Machine Learning
Data preparation

1

Incomplete data

- **Elimination** a col or a row
- **Inspection** someone need to check why data is missed and fix it.
- **Identification** figuring out the missing data kinda replacement by a reason like IMEI example.
- **Replacement**
  - **mean value of the attribute**
  - **mean value of the attribute for the same target class**
  - **value estimated using statistical inference**

less recommanded.

at the end we are supposed to have a dataset without missing value.

2

Anomaly Detection:

Noisy data

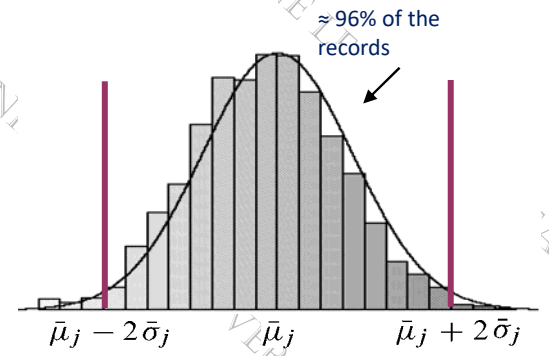
outliers value

only numerical values, not categorical.

if we take 95% of dispresion or 2 3 times of variance is outlier. observations out of this intervals are supecious to be outlier. need to be barresi bishter.

DISPERSION $\bar{\mu}_j \pm z_{\alpha/2}\bar{\sigma}_j$ **APPLY ONLY IF DISTRIBUTION IS NORMAL**

quantile



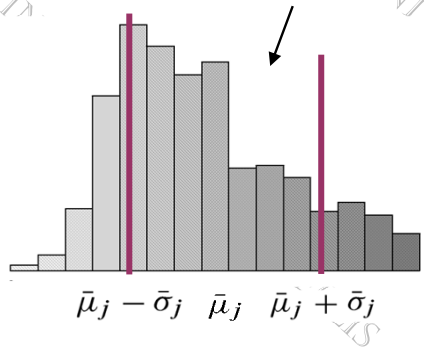
3

Noisy data

a percentage at least $1-1/\gamma^2$ **of the observations, with** $\gamma > 1$**, falls in the interval** $(\bar{\mu} \pm \gamma\bar{\sigma})$**(** **(TCHEBYSHEFF theorem)**

True for all types of distribution

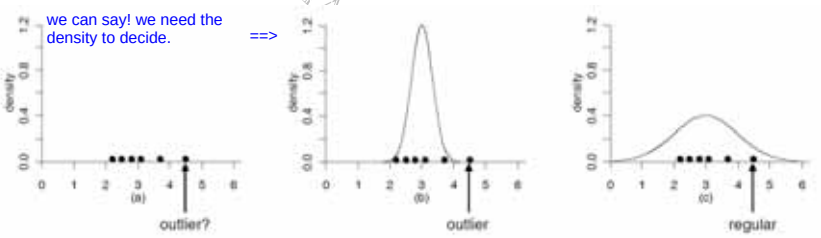
at least ($1-1/\gamma^2$) % of the observations



4

Z-index method for outlier detection

we can say! we need the density to decide. ==>



$$z_{ij}^{ind} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j}$$

sample mean

sample std

it's outlier if :

$$\left| z_i^{ind} \right| > 3$$

$$\left| z_i^{ind} \right| \gg 3$$

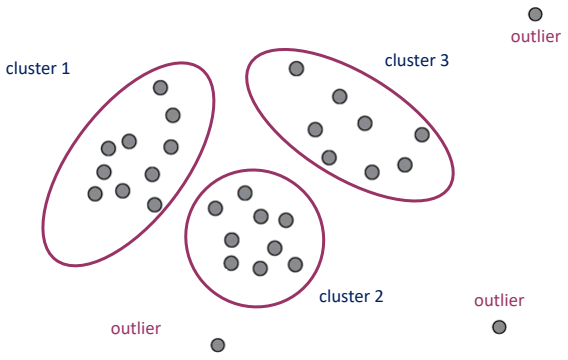
5

Multidi-mentional Outlier Detection

Noisy data

if we take the inequatliies for each of dimentions we'll see no outliers, but if we check them together, we see these are outliers.

Applying Clustering helps us finding them.

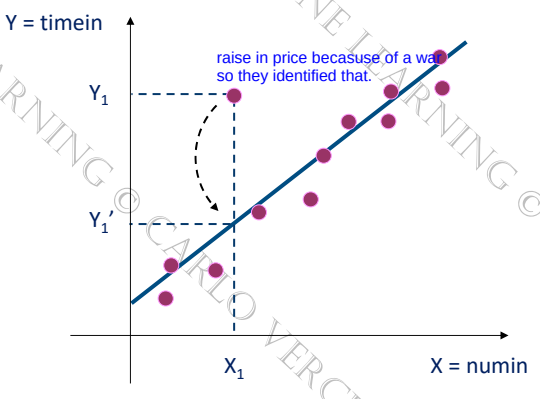


evaluate "density" in a neighborhood of a record: a record is an outlier if a percentage *p* of the records falls at a distance greater than *d*

6

Noisy data

identification example:



7

missing
anomaly
transformation:

Data transformation

Simplifing the learning process for algorithms.

DECIMAL SCALING

$$x'_{ij} = \frac{x_{ij}}{10^h}$$

MIN-MAX method Preffered

$$x'_{ij} = \frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}}(x'_{max,j} - x'_{min,j}) + x'_{min,j}$$

Z-INDEX method

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j}$$

8

Data standardization example

numout	decimal scaling	min-max method	z-index method
45	0.045	0.295	-0.006
20	0.02	0.078	-0.023
69	0.069	0.504	-0.01
66	0.066	0.478	0.008
11	0.011	0	-0.029
42	0.042	0.269	-0.008
126	0.126	1	0.049
original values			

9

Data reduction

can be in both col and rows

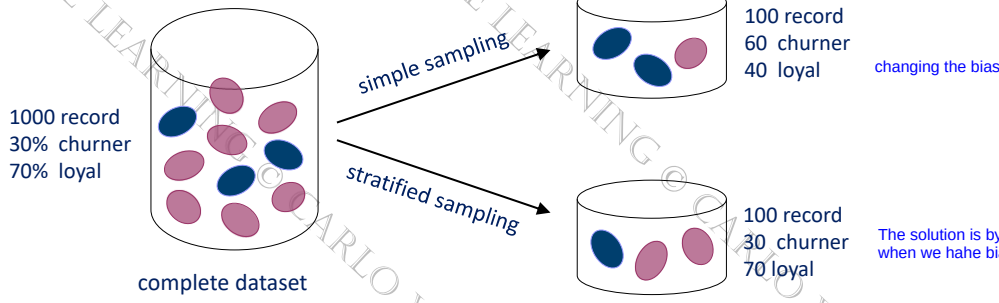
- improve efficiency in model identification in slow algorithm or heavy computing algorithm
- preserve model accuracy
- achieve simpler models

- record reduction by SAMPLING
- attribute reduction by SELECTION or PROJECTION
- values reduction by DISCRETIZATION or AGGREGATION

10

Dataset reduction

Sampling Rows



changing the biases in data even may happen

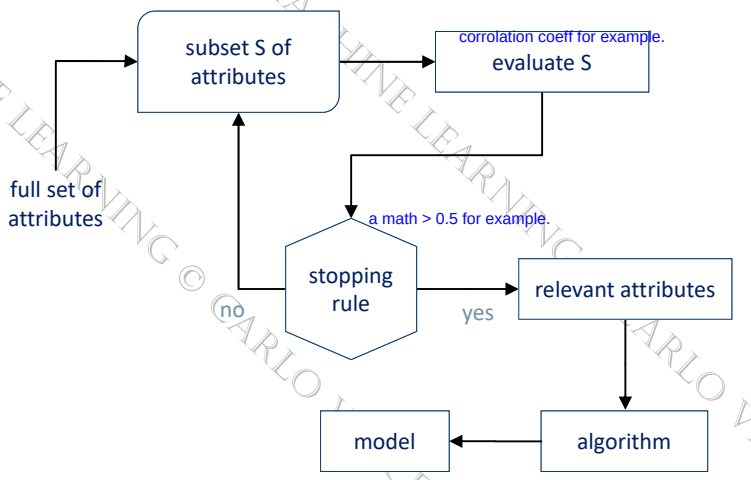
The solution is by keeping the target proportion when we hahe biases in target /

11

Attribute selection: filter methods

algorithm independent very fast, but potentially loss in accuracy.

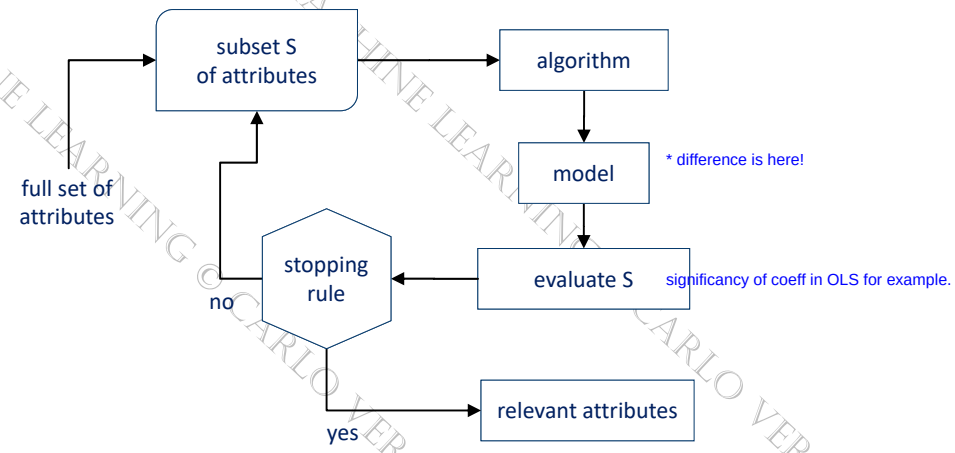
keeping some variables only.



12

Attribute selection: wrapper methods

Forward Selection
Backward Selection

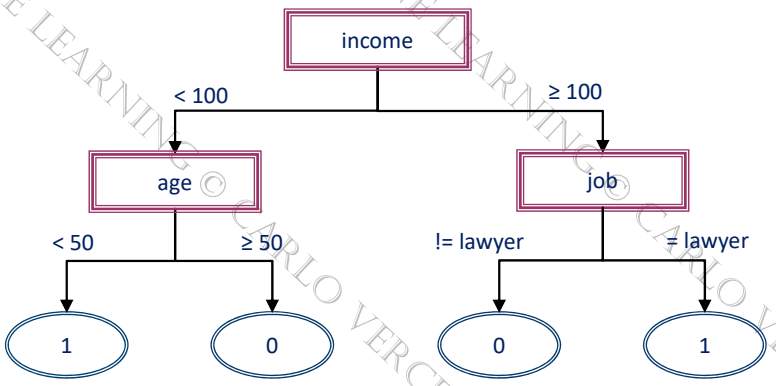


13

Attribute selection: embedded methods

نفهمیدم چیکار کرد ... ولی گویا خود مدل انجام میده این کار رو و فقط چندتا اتریبیوت رو انتخاب میکنه.

Set of attributes: {age, sex, education, income, job}
Target: life insurance policy (class = 1)



14

Are we sure that these attributes are the best? NO

Attribute selection

an NPR problem. exponentially difficult.

Given a dataset with *n* attributes, how many subsets to evaluate? ….2^n

subset S of attributes

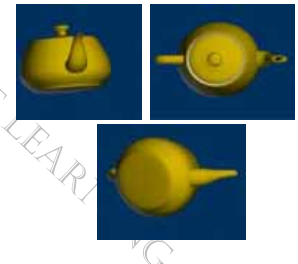
full set of attributes

greedy search scheme:

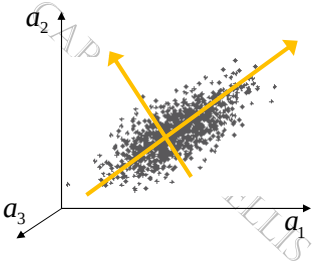
- FORWARD
- BACKWARD
- FORWARD-BACKWARD

Dimentionality reduction

Principal component analysis



Principal component analysis



17

Principal component analysis

- Principal component analysis (PCA) is the oldest and most popular technique of attribute reduction by means of projection.
- The n original attributes are replaced by n new attributes (principal components) obtained as a linear combination of the original ones.
- Principal components are generated in sequence by means of an iterative algorithm:
  - the first component is determined by solving an appropriate optimization problem, in order to explain the highest percentage of variation in the data.
  - at each iteration the next principal component is selected, among vectors orthogonal to components already determined, as the one which explains the maximum percentage of variance not yet explained by the previously generated components.
- In most cases, q principal components, with $q < n$, have information content almost equivalent to the original attributes.

18

Principal component analysis

$\mathbf{V} = \mathbf{X}'\mathbf{X}$ COVARIANCE matrix of data, after standardization: $\bar{x}_{ij} = x_{ij} - \frac{1}{m}\sum_{i=1}^{m} x_{ij}$.

The principal components are obtained as linear combinations of the original variables:

$$\mathbf{p}_j = \mathbf{X}\mathbf{w}_j$$

where the coefficients $\mathbf{w}_j$ have to be determined.

The variance of the projection $\mathbf{w}_j'\mathbf{x}_i$ can be calculated as:

$$\begin{aligned} \mathrm{E}[\mathbf{w}_j'\mathbf{x}_i - \mathrm{E}[\mathbf{w}_j'\mathbf{x}_i]]^2 &= \mathrm{E}[(\mathbf{w}_j'(\mathbf{x}_i - \mathrm{E}[\mathbf{x}_i]))^2] \\ &= \mathbf{w}_j' \mathrm{E}[(\mathbf{x}_i - \mathrm{E}[\mathbf{x}_i])'(\mathbf{x}_i - \mathrm{E}[\mathbf{x}_i])]\mathbf{w}_j \\ &= \mathbf{w}_j' \mathbf{V} \mathbf{w}_j . \end{aligned}$$

Principal component analysis

The FIRST PRINCIPAL COMPONENT $\mathbf{p}_j = \mathbf{X}\mathbf{w}_j$ is obtained by solving an optimization problem

$$\max_{\mathbf{w}_1} \quad \{ \mathbf{w}_1'\mathbf{V}\mathbf{w}_1 : \mathbf{w}_1'\mathbf{w}_1 = 1 \}$$

It can be shown that:

the vector that solves the problem is the eigenvector $\mathbf{u}_1$ associated to the largest eigenvalue λ_1 of the covariance matrix $\mathbf{V}$

the eigenvalue λ_1 is equal to the amount of explained variance

The SECOND PRINCIPAL COMPONENT is obtained by solving a similar optimization problem with the additional condition

$$\mathbf{w}_2'\mathbf{w}_1 = 0$$

X outside of the exam scope

Principal component analysis

$$\max_{\mathbf{w}_1} \{ \mathbf{w}_1' \mathbf{V} \mathbf{w}_1 : \mathbf{w}_1' \mathbf{w}_1 = 1 \}$$

Lagrangean function

$$L(\mathbf{w}_1, \lambda_1) = \mathbf{w}_1' \mathbf{V} \mathbf{w}_1 - \lambda_1 (\mathbf{w}_1' \mathbf{w}_1 - 1)$$

KKT conditions

$$\frac{\partial L(\mathbf{w}_1, \lambda_1)}{\partial \mathbf{w}_1} = 2\mathbf{V}\mathbf{w}_1 - 2\lambda_1 \mathbf{w}_1 = \mathbf{0}.$$

$$\frac{\partial L(\mathbf{w}_1, \lambda_1)}{\partial \lambda_1} = 1 - \mathbf{w}_1' \mathbf{w}_1 = 0.$$

$$\Rightarrow \quad (\mathbf{V} - \lambda_1 \mathbf{I})\mathbf{w}_1 = \mathbf{0}.$$

This implies

$$\mathbf{w}_1' \mathbf{V} \mathbf{w}_1 = \lambda_1 \mathbf{w}_1' \mathbf{w}_1 = \lambda_1 \mathbf{u}_1' \mathbf{u}_1 = \lambda_1$$

$$\mathbf{p}_j = \mathbf{X}\mathbf{u}_j$$

Principal component analysis

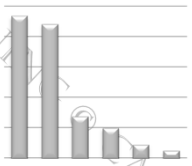
mige mamolan 2ta mishe...let's see

How many principal component to consider?

percentage of explained variance

$$I_q = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}$$

scree-plot



ELBOW method of selection

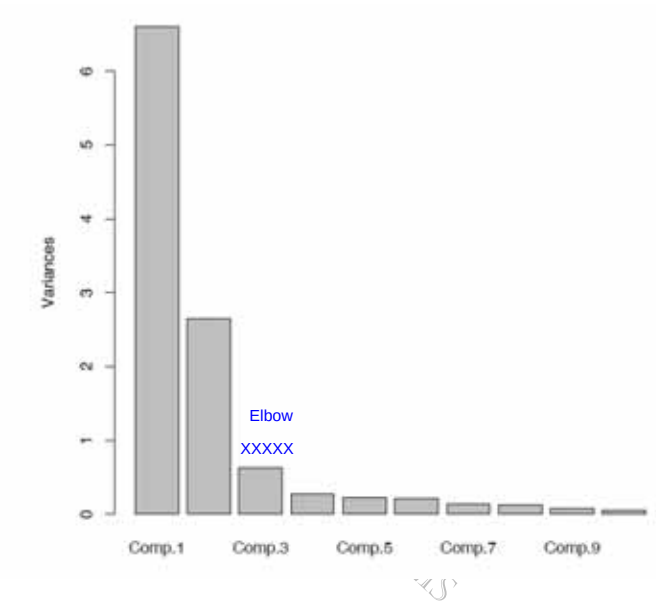
Principal component analysis

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
mpg	0.363	-0.226	-0.103	-0.109	0.368	0.754
cyl	-0.374	-0.175	0.169	0.231	-0.846	
disp	-0.368	0.257	-0.394	-0.336	0.214	0.198
hp	-0.33	-0.249	0.14	-0.54	0.222	-0.576
drat	0.294	-0.275	0.161	0.855	0.244	-0.101
wt	-0.346	0.143	0.342	0.246	-0.465	0.359
qsec	0.2	0.463	0.403	0.165	-0.33	0.232
vs	0.307	0.232	0.429	-0.215	-0.6	0.194
am	0.235	-0.429	-0.206	-0.571	-0.587	-0.178
gear	0.207	-0.462	0.29	-0.265	-0.244	0.605
carb	-0.214	-0.414	0.529	-0.127	0.361	0.184

	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
mpg	0.236	0.139			0.125
cyl					0.141
disp					-0.661
hp	0.248				0.256
drat					
wt					0.567
qsec	-0.528	-0.271			-0.181
vs	-0.266	0.359	-0.159		
am					
gear	-0.336	-0.214			
carb	-0.175	0.396	0.171		-0.32

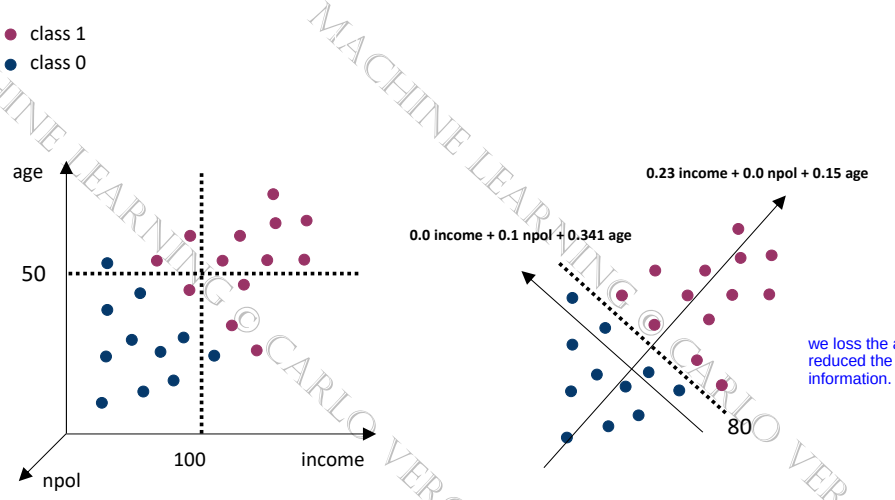
23

Principal component analysis



24

Principal component analysis



we loss the ability to intrpret the coefficients as we reduced the dimentions and we have some loss of information.

25

ML Trees are the most intrepratable algos.

Customer Satisfaction Analysis

in lab

A survey in order to evaluate 20 different healthcare structures. 2000 customers have evaluated, with a 1-10 scale, each of six features of the service:

1.Courtesy
2.Clarity
3.Competence
4.Condition (of the structure)
5.Promptness (of the service)
6.Opening times

The following table shows the average results for each structure and each variable:

Structure	Courtesy	Clarity	Competence	Condition	Promptness	Opening times
A	7	5	9	8	6	7
B	5	6	8	4	4	6
C	5	5	8	7	7	7
D	6	6	9	7	6	7
E	7	5	10	4	3	6
F	6	4	8	4	5	6
G	5	6	9	4	3	5
H	5	5	8	5	4	6
I	4	4	7	7	5	6
L	5	5	8	7	6	7

26

Customer Satisfaction Analysis

	Courtesy	Clarity	Competence	Condition	Promptness	Opening.times
Courtesy	1.0000000	0.54586083	0.694808334	-0.115503469	0.00000000	0.1430140
Clarity	0.5458608	1.00000000	0.544426373	-0.081651741	0.04523081	0.1010995
Competence	0.6948083	0.54442637	1.000000000	0.003778134	-0.05860090	-0.1216282
Condition	-0.1155035	-0.08165174	0.003778134	1.000000000	0.71911387	0.4394498
Promptness	0.0000000	0.04523081	-0.058600904	0.719113873	1.00000000	0.7982825
Opening.times	0.1430140	0.10109950	-0.121628193	0.439449764	0.79828247	1.0000000

Relevant values of correlation among our variables

27

Customer Satisfaction Analysis

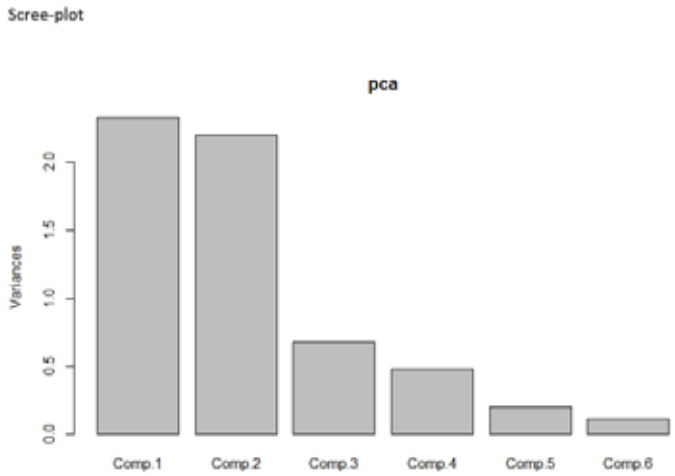
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.5257	1.4824	0.8224	0.69466	0.45322	0.33260
Proportion of Variance	0.3879	0.3662	0.1127	0.08043	0.03423	0.01844
Cumulative Proportion	0.3879	0.7542	0.8669	0.94733	0.98156	1.00000

The first two components account for about 75% of the variance.

28

Customer Satisfaction Analysis



29

Customer Satisfaction Analysis

$$Y_1 = 0.1086X_1 + 0.0830X_2 + 0.1663X_3 - 0.5278X_4 - 0.6168X_5 - 0.5428X_6$$

$$Y_2 = 0.5869X_1 + 0.5389X_2 + 0.5621X_3 - 0.0515X_4 - 0.1356X_5 - 0.1678X_6$$

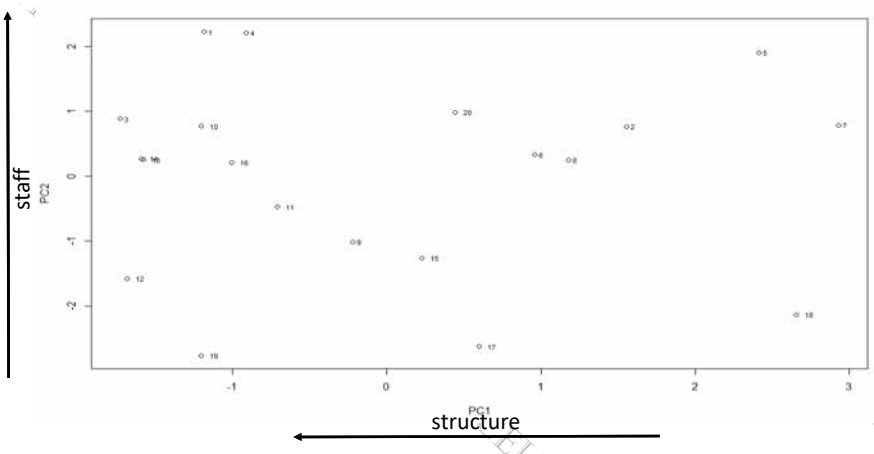
30

Customer Satisfaction Analysis

	PC1	PC2	PC3	PC4	PC5	PC6
[1,]	-1.1828628	2.2263798	-0.22186202	-0.91961937	-0.62306953	-0.028256138
[2,]	1.5546960	0.7654794	0.78344001	0.86684190	0.09347771	0.083846308
[3,]	-1.7240532	0.8856652	0.23333559	0.09592982	0.44168534	-0.549643694
[4,]	-0.9068263	2.2086511	0.27104063	0.28268453	-0.01769527	0.015156363
[5,]	2.4116996	1.9021137	0.13642506	-1.12399068	-0.19231098	0.556696687
[6,]	0.9589090	0.3290325	0.41935059	-1.18921989	0.09996898	-0.744915764
[7,]	2.9310884	0.7865377	-0.26154565	1.01955216	0.01685012	0.044668363
[8,]	1.1803912	0.2520779	0.17638250	0.15112222	-0.01811239	0.224943955
[9,]	-0.2186553	-1.0141899	-0.57707610	0.12265701	-0.04730614	-0.032665186
[10,]	-1.1967878	0.7697211	0.21721841	0.07567853	0.20680405	0.069612607
[11,]	-0.7035076	-0.4727002	1.47351814	0.28010514	-0.98973584	0.182605173
[12,]	-1.6784426	-1.5770088	0.22172883	-0.82491768	0.16201852	-0.005985052
[13,]	-1.5710926	0.2563197	-0.38783911	-0.64004115	0.09521395	0.210710254
[14,]	-1.5859659	0.2669501	-0.25381324	0.64140750	0.45630104	0.407718213
[15,]	0.2299627	-1.2612263	0.99588496	0.96178037	-0.49539272	-0.562470476
[16,]	-1.0041463	0.2102320	0.46394032	0.48311800	0.94487539	0.100527297
[17,]	0.6004681	-2.6257821	0.07073805	-0.26288977	0.43088624	0.333856443
[18,]	2.6569484	-2.1303118	-0.14433919	-0.75505323	0.34904884	-0.197022566
[19,]	-1.1987349	-2.7624412	-1.04413849	0.17521472	-0.71336311	0.211425882
[20,]	0.4469118	0.9845000	-2.57438929	0.55963983	-0.20014418	-0.320808668

31

Customer Satisfaction Analysis



32

Data discretization

For numerical attributes ⇒ DISCRETIZATION

introduce classes of spending!
like address => to postal codes

SUBJECTIVE SUBDIVISION

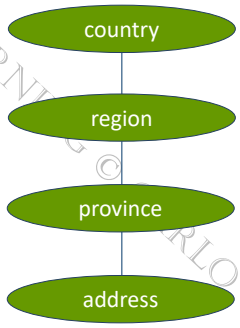
SUBDIVISION INTO CLASSES
of equal size or equal width

numin	
229	low
235	
237	
240	medium
245	
248	
253	high
360	
271	

33

Hierarchical discretization

For a categorical attribute



15 distinct values

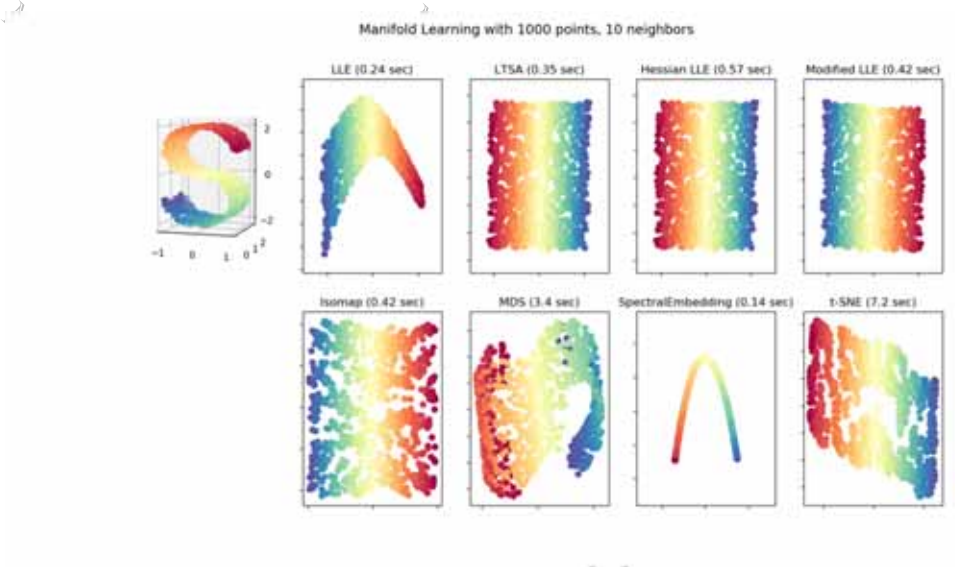
65 distinct values

3567 distinct values

674.339 distinct values

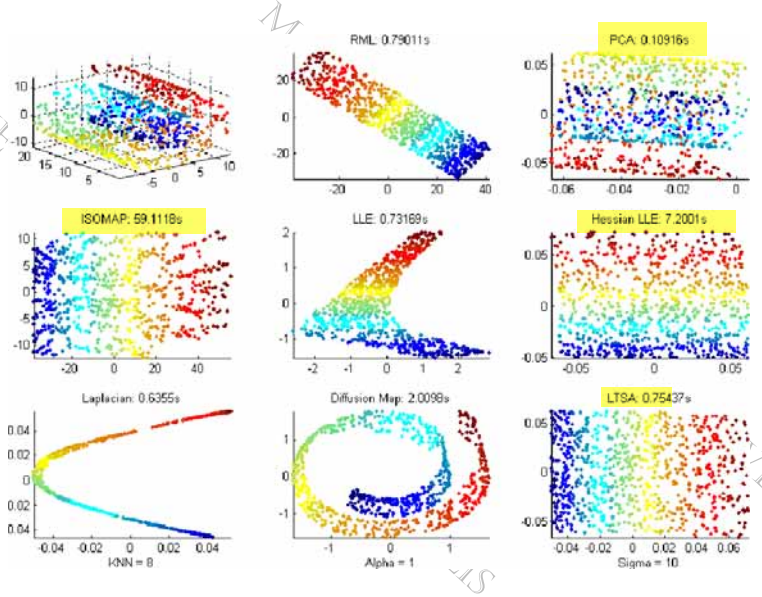
34

Nonlinear dimensionality reduction - embedding



35

Nonlinear dimensionality reduction - embedding



local linear embeding

جملات اخر درباره استفاده از پی سی ای برای دیتاها رو ببین.

36