Contents

1 Curs 5. Tipuri de date. descriptive

In [17]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

print(f'NumPy version: {np.__version__}
print(f'pandas version: {pd.__version__
```

executed in 5.42s, finished 19:45:20 2021-03-21

```
NumPy version: 1.19.2
pandas version: 1.2.3
```

1.1 Tipuri de date

- Un set de date este o colectie de obiecte-data
- Sinonime pentru obiecte-data: înregistrare, pu confuzie), observatie, entitate.
- Obiectele sunt descrise prin atribute
- Sinonime pentru atribut: **variabila**, **caracteris**

Definitie: Atribut: proprietate sau caracteristica a u

- Exemple: culoarea ochilor, temperatura.
- Trebuie facuta diferenta între proprietatile atrib
  - acelasi atribut poate avea valori diferite: î
  - diferite atribute pot fi masurate cu acelasi reprezentate prin numere întregi; în timp

înaltime (medie, max etc.) nu se aplica si

1.1.1 Tipuri de atribute

Exista diferite tipuri de atribute:

- Categoriale (calitative)
  - **nominale**: valori diferite care permit recu
  - **ordinale**: valorile permit ordonarea obiec operatii permise: =, $\neq$, <, >; functii aplicab
- Numerice (cantitative)
  - **interval**: se poate face diferenta între val Fahrenheit; pe lânga operatiile de mai sus
  - **multiplicabile**: permit împartiri si înmultir mai sus si inmultirea, impartirea; functii ap

1.1.2 Transformari la nivel de atribut

Exista niste transformari care se pot efectua sau n

- pentru atribute nominale: orice asociere unu-la datelor;
- pentru atribute ordinale: orice modificare de va monoton crescatoare; {bun, mai bun, cel mai b
- pentru atribute interval: transformari de forma
- pentru atribute multiplicabile: $val\_veche/val\_$

1.1.3 Descrierea atributelor prin nun

- Atribute discrete:
  - o multime cel mult numarabila de valori;
  - exemple: coduri postale, cuvinte într-un d
  - se reprezinta cel mai frecvent ca numere
  - caz special — atribute binare: {prezent, a
- Atribute continue:
  - valorile sunt exprimate prin numere reale
  - exemple: temperatura, masa
  - dpdv practic reprezentarea se face cu o p
  - reprezentare actuala: valori în virgula mol
- Valori asimetrice:
  - doar prezenta unei trasaturi (i.e. valoare r
  - exemple: vectorul care reprezinta daca ni
  - daca se iau in considerare doi astfel de v simultan

Contents

1.1.4 Tipuri de seturi de date

- Seturi de date de tip: înregistrare, de tip graf s
- Caracteteristici generale:
  - **dimensionalitatea** = numarul de atribute multe dimensiuni se pot aplica tehnici de
  - **caracterul rarefiat al datelor** = procentu poate reduce drastic necesarul de memo
  - **rezolutia** = scara la care se face raportar zile pot arata iminenta unei furtuni, dar la

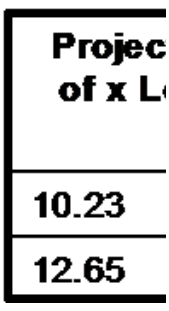
1.1.4.1 Seturi de date de tip înregistrare

- cel mai des furnizate si frecvent utilizate in apl
- nu exista legatura intre inregistrari distincte
- stocare: fisiere text (e.g. CSV), Excel, baze de

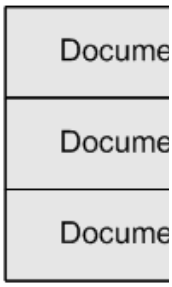
Cazuri remarcabile de seturi de date inregistrare:

- Tranzactii, date specifice cosurilor de cumpara
  - exemplu: intr-un magazin, setul de produ
  - se analizeaza asocierea intre produsele i
  - posibilitate de reprezentare: indicator boo
  - variatie: cate exemplare din produs au fo

- Matrice de date:
  - pentru cazul in care datele au acelasi set
  - fiecare data in parte poate fi considerata
  - fiecare atribut considerat este o dimensiu
  - este tipul de date standard pentru analiza
  - nota: intre conceptul de dimensiune asa



* Matrice de date rarefiate: * caz special a
trasaturilor * exemplu: tranzactii din cosur
dintr-un anumit lexic; pentru un document
sau nu = matricea document–termen * in
(https://docs.scipy.org/doc/scipy/reference



1.1.4.2 Seturi de date de tip graf

- reprezentare convenabila pentru cazurile:
  - graful reprezinta relatii între obiecte
  - obiectele însele sunt reprezentate ca graf

Caz 1: datele reprezinta relatii între obi

- obiectele sunt reprezentate ca noduri în graf
- relatiile dintre obiecte sunt reprezentate sub fo
- exemplu: pagini web care contin legaturi catre
- exemplu de algoritm ce foloseste structura de

Contents

Caz 2: obiectele-data sunt grafuri

- obiectele pot contine subobiecte care sunt leg
- uneori nu doar legaturile sunt importante, ci si
- exemplu: formulele chimice - benzen = C_6H_6
- utilitate: se poate detecta care substructura ap anumitor proprietati chimice.
- domeniu aparte: “mineritul” substructurilor

1.1.4.3 Seturi de tip secventa

- atributele au relatii care implica ordonare în tir
- subtipuri: date secventiale, secventa, serii de

Caz: date secventiale

- numite si date temporale
- fiecare înregistrare are un atribut suplimentar

Contents

Caz: date secventa

- setul de date entitati individuale, precum secv
- similare cu cele secventiale, dar fara timp incl
- pozitia din secventa este importanta
- exemplu: informatia genetica este o secventa
- aplicatie: predictia similaritatilor în structura si

Caz: serii de timp

- fiecare înregistrare e o serie de timp = o serie
- exemplu: seturi de date de tip financiar, repre
- exemplu: date meteo masurate lunar

Contents

- ***Caz: Date spatiale***
  - cazul datelor care au atribute spatiale sau are
  - exemplu: date climatice raportate pe regiuni
  - exemplu: date adunate pentru scurgerea unui

1.2 Calitatea datelor

1.2.1 Probleme legate de masurarea

Contents

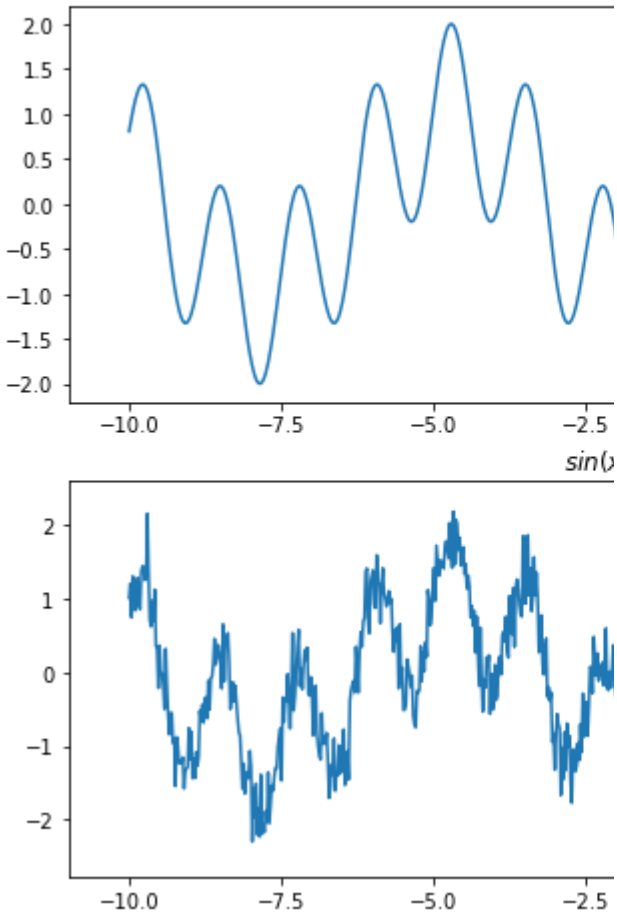
- Presupunerea ca datele pe baza carora se fac
- Prevenirea problemelor care duc la scaderea
- Abordari:
  - detectarea si corectarea erorilor = curatar
  - construirea de algoritmi care sa tolereze c
- surse de probleme în calitatea datelor:
  - procesele de masurare
  - aplicatiile folosite
- Zgomotul
  - componenta aleatoare care se adauga ur
  - Exemplu: distorsiunea vocii unei persoan
  - Daca eroarea apare mereu în acelasi loc:

In [16]:
```
x = np.linspace(-10, 10, 1000)
y = np.sin(x) + np.sin(5*x)

fig, (ax1, ax2) = plt.subplots(2, figsi
fig.suptitle('Date cu si fara zgomot')
ax1.set_title(r'$sin(x) + sin(5x)$')
ax1.plot(x, y)
ax2.set_title(r'$sin(x) + sin(5x) + 0.3
ax2.plot(x, y + 0.3 * np.random.randn(1
```

executed in 447ms, finished 17:33:07 2021-03-21

Da



1.2.2 Calitatea datelor: precizie, abat

- **Precizie**: apropierea valorilor rezultate prin ma
- **Abatere (eng: bias)**: o variatie sistematica a r
- Exemplu: se masoara o cantitate de 1 gram. \
  0.001.
- Precizia este considerata abaterea standard:
  - $\sigma = \sqrt{E[(X - E(X))^2]}$ deci pentru dat
- **Acuratetea**: apropierea masuratorilor fata de

1.2.3 Calitatea datelor: anomalii

- Anomaliile sunt obiecte cu caracteristici consi
- Anomaliile (outliers) nu sunt zgomote, ci obiec
- Utilitate: detectarea de nise pe piata, detectar

1.2.4 Calitatea datelor: valori lipsa

- Cazuri: una sau mai multe valori de atribute lip
- Motive pentru lipsa valorilor:
  - informatia nu este colectata — oamenii nu
  - atributele nu se pot aplica tot timpul tuturo
- Operarea în aceste situatii:
  - eliminarea obiectelor-data sau a atributelo
  - estimarea valorilor lipsa
  - ignorarea valorilor lipsa în timpul analizei

1.2.5 Calitatea datelor: valori incons

- Valori inconsistente:
  - Exemplu: oras si cod postal precizate, da
  - Exemplu: typos, marimi cu valori impropri
  - Operare: detectarea valorilor gresite si co
  - E necesara utilizarea surselor de date rec
- Date duplicate:
  - Duplicarea poate sa fie exacta sau aproa
  - Exemplu: aceeasi persoana cu adrese de
  - Procesul de curatare = deduplicare

1.2.6 Calitatea datelor din perspectiv

- Din perspectiva aplicatiilor, “datele au calitate
- Caracterul oportun al datelor — daca datele s
- Relevanta — in cazul in care se vrea crearea alta situatie este data de esantionarea neadec
- Cunostintele apriori despre date — de exempl reducerea redundantei si a dimensionalitatii; c

1.3 Preprocesarea datelor

- Strategii si tehnici complexe, ce pot cere pâna
- Doua variante:
  1. selectara obiectelor-data si a atributelor
  2. crearea/schimbarea de atribute
- Variante de preprocesare:
  - agregare
  - esantionare
  - reducerea dimensionalitatii
  - selectarea unui subset de atribute
  - crearea de atribute
  - discretizare si binarizare
  - transformarea variabilelor

1.3.1 Agregare

- Scop: combinarea a doua sau mai multe atribu
- Utilitate:
  - reducerea cantitatii de date
  - schimbarea scalei: orasele sunt agregate
  - date mai stabile: datele agregate au tendi

Contents

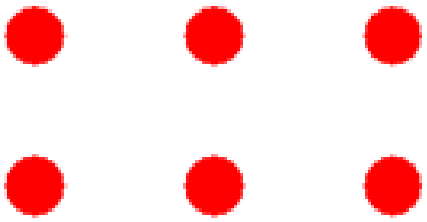
1.3.2 Esantionare

- principala tehnica folosita pentru selectarea da
- în statistica, a fost folosita atât pentru investig
- este folosita pentru ca obtinerea întregului set
- în DM esantionarea este folosita pentru ca pro
- un esantion este reprezentativ daca are aprox
- utilizarea unui esantion reprezentativ e aproap
- tipuri de esantionare:

- esantionare aleatoare uniforma: avem o p
- esantionare fara înlocuire: daca un obiect
- esantionare cu înlocuire: obiectele nu sur
- esantionarea stratificata: se divid datele în cu proportia din multimea initiala.

Esantion de 8000 de punc

- Determinarea marimii esantionului
  - problema: avem un set de date care cons
  - trebuie extras un esantion astfel încât din
  - “garantarea” se exprima probabilist: care anterior sa depaseasca un anumit prag?



1.3.3 Reducerea dimensionalitatii

- seturile de date pot avea un numar mare de a
- valorile pentru densitate de probabilitate si dis

- solutie: reducerea dimensionalitatii fara a pier
- beneficii: algoritmii de DM lucreaza mai eficier
- Scop:
  - evitarea blestemului dimensionalitatii
  - reducerea timpului de rulare necesar algo
  - datele devin mai usor de vizualizat
  - poate ajuta la eliminarea trasaturilor irelev
- Tehnici folosite:
  - analiza componentelor principale (Princip
  - descompunerea valorilor principale (Singu
  - alte metode: supervizate si transformari n

1.3.4 Selectarea subsetului de atribu

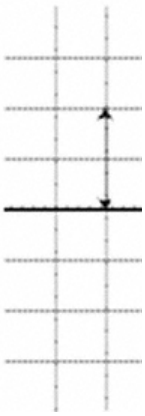
- motivatie: de multe ori, în seturile de date pot
- unele din aceste atribute pot fi eliminate prin "
- varianta ideala de lucru: se încearca toate cor force
- variante:

  1. metode încorporate — algoritmul de DM î
  2. metode de filtrare — atributele sunt selec
  3. metode bazate pe încercare — se foloses
- Pentru selectarea subsetului de trasaturi e nev
  - o masura pentru evaluarea unui subset d
  - o metoda de cautare pentru generarea ur
  - un criteriu de oprire
  - o procedura de validare
- Alternativa pentru selectarea de atribute: ponc
  - atribute importante -> ponderi mari; atribu
  - ponderile se pot asocia pe baza cunostint

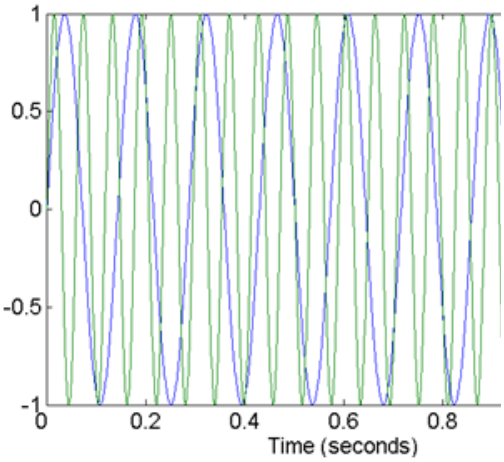
1.3.5 Crearea de trasaturi

- Scop: crearea de noi atribute pe baza celor ex
- Noile atribute pot releva mai eficient informatia
- Numarul atributelor noi obtinute poate fi mai m
- Metode:
  - Extragere de trasaturi
  - Transformarea datelor, utilizarea unui alt s
  - Construirea de trasaturi
- Exemplu: extragerea de trasaturi
  - crearea unui nou set de atribute pe baza
  - exemplu: fotografie -> detectarea muchiil
  - metodele sunt strâns legate de domeniu -

- Exemplu: transformarea datelor
  - trecerea de la coordonatele carteziene la



  - transformata Fourier



Doua functii sinusoidale de frecver

- construirea de noi atribute
  - uneori, atributele originare pot sa aib
  - exemplu: pentru retele neurale intrar
  - exemplu de trasatura derivata: densi
  - crearea de noi trasaturi necesita cun

```
### Discretizare si binarizare
* unii algoritmi cer ca datele sa fie în
* algoritmii de determinare a asocierilor
* apare nevoia de a transforma atribute c
* Discutie pentru binarizare:
    * $m$ valori -> fiecare valoare initi
transcrierea în baza 2
    * daca atributele sunt ordinale, atur
    * exemplu: {slab, mediu, bun} -> $\{(
    * $\lceil \log_2 m \rceil$ biti folos
    * one hot encoding: pentru m clase se
    exemplu: {slab, mediu, bun} -> $(x_1
* Discretizare a atributelor continue
    * procesul de discretizare este deper
    * pasi:
        * se decide numarul de categorii
        * se decide modul de asociere înt
        * exemplu: se sorteaza datele, se
eticheta
* Discretizare nesupervizata
    * intervalele pot fi de latime egala
    * prezinte frecvente egale
    * metode de clustering, e.g. K-means
    * inspectare vizuala a datelor
 <table>
  <tr>
      <td> <img src="./images/discretizar
      <td> <img src="./images/discretizar
  </tr>
  <tr>
      <td> <img src="./images/discretizar
      <td> <img src="./images/discretizar
```

```
  </tr>
</table>
* Discretizare supervizata
    * utilizarea de informatie suplimenta
    * conceptual: discretizarea sa se fac
    * metode statistice: intervale mai mi
* Atribute categoriale cu prea multe atri
    * exemplu: o multitudine de departame
    * se pot unifica pe domenii: ingineri
    * altfel: gruparea poate fi facuta în
clasificare
```

1.3.7 Transformarea de variabile

- transformare care se aplica la toate valorile ur
- exemplu: daca doar amplitudinea unei valori e
- functii folosite: $\exp(x)$, $\log(x)$, $\sqrt{x}$, $1/x$, func
- functii complexe: normalizare
- pentru datele care prezinta domeniu mare de
  - exemplu: transferuri de date de 109, 108, 9, 8, 3, 1.
- aplicarea trebuie facuta în cunostinta de cauz

1.3.8 Exemplu de transformare: star

- în statistica: daca $\overline{x}$ este media unui atribut si

  atunci transformarea $x \rightarrow \frac{x-\overline{x}}{s_x}$ creeaza o nou
- exemplu: persoane cu variabilele: venitul anua venituri ar domina distanta între doi oameni, d
- problema: abaterea patratica medie e influenta

  unde μ e media sau mediana

Explorarea datelor: Statisti

1.4.1 Explorarea datelor

- Explorarea datelor reprezintă investigarea pre
- Pasul de explorare poate fi de folos în alegere
- Se poate folosi abilitatea naturală a oamenilor
- Domeniul a fost introdus de către statisticianu
- AED este domeniu opus lui “Confirmatory Dat

- Curs de AED: aici (https://www.coursera.org/le
science&utm_source=gg&utm_medium=sem&
ROW&adgroupid=116274867101&device=c&k
yWJrboZGvrVzGL9QL0-8PazMsYrrQaAnG4E

```
* În AED, așa cum este definit de Tukey:
    * Focus-ul este pe vizualizare
    * Gruparea (clustering) și detectarea
    * Acestea două sunt subdomenii aparte
```

Setul de date Iris

```
* Constă în date măsurate pentru 150 de f
exemplare pe specie)
* Măsurătorile sunt pentru lungimea/lățim
* A cincea coloană este specia florii - a
* Datele se pot descărca [de aici](http:/
```

In [35]:
```
iris_url = 'http://archive.ics.uci.edu/
iris_columns = ['sepal_length', 'sepal_
data = pd.read_csv(iris_url, header = N
data.columns = iris_columns
data.describe(include='all')
```

executed in 724ms, finished 20:37:59 2021-03-21

Out[35]:

	sepal_length	sepal_width	petal_length	
count	150.000000	150.000000	150.000000	
unique	NaN	NaN	NaN	
top	NaN	NaN	NaN	
freq	NaN	NaN	NaN	
mean	5.843333	3.054000	3.758667	
std	0.828066	0.433594	1.764420	
min	4.300000	2.000000	1.000000	
25%	5.100000	2.800000	1.600000	
50%	5.800000	3.000000	4.350000	
75%	6.400000	3.300000	5.100000	
max	7.900000	4.400000	6.900000	

Sursa: morioh.com

1.4.3 Statistici de sumarizare

- Statisticile de sumarizare sunt numere care sc
- Reprezintă manifestarea cea mai vizibilă a sta
- Exemple: frecvenţa, media, dispersia

1.4.3.1 Frecvenţa şi valoarea modală

Pentru un set de m date categoriale cu valorile $\{v$

iar frecventa absoluta este numaratorul fractiei de

In [30]:
```
from collections import Counter
freq_iris = Counter(data['class'])
rel_freq_iris = {iris_type:freq / len(d
```

executed in 5ms, finished 20:28:35 2021-03-21

In [31]:
```
print(f'Frequency: {freq_iris}')
print(f'Relative Frequency: {rel_freq_i
```

executed in 6ms, finished 20:29:02 2021-03-21

```
Frequency: Counter({'Iris-setosa': 50, 'I
Relative Frequency: {'Iris-setosa': 0.33
33}
```

```
* Valoarea modală (sau moda) este valoare
$$
moda = \arg\max\limits_{v_i} frecventa(v_
$$
* Atenție la situația când o anume valoar
* Pot exista seturi de date pentru care f
* Pentru valori continue, conceptele de m
```

Percentile

```
* Pentru cazul valorilor ordonate se pot
* Pentru un atribut continuu sau ordinal
șirul de valori ale lui $x$ astfel încât
* Nu există o definiție standardizată per
* Pentru cazul în care se calculează perc
devin neesențiale
* Tradițional se consideră $x_{0\%} = \mi
* Mod de calcul pentru determinarea celei
cea mai apropiată de $\frac{m}{100} \cdot
sortat
* Functie utilizabila: [numpy.percentile]
pandas [pandas.Dataframe.quantile](https:
```

In [34]:
```
help(np.percentile)
```

executed in 182ms, finished 20:36:59 2021-03-21

```
        a = np.arange(4)
        p = np.linspace(0, 100, 6001)
        ax = plt.gca()
        lines = [
            ('linear', None),
            ('higher', '--'),
            ('lower', '--'),
            ('nearest', '-.'),
            ('midpoint', '-.'),
        ]
        for interpolation, style in lines
            ax.plot(
                p, np.percentile(a, p, in
                label=interpolation, line
        ax.set(
            title='Interpolation methods
            xlabel='Percentile',
            ylabel='List item returned',
            yticks=a)
```

In [53]:
```
_25th_percentile_sepal_length = np.perc

print(f'Cate valori sunt strict mai mic
print(f'Cate valori sunt strict mai mic
print(f'lista valorilor sortate: {list(

# se poate intelege pentru aceste date
```

executed in 9ms, finished 20:44:19 2021-03-21

```
Cate valori sunt strict mai mici decat a
Cate valori sunt strict mai mici sau egal
lista valorilor sortate: [4.3, 4.4, 4.4,
9, 4.9, 5.0, 5.0, 5.0, 5.0, 5.0, 5.0, 5.0
5.4, 5.4, 5.4, 5.4, 5.4, 5.4, 5.5, 5.5, 5
5.7, 5.8, 5.8, 5.8, 5.8, 5.8, 5.8, 5.8, 5
6.2, 6.3, 6.3, 6.3, 6.3, 6.3, 6.3, 6.3, 6
6.7, 6.7, 6.7, 6.7, 6.7, 6.7, 6.8, 6.8, 6
```

In [54]:
```
data['sepal_length'].quantile(q=0.25)
```

executed in 15ms, finished 20:48:41 2021-03-21

Out[54]: 5.1

Media si mediana

* Pentru un set de valori ${x_1, x_2, . .
$$
\overline{x} = E(x) = mean(x) = \frac{1}{
$$
* Mediana este a 50-a percentila. O varia
(reordonarea) $\{x_{(1)}, x_{(2)}, \dots,
$$
mediana(x) = \begin{cases}
x_{r+1} & \text{daca }m=2r+1
\\
\frac{x_{(r)} + x_{(r+1)}}{2} & \text{dac
\end{cases}
$$
O varianta mai rafinata (si care se poate
(complexitate liniara in loc de $O(n\log

Obtinere: [numpy.median](https://numpy.or
(https://pandas.pydata.org/pandas-docs/st

In [57]:
```
np.median(data.sepal_length.values)
```

executed in 13ms, finished 21:18:38 2021-03-21

Out[57]: 5.8

Diferente intre medie si mediana:

- Media este valoare "de mijloc" doar dacă distr
- Dacă distribuţia este asimetrică, atunci media

- Media este influenţată de outliers, în timp ce n
- Medie retezată (eng: trimmed mean) se utilize din date; se calculează media pentru ceea ce
- media standard se obţine din media retezată

```
Exemple:
* Considerăm valorile {1, 2, 3, 4, 5, 90}
considerabil diferită faţă de media setul
* Media, medianele şi valoarea de trimmec
<img src="./images/stats_iris.png" alt="d
* Exerciţiu: dacă valoarea medianei este
```

Măsurari ale împrăştierii datelor

```
* Sunt măsuri care cuantifică concentrare
* Diametrul domeniului de valori (eng: ra
$$
range(x) = \max(x) - \min(x) = x_{(m)} -
$$
* Range-ul este nerelevant, deoarece pute
outlier măresc artificial diametrul setul
* Varianţa (dispersia; eng: variance) unu
$$
varianta(x) = s_x^2 = \frac{1}{m-1} \sum\
$$
* Utilizarea numitorului $m - 1$ în loc d
şi are ca scop corectarea abaterii din es
* Abaterea standard este $s_x = \sqrt{s_x
* Python: [numpy.std](https://numpy.org/d
```

In [59]:
```
np.std(data.petal_length), np.std(data.
```

executed in 17ms, finished 21:40:36 2021-03-21

Out[59]:
```
(1.7585291834055201, 0.8253012917851409)
```

- Deoarece media poate să fie distorsionată de
- Se preferă considerarea altor trei măsuri:
  - absolute average deviation, AAD:
  - median absolute deviation, MAD:
  - interquartile range:

```
Verificam ca prin standardizare de date,
```

In [69]:

```
med_petal_length = np.mean(data.petal_l
# centram datele: media lor devine 0
centered_petal_length = data.petal_leng
# assert 0 == np.mean(centered_petal_le
assert np.allclose(0, np.mean(centered_
std_petal_length = np.std(data.petal_le
std_centered_petal_length = np.std(cent
assert np.allclose(std_petal_length, st
standardized_petal_length = centered_pe
# verificare: media 0...
assert np.allclose(0, np.mean(standardi
# si standard deviation ~ 1
assert np.allclose(1, np.std(standardiz
```

executed in 52ms, finished 21:53:56 2021-03-21

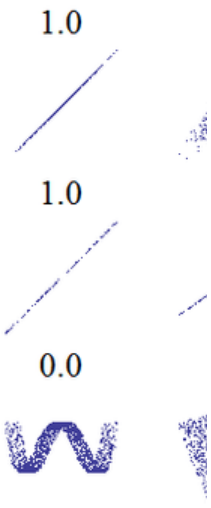
1.4.4 Statistici de sumarizare a datel

- Date multivariate: date cu mai multe atribute
- Pentru atributul x_i calculăm media $\overline{x}_i$
- Media setului de obiecte este $\mathbf{x} = (\overline{x}_1, \ldots, \overline{x}$
- Analog se poate calcula dispersia, mediana et
- Matricea de covarianţă: elementul s_{ij} este cov

  unde x_{pq} este a p-a valoare a atributului x_q
- s_{ij} este măsură a gradului în care două atribu valorilor atributelor
- $s_{ij} = 0$ arata ca atributele s_i si s_j nu sunt lini
- pentru a elimina dependenta de magnitudinea

- r_{ij} se mai numeste corelatia Pearson a atribu
- $r_{ij} = \pm 1$ indica faptul ca x_i e in relatie liniara

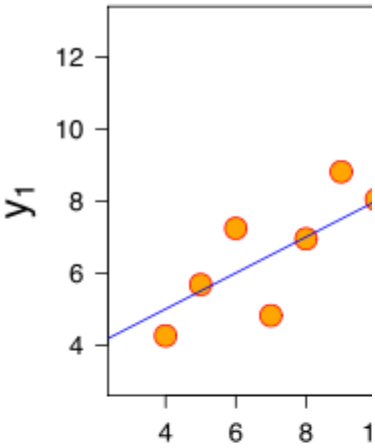
  cu $sgn(a) = sgn(r_{ij})$

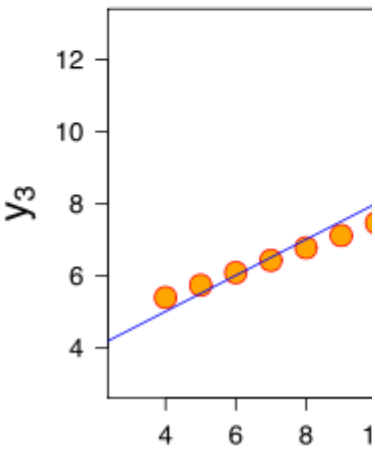


Seturi de date (x, y) împreună cu coeficientul de co
(sus) precum şi modul în care ele sunt legate liniar
relaţii mai complexe între date (rândul de jos). Surs

Legat de coeficientul de corelaţie, câteva observaţ

- “Corelaţia nu înseamnă cauzalitate” – nu se p
  Corelaţie mare poate fi o condiţie necesară pe
  punct de pornire în cercetarea unei legături în
- Corelaţia şi liniaritatea – coeficientul Pearson
  date.
- Exemplu: 4 seturi de date cu două atribute; în
  fiecare caz (0.816); cu toate acestea, legătura





Date cu caracteristici numerice identice (medi

Contents

Contents