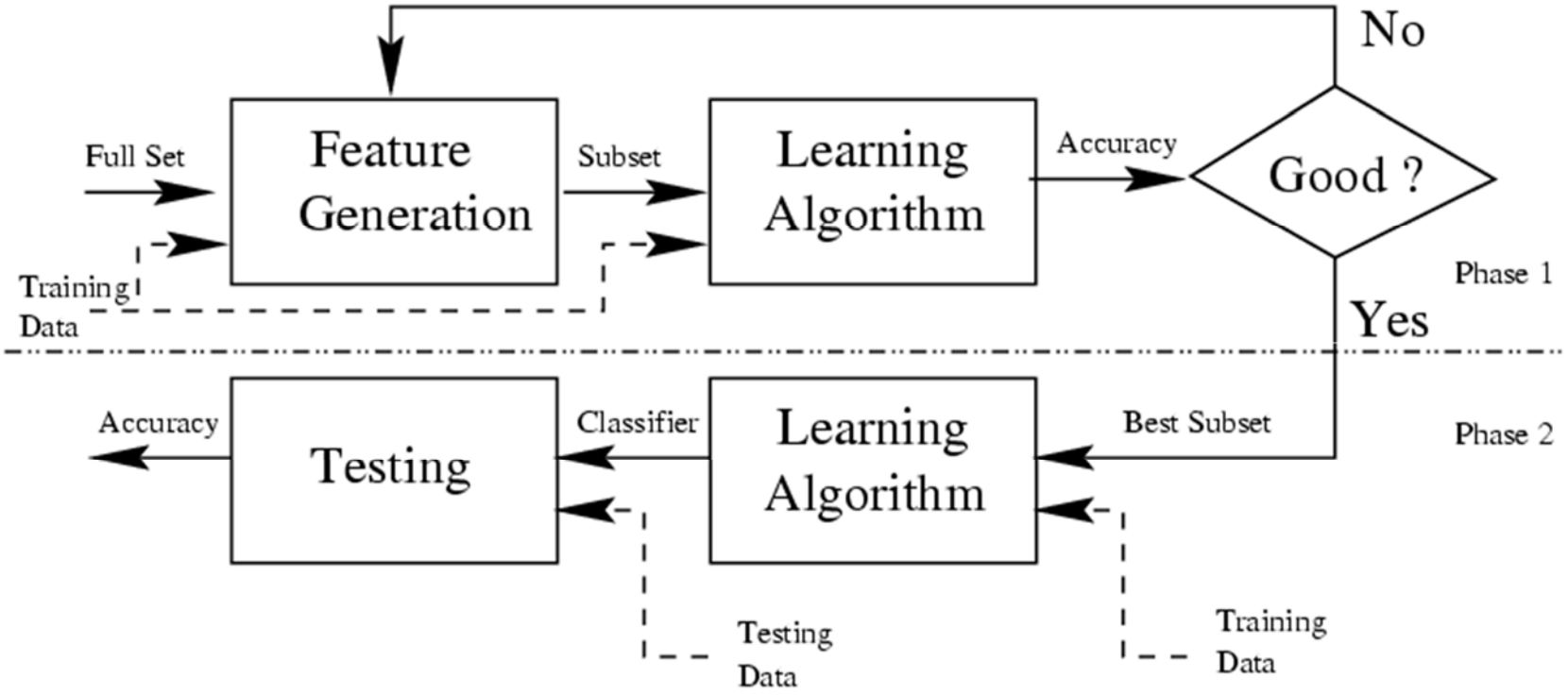
Methods of Feature Selection

- Statistic methods
- Filter methods
  - Separating feature selection from classifier learning
  - Relying on general characteristics of data (*information, distance*, *dependence*, *consistency*)
  - No bias toward any learning algorithm, fast
- Wrapper methods
  - Relying on a predetermined classification algorithm.
  - Using predictive accuracy as goodness measure
  - High accuracy, computationally expensive
- Embedded methods
  - Combine Filter and Wrapper approaches

Filter Model



Wrapper Model



Representative Algorithms for Feature Selection

- Filter algorithms
  - Feature ranking algorithms
    - Example: Relief (*Kira & Rendell 1992*)
- Wrapper algorithms
  - Feature ranking algorithms
    - Example: SVM

Feature Selection: Relief

Basic algorithm construct :

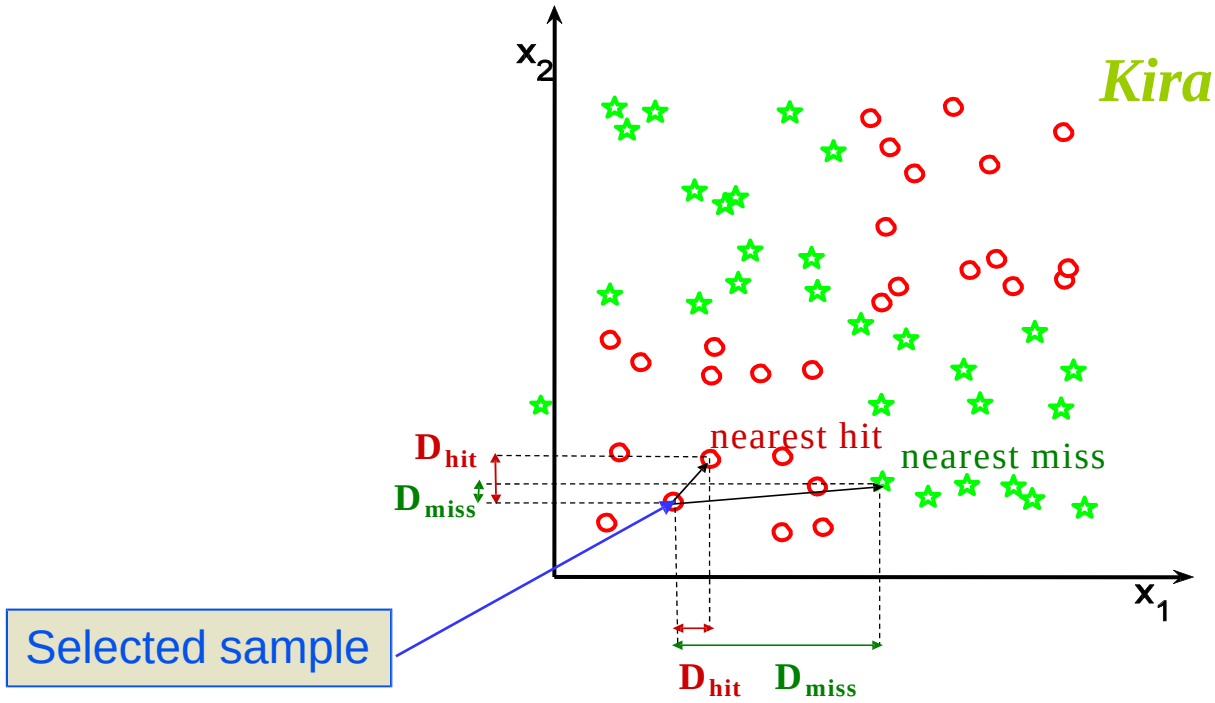
- Each feature is assigned with cumulative weight computed over a subset of the training data set.
- Feature with weight over a certain threshold is the selected feature subset.

Assignment of weightage :

- X randomly selected sample
- near-hit sample = the closest sample from the same class.
- near-miss sample =the closest sample from the different class.

$$W = - \text{diff}(X,\text{near-hit})^2 + \text{diff}(X,\text{near-miss})^2$$

Feature Selection: Relief



Kira and Rendell, 1992

Feature Selection: Relief

Relief

Input: N - features
m - number of instances sampled
τ - adjustable relevance threshold

initialize: $\mathbf{w} = 0$
for $i = 1$ to m
begin
randomly select an instance I
find nearest-hit H and nearest-miss J
for $j = 1$ to N
$\mathbf{w}(j) = \mathbf{w}(j) - \mathtt{diff}(j, I, H)^2/m + \mathtt{diff}(j, I, J)^2/m$
end

Output: $\mathbf{w}$ greater than τ

Feature Selection: Relief

Relief

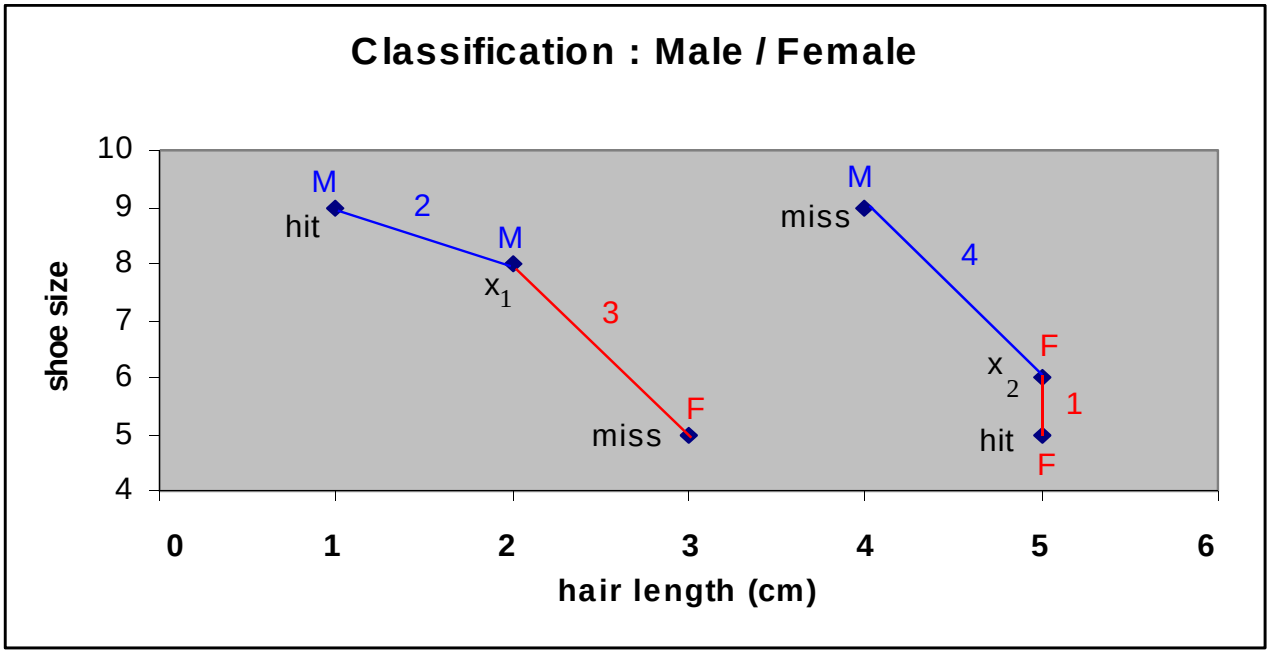
Input: N - features
m - number of instances sampled
τ - adjustable relevance threshold

initialize: $\mathbf{w} = 0$
for $i = 1$ to m
begin
randomly select an instance I
find nearest-hit H and nearest-miss J
for $j = 1$ to N
$\mathbf{w}(j) = \mathbf{w}(j) - \mathtt{diff}(j, I, H)^2/m + \mathtt{diff}(j, I, J)^2/m$
end

Output: $\mathbf{w}$ greater than τ

Time consuming:
Tree structured data!

Feature Selection: Relief



Relief -Advantages:

- noise-tolerant.
- unaffected by feature interaction.
- fast – linear complexity

feature	x	w	$-(x-hit)^2$	$+(x-miss)^2$	=w	x	w	$-(x-hit)^2$	$+(x-miss)^2$	=w
shoe size	x_1	0	$-(8-9)^2$	$+(8-5)^2$	+8	x_2	8	$-(6-5)^2$	$+(6-9)^2$	+8
hair length	x_1	0	$-(2-1)^2$	$+(2-3)^2$	0	x_2	0	$-(5-5)^2$	$+(5-4)^2$	+1

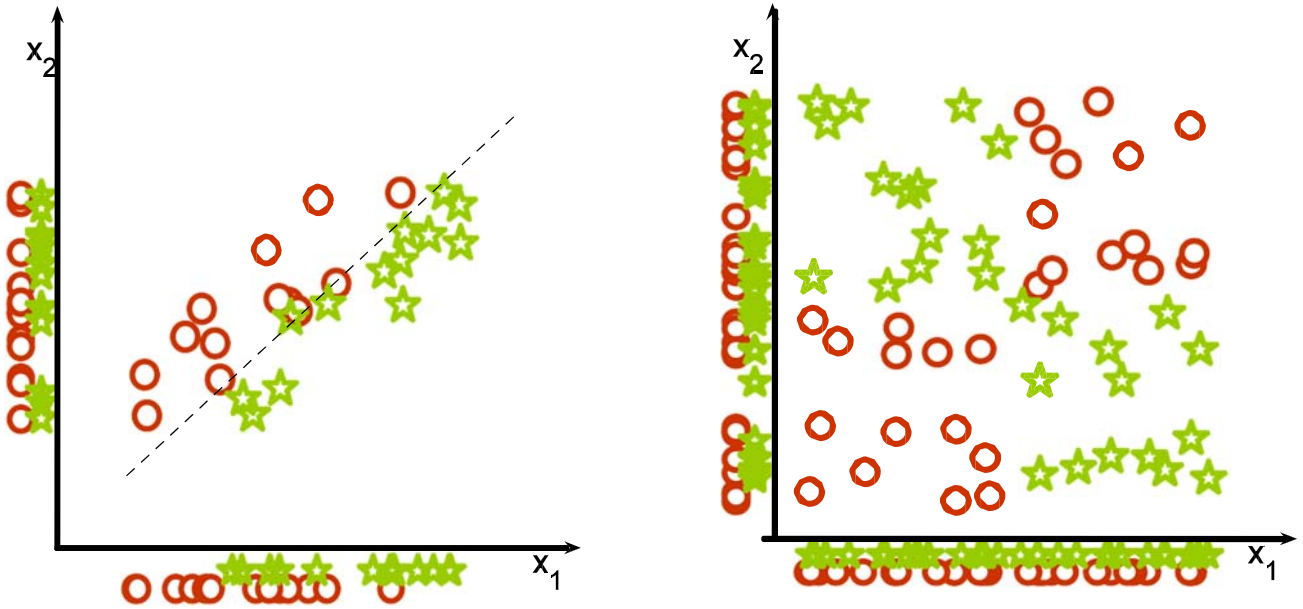
** if (threshold=5), the feature "shoe size" will be selected (8+8)/2 = 8.*

Features Selection

Altres solucions:

Mesures d'entropia (ho veurem quan veiem els arbres de decisió)

Feature selection may fail!!!



Guyon-Elisseeff, JMLR 2004; Springer 2006

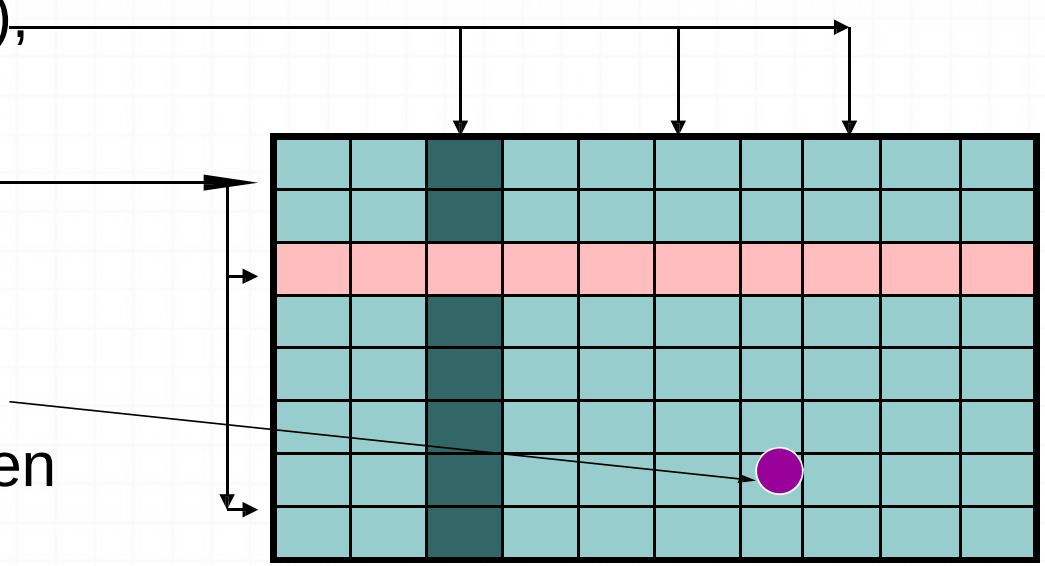
Índex

- Tractament d'atribut
  - Reducció d'atributs (feature selection)
    - Estadistics
    - Filter (Reaper)
    - Wrappers
  - **Extracció (o formació) de característiques**
    - PCA
- Reducció de casos
- Reducció de valors

Dimensions Reduction Of Large Data Sets

Main dimensions:

- **columns** (features),
- **rows** (cases or samples),
- **values** of the features for the given sample



Features Extraction

Principal Components Analysis:

- PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the **greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on..**

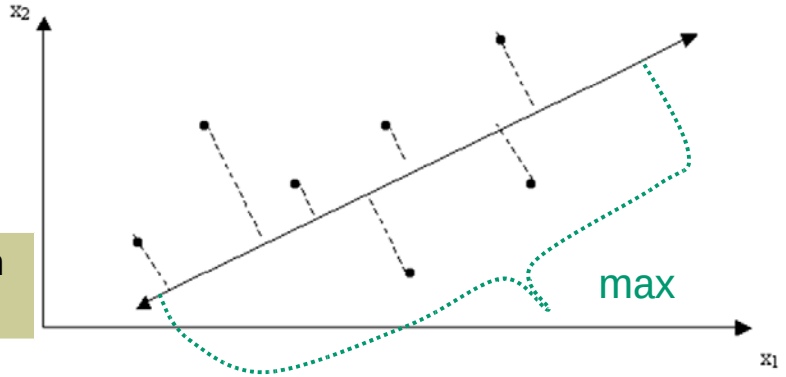
Features Extraction

Principal Components Analysis:

- The features are examined collectively, merged and transformed into a new set of features that hopefully retain the original information content in a reduced form.
- Dot product $a \cdot b$ as projection of point a in normalized vector b
- Given m features, they can be transformed into a single new feature F', by the simple application of weights w:

$$F' = \sum_{j=1}^{m} w(j) \cdot f(j)$$

The first principal component is an axis in the direction of maximum variance.



Features Extraction

Principal Components Analysis:

- Most likely a single set of weights **w(j)** will not be adequate transformation.
- Up to **m** transformations are generated, where each vector of m weights is called a **principal component** and it generate a new feature.
- Eliminating the bottom ranked transformation will cause dimensions reduction.

Features Extraction

Principal Components Analysis Algorithm:

1. We use **covariance matrix S** computation, as a first step in features transformation.

$$S_{n\times n} = \frac{1}{(n-1)}\left[\sum_{j=1}^{n}(x_j - \bar{x}_j)^T(x_j - \bar{x}_j)\right]$$

where $\bar{x}_j$ is the mean of feature j

2. The **eigenvalues** of the covariance matrix S for the given data should be calculated in the next step and the eigenvalues of $S_{n.n}$ are sorted:

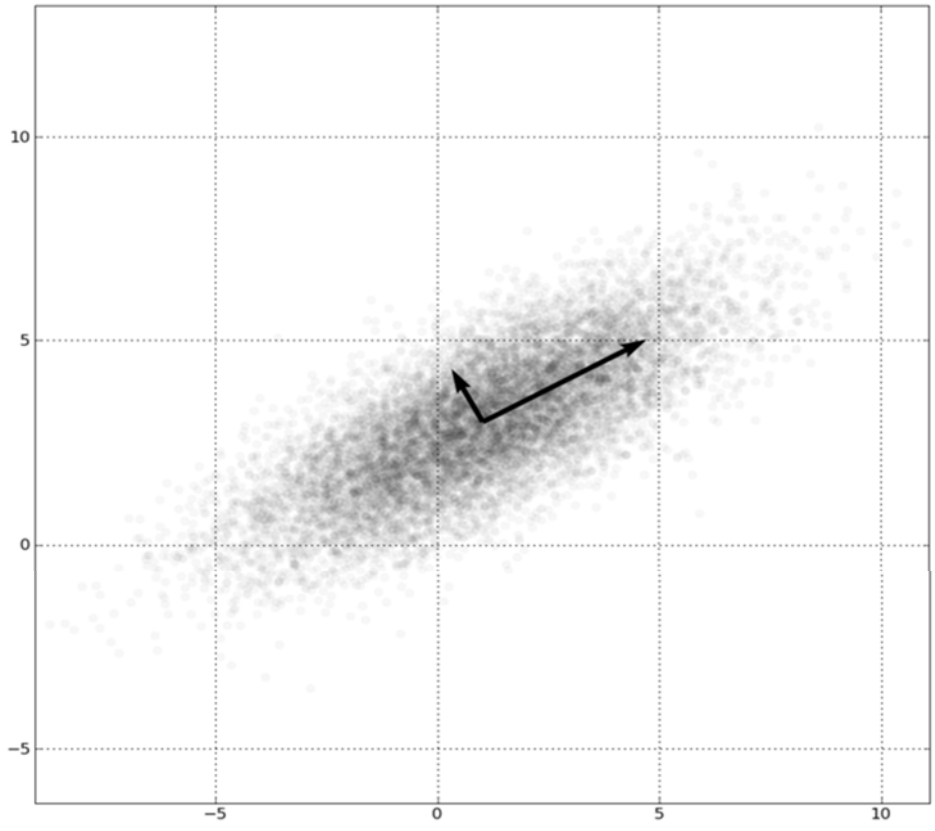
$$\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \text{ where } \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0.$$

Features Extraction

Principal Components Analysis Algorithm:

3. The **eigenvectors e_1, e_2, …, e_n** correspond to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and they are called the **principal axes**.

Eigenvectors for PCA example



Features Extraction

Principal Components Analysis Algorithm:

3. The **eigenvectors e_1, e_2, …, e_n** correspond to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and they are called the **principal axes**.

4. The **criterion** for features selection is based on the ratio **R** of the sum of the m largest eigenvalues of S to the trace of S (for example R>90%):

$$R = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

Features Extraction

Principal Components Analysis – IRIS Data:

	Feature 1	Feature 2	Feature 3	Feature 4
Feature 1	1.0000	-0.1094	0.8718	0.8180
Feature 2	-0.1094	1.0000	-0.4205	-0.3565
Feature 3	0.8718	-0.4205	1.0000	0.9628
Feature 4	0.8180	-0.3565	0.9628	1.0000

The correlation matrix for Iris data

Features	Eigenvalues
Feature 1 *	2.91082
Feature 2 *	0.92122
Feature 3 *	0.14735
Feature 4 *	0.02061

The eigenvalues for Iris data

By setting a threshold value for **R* = 0.95**

R = ((2.91082+0.92122) / (2.91082+0.92122+0.14735+0.02061) = 0.958 > 0.95

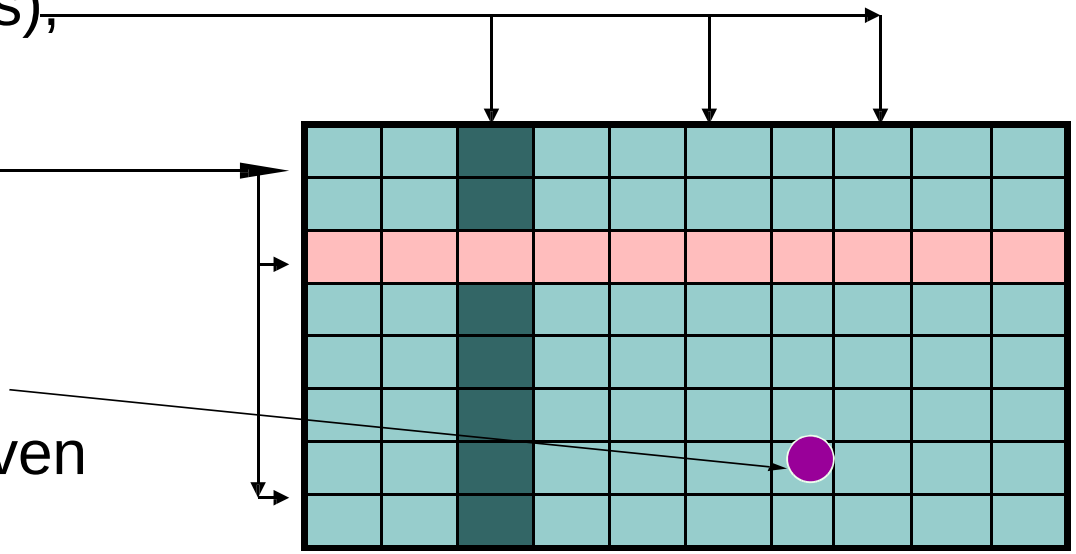
Índex

- Tractament d'atribut
  - Reducció d'atributs (feature selection)
    - Estadistics
    - Filter (Reaper)
    - Wrappers
  - Extracció (o formació) de característiques
    - PCA
- **Reducció de casos**
- Reducció de valors

Dimensions Reduction Of Large Data Sets

Main dimensions:

- **columns** (features),
- **rows** (cases or samples),
- **values** of the features for the given sample



Cases Reduction

Cases are the largest and the most critical dimension.

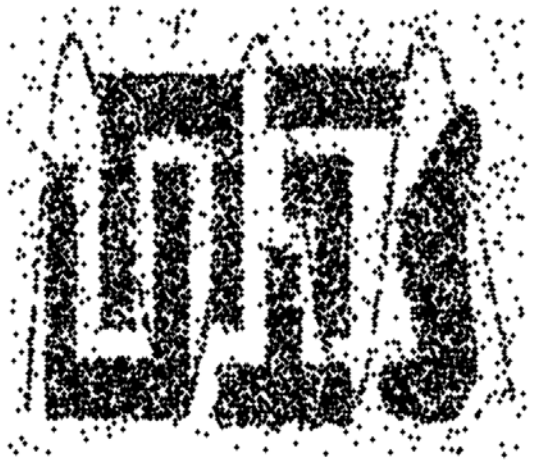
- The most complex task in data reduction.

Sampling is a solution ? **:** After the subset of data is obtained, it is used to estimate information about entire data set.

•Sampling process causes always a **sampling error** - inherent and unavoidable for every approach and every strategy.

•Advantages: reduce cost, greater speed, greater scope, and sometimes even better model.

Cases Reduction: Sample Size



8000 points

2000 Points

500 Points

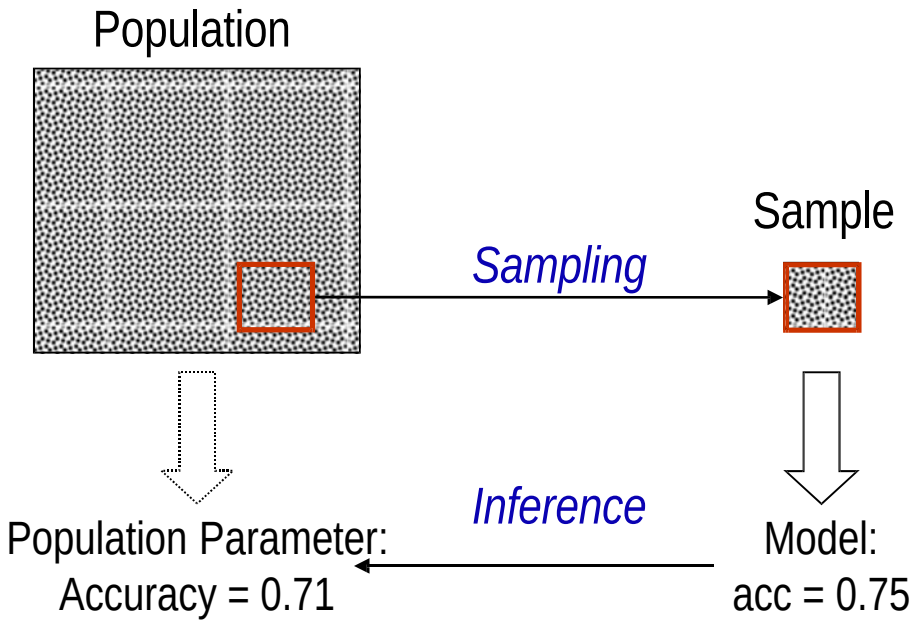
Cases Reduction: Sampling ...

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative.
  - A sample is representative if it has approximately the same property (of interest) as the original set of data.

Cases Reduction:

Accuracy Parameter Estimation

- *Challenging task*: Infer the value of a population parameter based on a sample model.



Cases Reduction

Classifications for sampling methods:

1) general-purpose sampling methods, and
2) sampling methods for specific domains.

General-purpose sampling methods

1. Systematic sampling:

- simplest
- for example 50% of a data set (every second sample)
- built in most of Data Mining tools
- problem: regularities in data set!

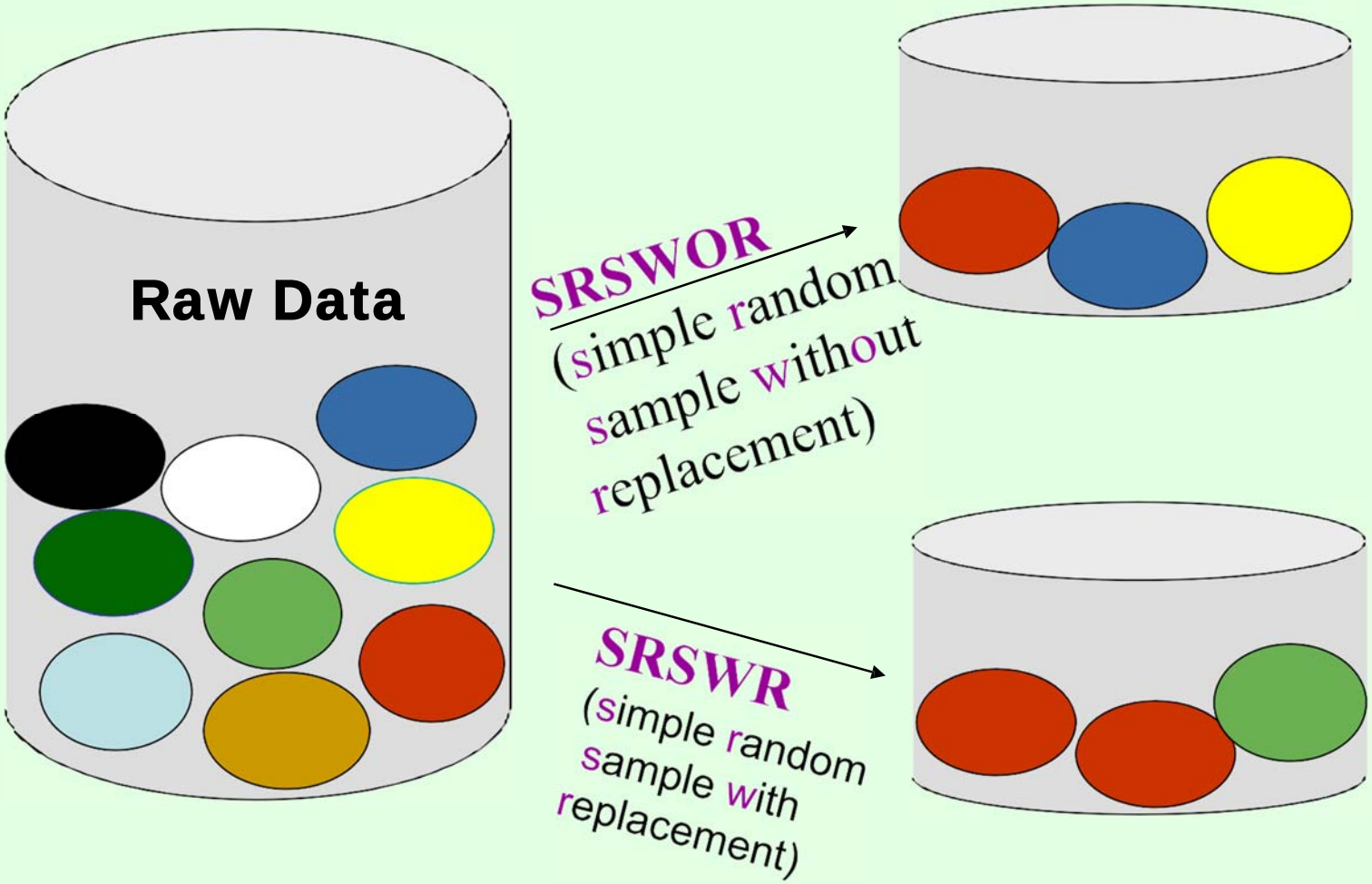
Cases Reduction

General-purpose sampling methods

2. Random sampling

- *random sampling without replacement*,
- *random sampling with replacement*.

Cases Reduction: Random Sampling



Cases Reduction

General-purpose sampling methods

2. *Random sampling*

- *random sampling without replacement*,
- *random sampling with replacement*.

Random Sampling is a process!

2.1 Incremental sampling:

Subsets :10%, 20%, 33%, 50%,..
No improvements in solution –> stop iterations.

Cases Reduction

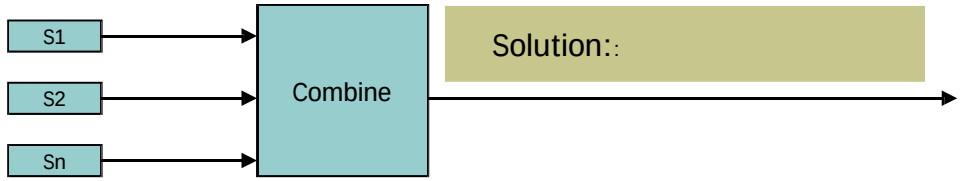
General-purpose sampling methods

2. Random sampling

- *random sampling without replacement*,
- *random sampling with replacement*.

2.2 Average sampling:

Combined solution from several subsets (randomly selected).



Cases Reduction

General-purpose sampling methods

2. Random sampling

- *random sampling without replacement*,
- *random sampling with replacement*.

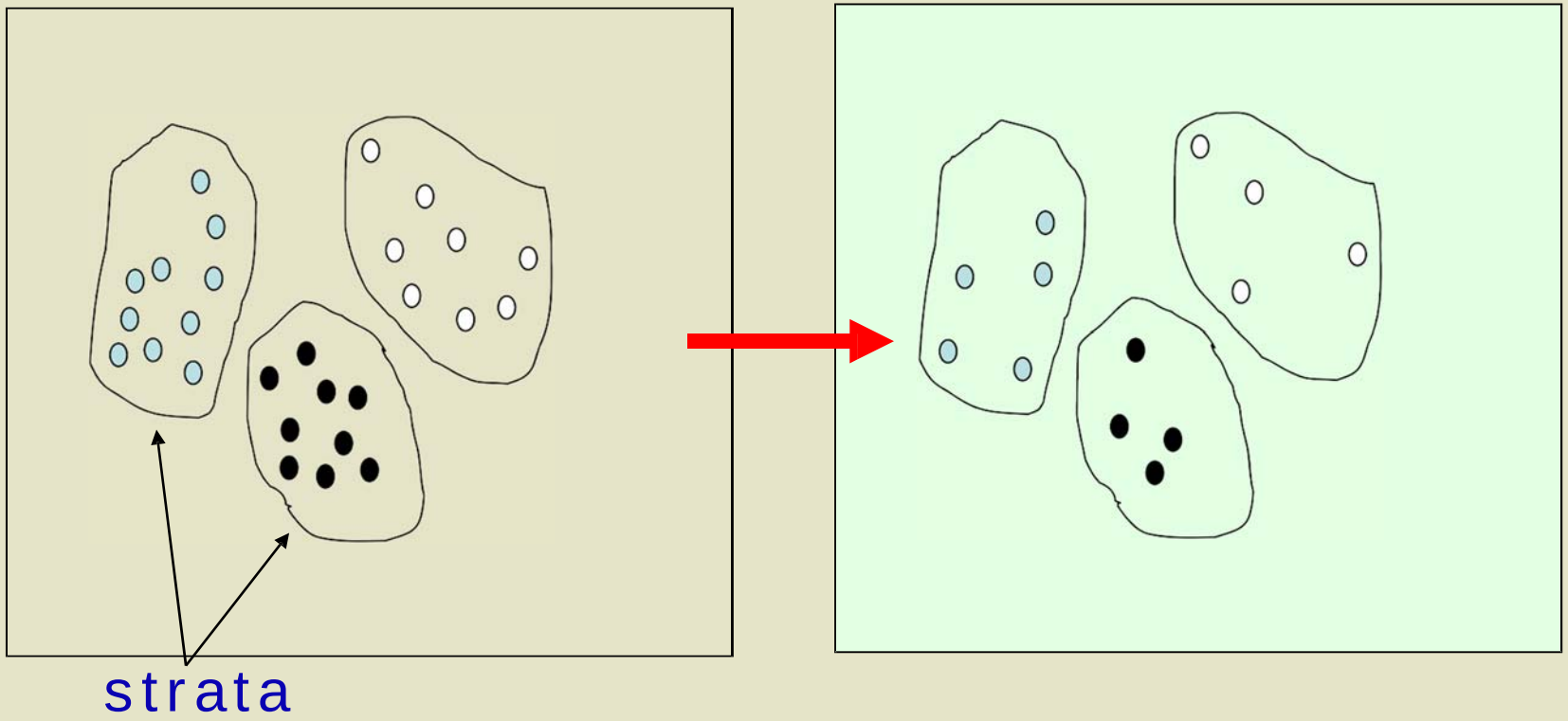
2.3 Stratified sampling:

- Split data set into non-overlapping subsets = strata.
- Combine strata results.

Cases Reduction: Stratified Sampling

Raw Data

Stratified Sample



Cases Reduction

General-purpose sampling methods

2. Random sampling

- *random sampling without replacement*,
- *random sampling with replacement*.

2.4 Inverse Sampling:

- For features with rare frequencies of values (skewed distributions).
- Dynamic increase of subsets until some condition about feature are satisfied.

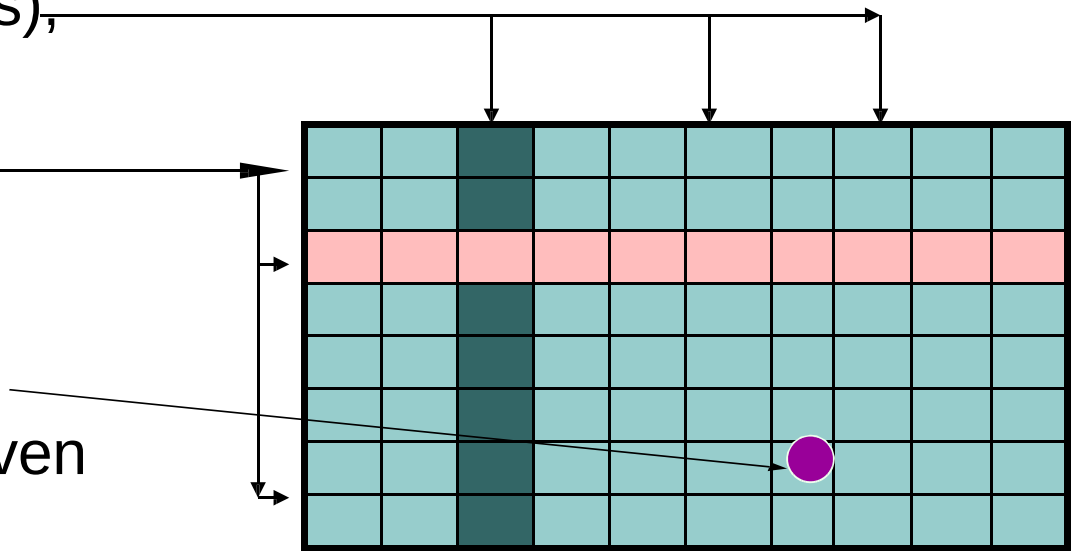
Índex

- Tractament d'atribut
  - Reducció d'atributs (feature selection)
    - Estadistics
    - Filter (Reaper)
    - Wrappers
  - Extracció (o formació) de característiques
    - PCA
- Reducció de casos
- **Reducció de valors**

Dimensions Reduction Of Large Data Sets

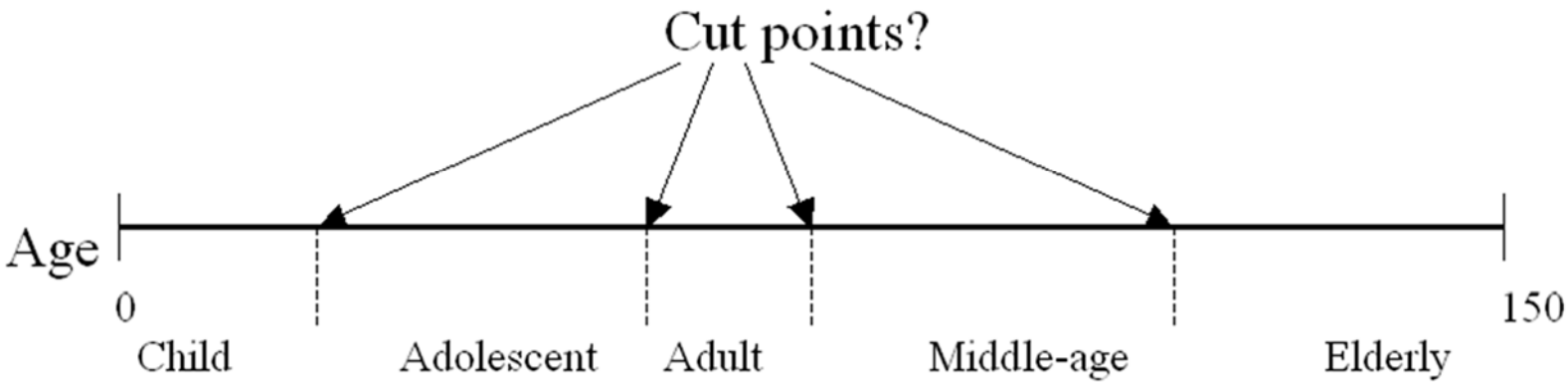
Main dimensions:

- **columns** (features),
- **rows** (cases or samples),
- **values** of the features for the given sample



Values Reduction

Feature discretization techniques:



What are the cut points?

How to select representatives of intervals?

Values Reduction

1. Bins Method

a) Sort all values v_i for a given feature.

b) Assign approximately equal numbers of adjacent sorted values v_i to each **bin**, where the number of bins **k** is given in advance.

c) Move a border element v_i from one bin to the next (or previous) to minimize the global distance error **ER:**

$$ER = \sum_{n=1}^{k} \left(\sum_{i \in Bin_k} \left| v_{i,n} - v_{\text{moda},n} \right| \right)$$

Values Reduction

1.Bins Method Example

f:{ 5, 1, 8, 2, 2, 9, 2, 1, 8, 6} , (k = 3), bins will be represented by its modes.

a) Sorted set of values for feature f is:

{1, 1, 2, 2, 2, 5, 6, 8, 8, 9}

b) Initial bins (k=3) are:

{1, 1, 2, 2, 2, 5, 6, 8, 8, 9}
BIN_1 BIN_2 BIN_3

c1) Modes for three selected bins are: {1, 2, 8}. The total error, using absolute distance for modes, is:

ERR = 0+0+1+0+0+3+2+0+0+1 = 7.

.
.
.

c4) After moving two elements from BIN_2 into BIN_1, and one element from BIN_3 to BIN_2, the new and final distribution of elements will be:

f = {1, 1, 2, 2, 2, 5, 6, 8, 8, 9}
Final bins BIN1 BIN2 BIN3

Final modes are: {2, 5, 8}, and the total minimized error is: ER = 4., and transformed set of values is:

f* = {2, 2, 2, 2, 2, 5, 5, 8, 8, 8} **or non-sorted** : { 5, 2, 8, 2, 2, 8, 2, 2, 8, 5}.

Values Reduction

2. Number Approximation by Rounding

1. Integer division: $Y = int\ (X/10^k)$

2. Rounding: *If* $(mod\ (X,10^k) >= (10^k/2))$ *then* $Y=Y+1$

3. Integer multiplication: $X = Y * 10^k$

where k is the number of rightmost decimal places to round.

For example,	1453 is	rounded	to	1450	with k=1,
		rounded	to	1500	with k=2, and
		rounded	to	1000	with k=3.

Values Reduction

3. ChiMerge Technique

(for classification problems only!!!)

1. **Sort** the data for the given feature in ascending order;

2. **Define initial intervals** so that every value of the feature is in a separate interval.

3. **Repeat until** no χ^2 test of any two adjacent intervals is less then threshold value:

    1. Compute χ^2 tests for each pair of adjacent intervals.
    2. Merge two adjacent intervals with the lowest χ^2 value, if calculated χ^2 is less then threshold.

Values Reduction

A ChiMerge requires computation of χ^2 **test** for the contingency table 2x2 of categorical data:

	Class1	Class2	Σ
Interval-1	A_{11}	A_{12}	R_1
Interval-2	A_{21}	A_{22}	R_2
Σ	C_1	C_2	N

χ^2 test is:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- **k = number of classes,**
- **A_{ij} = number of instances in the i-th interval, j-th class,**
- **E_{ij} = expected frequency of A_{ij}, which is computed as (R_i C_j) / N,**
- **R_i = number of instances in the i-th interval = Σ A_{ij} , j = 1,…k,**
- **C_j = number of instances in the j-th class = Σ A_{ij} , i = 1,2,**
- **N = total number of instances = Σ R_i , i = 1,2.**

Values Reduction

3. ChiMerge Technique – An Example

Data Set

Sample:	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

Initial interval points

Values Reduction

3. ChiMerge Technique - An Example

χ^2 was minimum for intervals: [7.5, 8.5] and [8.5, 10]

Sample:	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

	K=1	K=2	Σ
Interval [7.5, 8.5]	$A_{11}=1$	$A_{12}=0$	$R_1=1$
Interval [8.5, 10]	$A_{21}=1$	$A_{22}=0$	$R_2=1$
Σ	$C_1=2$	$C_2=0$	N=2

Based on the table's values, we can calculate expected values:

E11 = 2/2 = 1, **E12 = 0/2** H **0.1, E22 = 0/2** H **0.1**

E21 = 2/2 = 1, and

and corresponding χ^2 test:

χ^2 **= (1–1)2 /1 + (0–0.1)2 / 0.1 + (1–1)2 / 1 + (0–0.1)2 / 0.1 = 0.2**

For the degree of freedom d=1, and χ^2= 0.2 < 2.706 (**MERGE !)**

Values Reduction

Sample:	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

3. ChiMerge Technique–An Example

.
.
.
One of the additional Iterations:

	K=1	K=2	Σ
Interval [0, 7.5]	$A_{11}=2$	$A_{12}=1$	$R_1=3$
Interval [7.5, 10]	$A_{21}=2$	$A_{22}=0$	$R_2=2$
Σ	$C_1=4$	$C_2=1$	N=5

E11 = 12/5 = 2.4, **E12 = 3/5 = 0.6,**
E21 = 8/5 = 1.6, and **E22 = 2/5 = 0.4**

χ^2 **= (2–2.4)2 / 2.4 + (1–0.6)2 / 0.6 + (2–1.6)2 / 1.6 + (0–0.4)2 / 0.4**

χ^2 **= 0.834**

For the degree of freedom d=1, χ^2 = 0.834 < 2.706 **(MERGE!)**

Values Reduction

Sample:	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

3. ChiMerge Technique - An Example

.
.
.
One of the additional Iterations:

	K=1	K=2	Σ
Interval [0, 10.0]	$A_{11}=4$	$A_{12}=1$	$R_1=5$
Interval [10.0, 42.0]	$A_{21}=1$	$A_{22}=3$	$R_2=4$
Σ	$C_1=5$	$C_2=4$	N=9

E11 = 2.78, E12 =2.22, E21 = 2.22, E22 = 1.78,

χ^2 = 2.72 > 2.706 **(NO MERGE !)**

Final discretization: [0, 10], [10, 42], and [42, 60]

Interval representatives: 5 (low) 26 (medium) 51 (high)

Values Reduction

3. ChiMerge Technique An Example

Final data set with reduced set of values for the future F:

Sample:	F	K
1	5	1
2	5	2
3	5	1
4	5	1
5	5	1
6	26	2
7	26	2
8	26	1
9	26	2
10	51	1
11	51	1
12	51	1