Sisteme de Recunoastere a Formelor – Laborator 5

Analiza statistica a datelor

1. Obiective

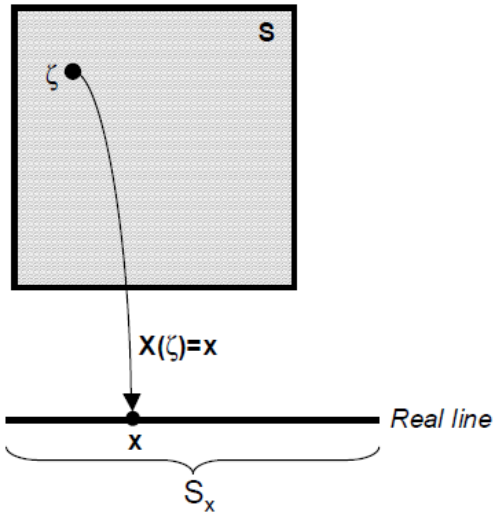
Scopul acestei lucrari este de a explora metodele de analiza statistica a datelor, folosite pentru clasificare si recunoastere. Vom studia media, deviatia standard si covarianta. Experimentele vor fi effectuate pe un set de imagini ce contin fete umane. Folosind matricea de covarianta, vom studia corelatia dintre diferiti pixeli.

2. Fundamente teoretice

2.1 Definitii

O variabila aleatoare X este o functie care ataseaza un numar real $X(\zeta)$ pentru fiecare posibil rezultat ζ din spatiul posibilelor rezultate ale unui experiment aleator (vezi figura de mai jos). Aceasta functie $X(\zeta)$ face o relationare a tuturor posibilelor elemente din spatiul rezultatelor (esantioanelor) cu domeniul numerelor reale (dreapta numerelor reale). Variabilele aleatoare pot fi:

- Discrete: numarul rezultat din aruncarea unui zar
- Continue: greutatea unui individ.



Un *vector de variabile aleatoare X* este o functie care ataseaza un vector de numere reale pentru fiecare rezultat $X(\zeta)$ din spatiul esantioanelor *S*. Notiunea de vector aleator este o extensie a notiunii de variabila aleatoare.

$$\boldsymbol{X} = [X_1, X_2, \dots, X_N]^T$$

2.2 Caracterizarea statistica a variabilelor aleatoare

O variabila aleatoare poate fi caracterizata partial prin:

1. Media: reprezinta centrul de masa al unei densitati de probabilitate.

$$E[X] = \mu = \int_{-\infty}^{\infty} x f_x(x) dx$$

2. Varianta sau dispersia: reprezinta "imprastierea" in jurul mediei

$$VAR[X] = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_x(x) dx$$

3. Deviatia standard: Radacina patrata a variantei, se exprima in aceleasi unitati ca variabila aleatoare.

$$STD[X] = VAR[X]^{1/2}$$

$f_x(x)$ este functia densitate de probabilitate pentru variabila *X*.

2.3 Caracterizarea statistica a vectorilor aleatori

Putem descrie partial un vector aleator prin urmatoarele valori:

1. Vectorul mediu:

$$E[\boldsymbol{X}] = \left[E[X_1], E[X_2], \dots, E[X_N]\right] = [\mu_1, \mu_2, \dots, \mu_N] = \boldsymbol{\mu}$$

2. Matricea de covarianta:

$$COV[\boldsymbol{X}] = \boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)^T] & \dots & E[(X_1 - \mu_1)(X_N - \mu_N)^T] \\ \dots & \dots & \dots \\ E[(X_N - \mu_N)(X_1 - \mu_1)^T] & \dots & E[(X_N - \mu_N)(X_N - \mu_N)^T] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \dots & c_{1N} \\ \dots & \dots & \dots \\ c_{N1} & \dots & \sigma_N^2 \end{bmatrix}$$

Matricea de covarianta indica tendinta fiecarei perechi de trasaturi (pozitii in vector) sa varieze impreuna, sau sa co-varieze. Covarianta are cateva proprietati importante:

- Daca X_i si X_k cresc impreuna, atunci $c_{ik} > 0$
- Daca X_i tinde sa descreasca atunci cand X_k creste, atunci $c_{ik} < 0$
- Daca X_i si X_k sunt necorelate, atunci $c_{ik} = 0$
- $|c_{ij}| < \sigma_i \sigma_j$, unde σ_i este deviatia standard a lui X_i
- $c_{ii} = VAR[X_i]$
- $c_{ij} = c_{ji}$

Termenii matricii de covarianta pot fi scrisi ca:

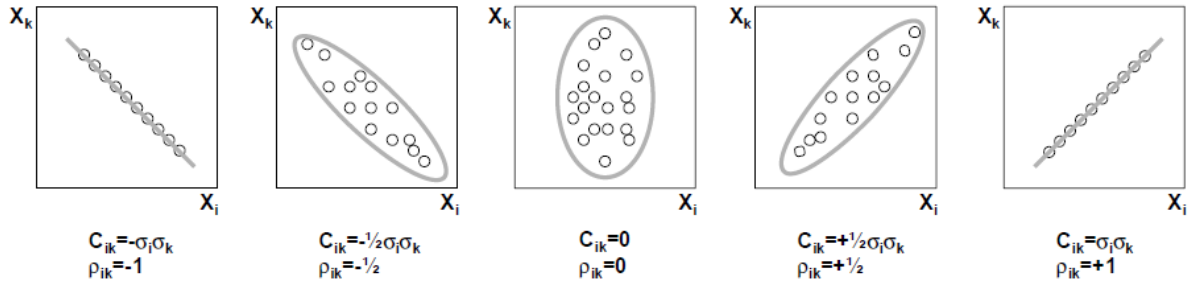
$$c_{ik} = E[(X_i - \mu_i)(X_k - \mu_k)]$$

$$c_{ii} = \sigma_i^2$$

$$c_{ik} = \rho_{ik} \sigma_i \sigma_k$$

unde ρ_{ik} este numit **coeficientul de corelatie**.

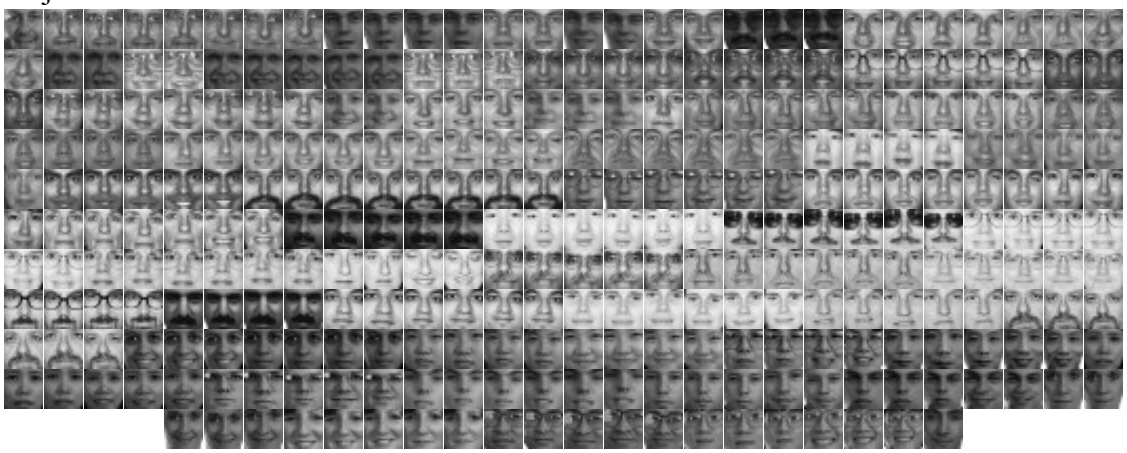
Figurile urmatoare prezinta **graficele de corelatie** dintre trasaturile X_i si X_k.



3. Aspecte practice

In sedinta de laborator va trebui sa studiati corelatia dintre pixeli apartinand unor fete umane.

Se dau p imagini, fiecare imagine continand o fata umana ($p = 400$), ca in imaginile de mai jos.



Se formeaza matricea de trasaturi $\boldsymbol{I}$ care va contine intensitatile din toate imaginile de intrare. $\boldsymbol{I}$ are dimensiunea $p \times N$, unde p este numarul de imagini si N este numarul de pixeli din imagine. Randul k contine toti pixelii din imaginea k rearanjate rand dupa rand in urmatorul mod:

$$\begin{bmatrix} A_{00} & A_{01} & A_{02} \\ A_{10} & A_{11} & A_{12} \\ A_{20} & A_{21} & A_{22} \end{bmatrix} \rightarrow [A_{00}, A_{01}, A_{02}, A_{10}, A_{11}, A_{12}, A_{20}, A_{21}, A_{22}]$$

Fiecare imagine din set are dimensiunea 19x19 pixeli. Interpretarea matricii $\boldsymbol{I}$ este ca fiecare rand contine un esantion al variabilei aleatoare N dimensionale $\boldsymbol{X}$, care urmareste distributia setului de date.

Sarcina voastra este sa calculati matricea de covarianta pentru un set dat de imagini si sa studiati cum variaza diferite trasaturi impreuna.

Valoarea medie al unei trasaturii de la pozitia i in imagine este:

$$\mu_i = \frac{1}{p}\sum_{k=1}^{p} I_{ki}$$

Unde I_{ki} reprezinta valoarea trasaturii *i* din imaginea *k*.

Deviatia standard a trasaturii *i* este:

$$\sigma_i = \sqrt{\frac{1}{p}\sum_{k=1}^{p}(I_{ki} - \mu_i)^2}$$

Elementele matricii de covarianta c_{ij} pot fi calculate ca:

$$c_{ij} = \frac{1}{p}\sum_{k=1}^{p}(I_{ki} - \mu_i)(I_{kj} - \mu_j)$$

Iar coeficientul de corelatie este:

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j}$$

Se verifica $c_{ii} = \sigma_i^2$ si $\rho_{ii} = 1$.

4. Activitate practica

1. Incarcati cele 400 de imagini si stocati valorile de intensitate din fiecare imagine ca si randuri in matricea de trasaturi ***I***. Cod exemplu care incarca setul de imagini este:

```
char folder[256] = "faces";
char fname[256];
for(int i=1; i<=400; i++){
      sprintf(fname,"%s/face%05d.bmp", folder, i);
      Mat img = imread(fname, 0);
}
```

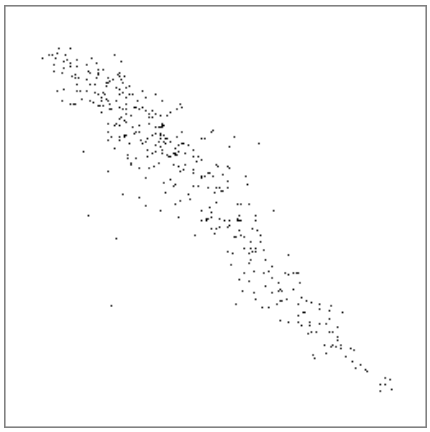
2. Calculati vectorul cu valorile medii si salvati-l intr-un fisier text de tip csv (comma separated values). Se scriu componentele cu virgule intre ele si se salveaza intr-un fisier text cu extensia csv. Acest tip de fisier poate fi deschis cu Microsoft Excel sub forma unui tabel.

3. Calculati matricea de covarianta si salvati-o intr-un fisier csv.

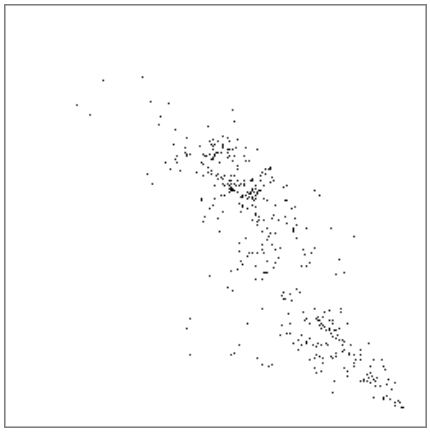
4. Calculati matricea coeficientilor de corelatie si salvati-o intr-un fisier csv.

5. Afisati coeficientul de corelatie si graficul de corelatie pentru urmatoarele pozitii (linie, coloana). Graficul de corelatie este o imagine alba de dimensiune 256x256 care contine puncte negre la fiecare pozitie (I_{kj}, I_{ki}), unde *k=1:p* iar *i* si *j* au fost fixate si reprezinta pozitii liniarizate din imagine.

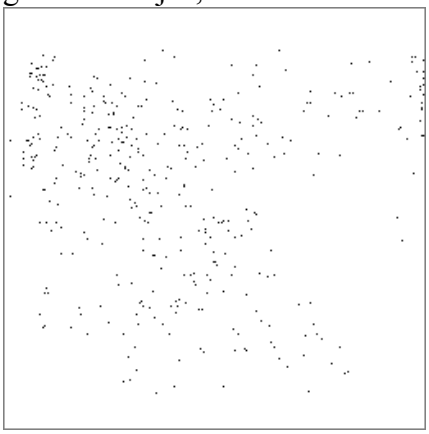
a. (5,4) si (5,14). Aceste puncte corespund unor pixeli apartinand ochiului stang si drept. Rezultatul trebuie sa fie asemanator cu cel din figura de mai jos, si coeficientul de corelatie trebuie sa fie ~ 0.94.



b. (10,3) si (9,15). Aceste puncte corespund pixelilor de pe obrazul stang si obrazul drept. Rezultatul trebuie sa arate ca in figura de mai jos, cu un coeficient de corelatie ~ 0.84.



c. (5,4) si (18,0). Aceste puncte corespund pixelilor care apartin ochiului stang si coltul din stanga jos al imaginii – deci puncte necorelate. Rezultatul ar trebui sa arate ca in figura de mai jos, avand coeficientul de corelatie ~ 0.07.



6. Afisati graficul functiei de densitate de probabilitate unidimensionala pentru o trasatura aleasa. Formula functiei de densitate Gaussiana este:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

unde μ este valoarea medie si σ este deviatia standard pentru trasatura selectata. Se normalizeaza valorile astfel incat maximul sa fie egal cu inaltimea imaginii.

7. Optional, afisati densitatea de probabilitate 2D sub forma unei imagini grayscale pentru doua trasaturi. Forma functiei de densitate Gaussiana este:

$$p(x_i, x_j) = \frac{1}{2\pi\sqrt{\det(C_{ij})}} exp\left(-0.5\left([x_i - \mu_i, x_j - \mu_j] C_{ij}^{-1} \begin{bmatrix} x_i - \mu_i \\ x_j - \mu_j \end{bmatrix}\right)\right)$$

unde μ_i este valoarea medie pentru trasatura *i* iar C_{ij} este matricea de covarianta pentru trasaturile *i* si *j*. Se normalizeaza valorile pentru a obtine intervalul 0:255.

5. Bibliografie

MIT CBCL FACE dataset http://www.ai.mit.edu/courses/6.899/lectures/faces.tar.gz