

Generarea setului de date de antrenament

Generalități

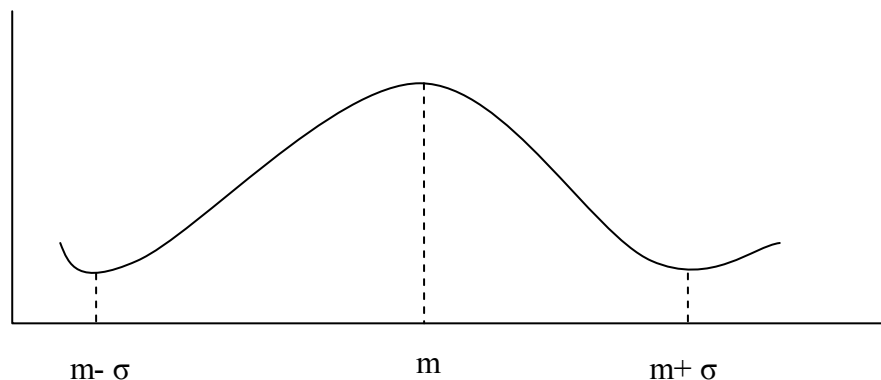
Pentru o mai bună analiză a rezultatelor algoritmilor de învățare care vor fi dezvoltați datele de antrenament (test) vor reprezenta o mulțime de puncte care să poată fi reprezentată grafic într-un spațiu bidimensional. Astfel fiecare dată de intrare va avea două caracteristici: coordonata pe x respectiv coordonata pe y a punctului astfel încât acel punct să fie afișabil pe ecran. Se va propune un algoritm care pe baza funcției lui Gauss va genera aleator aceste puncte.

Probabilitatea Gausseană

Probabilitatea Gausseană reprezintă probabilitatea ca un punct să se găsească în intervalul $[m-\sigma, m+\sigma]$, unde m este mijlocul intervalului (curbei), iar σ reprezintă dispersia datelor (probabilitatea ca punctul respectiv să se găsească sub clopotul lui Gauss)

Funcția lui Gauss:

$$Gauss(x) = e^{-\frac{(m-x)^2}{2\sigma^2}}$$



Algoritmul de generare a unei coordonate aleatoare

Acest algoritm se va rula independent pentru fiecare caracteristică a datei de antrenament (coordonata pe x și pe y). Parametrii m și σ se specifică pentru fiecare grupare de date în parte. Valorile pentru x și y vor fi reprezentate în coordonate carteziane nu coordonate ecran.

1. Se alege aleator o valoare în domeniul de coordonate pentru coordonata pe x;
2. Se calculează probabilitatea dacă valoarea aleasă aleator este sau nu apropiată de centru m (folosind funcția lui Gauss);
3. Se generează aleator o probabilitate în domeniul $[0,1]$;

4. Se verifică dacă probabilitatea calculată pentru coordonata punctului nostru este mai mare decât probabilitatea aleasă aleator;
 - a. Dacă da punctul ales aleator se ia în considerare și se trece la pasul 5;
 - b. Dacă nu se trece din nou la pasul 1;
5. Se reia algoritmul pentru coordonata y , după care la pasul 4.a algoritmul se termina pentru o dată de intrare (un punct).

Pentru o medie m și o dispersie σ dată algoritmul va genera aleator un număr de puncte în jurul lui m .

Problemă. Să se realizeze un program care scrie într-un fișier pe fiecare linie caracteristicile unui punct (dată de intrare). Se vor genera puncte pentru trei centre și dispersii diferite. La final în fișier vor fi 3000 de linii (se va genera aproximativ 1000 de puncte pentru fiecare centru). Liniile cu caracteristicile pentru un centru și o dispersie nu vor fi consecutive (pe cât posibil).

Observație:

Se va considera centrul ecranului ca fiind centru axelor de coordonate (0,0), iar media și punctele vor fi în coordonate carteziane (nu coordonate ecran).