5 Analiza statistică a datelor

5.1 Obiective

Scopul acestei lucrări este de a explora metodele de analiză statistică a datelor, folosite pentru clasificare și recunoaștere. Vom studia media, deviația standard si covarianța. Experimentele vor fi efectuate pe un set de imagini care conține fețe umane. Folosind matricea de covarianță, vom studia corelația dintre diferiți pixeli.

5.2 Fundamente teoretice

5.2.1 Definiții

În teoria probabilității, *spațiul rezultatelor* S reprezintă setul tuturor rezultatelor unui experiment. De exemplu, pentru experimentul de aruncare a unei monede, spațiul rezultatelor este S = {cap, pajură}. Un subset al spațiului rezultatelor se numește *eveniment*.

În multe cazuri, rezultatele sunt numerice, de exemplu atunci când corespund rezultatului măsurării cu un instrument, dar deseori acestea nu sunt numerice, dar pot fi asociate cu numere reale.

Având un experiment și un spațiu al rezultatelor, o *variabilă aleatorie* X este o funcție care atașează un număr real $X(\zeta)$ pentru fiecare posibil rezultat ζ din spațiul rezultatelor S al unui experiment aleatoriu, după cum se vede în Figura 5.1. Această funcție $X(\zeta)$ face o relaționare a tuturor posibilelor elemente din spațiul rezultatelor (eșantioanelor) cu domeniul numerelor reale (dreapta numerelor reale). Variabilele aleatorii pot fi:

- Discrete: numărul rezultat din aruncarea unui zar, numărul de capete obținut prin aruncarea de 2 ori a unei monede.
- Continue: greutatea unui individ.

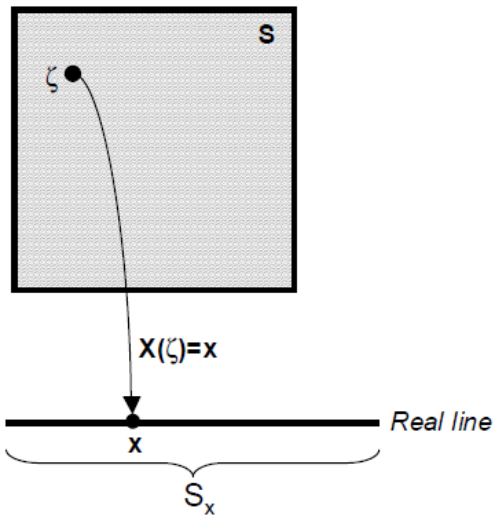


Figura 5.1 Exemplu de variabilă aleatorie.

Un *vector de variabile aleatorii X,* denumit și variabilă aleatorie *multivariată*, este un vector de variabile aleatorii asociate aceluiași experiment.

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

5.2.2 Caracterizarea statistică a variabilelor aleatorii discrete

O variabilă aleatorie *X* se poate caracteriza prin probabilitățile valorilor pe care le poate lua. Pentru o variabilă aleatorie continuă, **funcția de densitate de probabilitate** (FDP) exprimă acest lucru.

Pentru o variabilă aleatorie discretă, **funcția de masă de probabilitate** (FMP) notată prin p_X exprimă acest lucru. Astfel, dacă x este orice valoare posibilă a lui X, funcția de masă de probabilitate $p_X(x)$ reprezintă probabilitatea evenimentului $\{X=x\}$*:*

$$p_X(x) = P(\{X = x\})$$

O proprietate importantă a funcției de masă de probabilitate este:

$$\sum_x p_X(x) = 1$$

unde x ia toate valorile posibile ale lui X.

Este deseori de dorit să sumarizăm funcția de masă de probabilitate printr-un singur număr. Astfel se pot calcula următoarele cantități:

1. *Media* reprezintă o medie ponderată de probabilități a posibilelor valori ale lui X
$$E[X] = \mu = \sum_x x p_X(x)$$
Un caz particular este acela al unei variabile distribuite uniform, unde probabilitățile sunt egale (*1/n*) pentru fiecare valoare a acesteia (în total X poate lua *n* valori). Un exemplu ar fi cel de aruncare a unui zar, iar variabila distribuită uniform este numărul de pe fața zarului. Astfel media se reduce la:
$$E[X] = \mu = \frac{1}{n}\sum_x x$$

2. *Varianța (σ^2)*sau dispersia reprezintă "împrăștierea" in jurul mediei:
$$VAR[X] = E[(X - E[X])^2] = \sum_x (x - \mu)^2 p_X(x)$$

3. *Deviația standard (σ)* este rădăcina pătrată a variației, se exprimă in aceleași unități ca variabila aleatorie:
$$STD[X] = VAR[X]^{1/2}$$

5.2.3 Caracterizarea statistică a vectorilor aleatorii discreți

Putem descrie parțial un vector aleatoriu prin următoarele valori:

1. *Vectorul mediu*:

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

2. *Matricea de covarianță*:

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

$$COV[\boldsymbol{X}] = \begin{bmatrix} VAR(X_1) & \dots & COV[(X_1, X_N)] \\ \dots & \dots & \dots \\ COV[(X_N, X_1)] & \dots & VAR(X_N) \end{bmatrix}$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)^T] & \dots & E[(X_1 - \mu_1)(X_N - \mu_N)^T] \\ \dots & \dots & \dots \\ E[(X_N - \mu_N)(X_1 - \mu_1)^T] & \dots & E[(X_N - \mu_N)(X_N - \mu_N)^T] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \dots & c_{1N} \\ \dots & \dots & \dots \\ c_{N1} & \dots & \sigma_N^2 \end{bmatrix}$$

Pe diagonala matricei de covarianță avem varianța unei variabile X_i din vectorul aleatoriu X unde $i \in [1, N]$, iar pe celelalte poziții din matrice avem covarianța dintre două perechi de variabile (X_i, X_k) din vectorul aleatoriu.

Definim covarianța ca:

$$COV[X_i, X_k] = \sum_{x_i} \sum_{x_k} (x_i - \mu_i)(x_k - \mu_k) p_{X_i, X_k}(x_i, x_k)$$

Unde $p_{X_i, X_k}(x_i, x_k)$ este *funcția de masă de probabilitate comună* a variabilelor aleatorii X_i și X_k.

Matricea de covarianță indică tendința fiecărei perechi de variabile aleatorii să varieze împreună, sau să co-varieze.

Covarianța are câteva proprietăți importante:

- Dacă X_i și X_k cresc împreună, atunci $c_{ik} > 0$
- Dacă X_i tinde sa descrească atunci când X_k creste, atunci $c_{ik} < 0$
- Dacă X_i și X_k sunt necorelate, atunci $c_{ik} = 0$
- $|c_{ij}| < \sigma_i \sigma_j$, unde σ_i este deviația standard a lui X_i
- $c_{ii} = VAR[X_i]$
- $c_{ij} = c_{ji}$

Termenii matricei de covarianță pot fi scriși 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 corelație Pearson** .

Figurile următoare prezintă **graficele de corelație** dintre variabilele X_i si X_k.

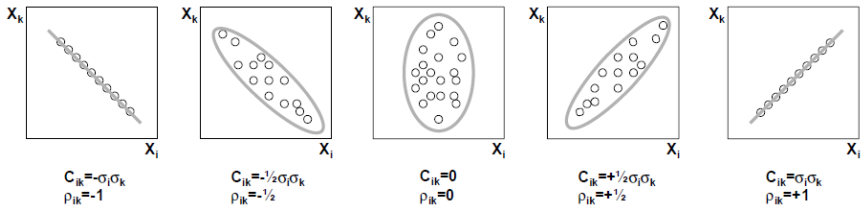


Figura 5.2 – Grafice de corelație între două variabile aleatorii X_i si X_k. De la stânga la dreapta: corelație negativă puternică, corelație negativă medie, nicio corelație, corelație pozitivă medie și corelație pozitivă puternică.

Coeficientul de corelație ρ_{ik} dintre două variabile reprezintă de fapt covarianța normalizată. Acesta ia valori între [-1,1]. Cu cât valoarea este mai apropiată de -1 sau 1, cu atât corelația este mai puternică. Covarianța se folosește atunci când scalele celor două variabile sunt la fel, iar coeficientul de corelație atunci când scalele sunt diferite.

5.3 Aspecte practice

Se dau p imagini, fiecare imagine conține o față umană (p = 400), ca în imaginile de mai jos din setul de date MIT CBCL FACE [1].

Se formează matricea de trăsături ***I*** care va conține intensitățile din toate imaginile de intrare. ***I*** are dimensiunea $p \times N$, unde p este numărul de imagini și N este numărul de pixeli din imagine. Rândul k conține toți pixelii din imaginea k rearanjați rând după rând în următorul 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 *N*=19x19 pixeli. Interpretarea matricei ***I*** este că fiecare rând conține un eșantion al variabilei aleatorii *N* dimensionale ***X***, care urmărește distribuția setului de date.

Obiectivul este calculul matricei de covarianță pentru un set dat de imagini și observarea modului în care variază împreună diferitele trăsături.

Valoarea medie a unei trăsături de la poziția *i* in imagine este:

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

Unde I_{ki} reprezintă valoarea trăsăturii *i* din imaginea *k*.

Deviația standard a trăsăturii *i* este:

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

Elementele matricei de covarianță 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 corelație este:

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

Se observă că $c_{ii} = \sigma_i^2$ și $\rho_{ii} = 1$.

5.4 Activitate practică

1. Încărcați cele 400 de imagini și stocați valorile de intensitate din fiecare imagine ca și rânduri in matricea de trăsături ***I***. Secțiunea de cod exemplu care încarcă 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, IMREAD_GRAYSCALE);
}
```

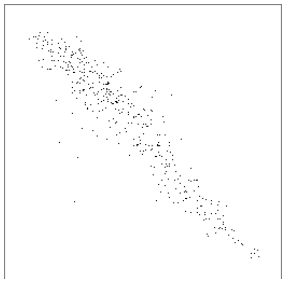
2. Calculați vectorul cu valorile medii si salvați-l într-un fișier text de tip CSV (comma separated values). Se scriu componentele cu virgule între ele și se salvează într-un fișier text cu extensia CSV. Acest tip de fișier poate fi deschis cu Microsoft Excel sub forma unui tabel.

3. Calculați matricea de covarianță si salvați-o într-un fișier CSV.

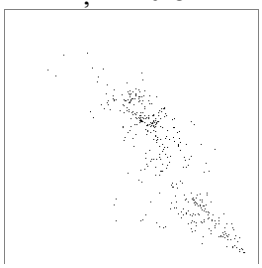
4. Calculați matricea coeficienților de corelație și salvați-o într-un fișier CSV.

5. Afișați coeficientul de corelație si graficul de corelație pentru următoarele poziții (linie, coloana). Graficul de corelație este o imagine albă de dimensiune 256x256 care conține puncte negre la fiecare poziție (I_{kj}, I_{ki}) , unde *k=1:p* iar *i* și *j* au fost fixate și reprezintă pozițiile liniarizate din imagine.

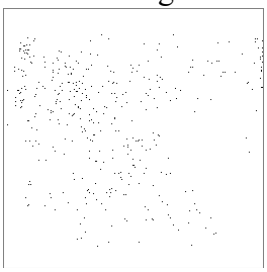
a. (5,4) și (5,14). Aceste puncte corespund unor pixeli aparținând ochiului stâng și drept. În acest exemplu, $i = 5 * 19 + 4$, iar $j = 5 * 19 + 14$. Rezultatul trebuie să fie asemănător cu cel din figura de mai jos, și coeficientul de corelație trebuie sa fie ~ 0.94.



b. (10,3) și (9,15). Aceste puncte corespund pixelilor de pe obrazul stâng și obrazul drept. Rezultatul trebuie să arate ca în figura de mai jos, cu un coeficient de corelație ~ 0.84.



c. (5,4) și (18,0). Aceste puncte corespund pixelilor care aparțin ochiului stâng și coltul din stânga jos al imaginii – deci puncte necorelate. Rezultatul ar trebui să arate ca în figura de mai jos, având coeficientul de corelație ~ 0.07.



6. Afișați graficul funcției de densitate de probabilitate unidimensională pentru o trăsătură aleasă. Formula funcției de densitate gaussiană 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 și σ este deviația standard pentru trăsătura selectată. Se normalizează valorile astfel încât maximul sa fie egal cu înălțimea imaginii.

7. Opțional, afișați densitatea de probabilitate 2D sub forma unei imagini cu niveluri de gri pentru două trăsături. Forma funcției de densitate gaussiană este:

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

unde μ_i este valoarea medie pentru trăsătura *i* iar C_{ij} este matricea de covarianță pentru trăsăturile *i* și *j*. Se normalizează valorile pentru a obține valori în intervalul 0:255.

5.5 Exemple

1. Fie experimentul de aruncare independentă a unei monede de două ori și fie X numărul de capete obținute. Spațiul rezultatelor este in acest caz S = {0, 1, 2}. Funcția de masă de probabilitate a lui X este:

$$p_X(x) = \begin{cases} \frac{1}{4}, & dacă\ x = 0\ sau\ x = 2 \\ \frac{1}{2}, & dacă\ x = 1 \end{cases}$$

Media lui X este:

$$E[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

Varianța lui X este:

$$VAR[X] = (0-1)^2 \cdot \frac{1}{4} + (1-1)^2 \cdot \frac{1}{2} + (2-1)^2 \cdot \frac{1}{4} = \frac{1}{2}$$

2. Considerăm experimentul de aruncare a unui zar, pentru care avem probabilități egale de apariție a fiecărei fețe (1/6). Fie vectorul de variabile aleatorii [X,Y]. Variabila aleatorie X este egală cu 1 dacă numărul de pe față este par (2, 4, 6) și 0 altfel:

$$X = \begin{cases} 1, & dacă\ număr\ par \\ 0, & altfel \end{cases}$$

Variabila aleatorie Y este egală cu 1 dacă numărul de pe față este prim (2, 3, 5) și 0 altfel:

$$Y = \begin{cases} 1, & dacă\ număr\ prim \\ 0, & altfel \end{cases}$$

Funcția de masă de probabilitate a lui X este:

$$p_X(x) = \begin{cases} \frac{3}{6} = \frac{1}{2}, & dacă\ x = 1 \\ \frac{3}{6} = \frac{1}{2}, & dacă\ x = 0 \end{cases}$$

Media lui X este:

$$E[X] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

Varianța lui X este:

$$\text{VAR}[X] = \left(0 - \tfrac{1}{2}\right)^2 \cdot \tfrac{1}{2} + \left(1 - \tfrac{1}{2}\right)^2 \cdot \tfrac{1}{2} = \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{1}{4}$$

Funcția de masă de probabilitate a lui Y este:

$$p_Y(y) = \begin{cases} \frac{3}{6} = \frac{1}{2}, & dacă\ y = 1 \\ \frac{3}{6} = \frac{1}{2}, & dacă\ y = 0 \end{cases}$$

Media lui Y este:

$$E[Y] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

Varianța lui Y este:

$$\text{VAR}[Y] = (0 - \tfrac{1}{2})^2 \cdot \tfrac{1}{2} + \left(1 - \tfrac{1}{2}\right)^2 \cdot \tfrac{1}{2} = \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{1}{4}$$

Funcția de masă de probabilitate comună a lui X și Y definește probabilitățile pentru fiecare pereche de valori ale lui X și Y:

(X = 0, Y=0), (X=0, Y=1), (X=1, Y=0), (X=1, Y=1)

Vom calcula *funcția de masă de probabilitate comună* a variabilelor X și Y astfel:

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{6}, & dacă\ x = 0\ și\ y = 0 \\ \frac{1}{6}, & dacă\ x = 1\ și\ y = 1 \\ \frac{2}{6}, & dacă\ x = 0\ și\ y = 1 \\ \frac{2}{6}, & dacă\ x = 1\ și\ y = 0 \end{cases}$$

Putem calcula covarianța dintre X și Y:

$$COV[X_i, X_k] = \sum_{x_i} \sum_{x_k} (x_i - \mu_i)(x_k - \mu_k) p_{X_i,X_k}(x_i, x_k)$$

Matricea de covarianță este:

	X	**Y**
X	VAR(X) = 1/4	COV(X,Y)= –1/12
Y	COV(Y,X)= –1/12	VAR(Y) = 1/4

$$COV[X,Y] = COV[Y,X]$$
$$= \left(0-\frac{1}{2}\right) * \left(0-\frac{1}{2}\right) * \frac{1}{6} + \left(0-\frac{1}{2}\right) * \left(1-\frac{1}{2}\right) * \frac{2}{6} + \left(1-\frac{1}{2}\right) * \left(0-\frac{1}{2}\right) * \frac{2}{6}$$
$$+ \left(1-\frac{1}{2}\right) * \left(1-\frac{1}{2}\right) * \frac{1}{6} = -\frac{2}{24}$$

5.6 Referințe

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