Clasificatorul Naive Bayes

În acest laborator vom clasifica cifrele scrise de mână din subsetul **MNIST** (introdus în laboratorul precedent) folosind clasificatorul Naive Bayes. Revedeți slide-urile de la curs (cursul 3) pentru a vă reaminti în detaliu în ce constă acest model.

Scopul acestui clasificator este de a clasifica o imagine X ce conține o cifră într-una din cele 10 clase din mulțimea {0, 1, …., 9}. Imaginea X o reprezentăm ca un vector cu 784 de componente $X = (X_1, X_2,\ldots, X_{784})$, unde fiecare componentă X_j reprezintă intensitatea pixelului de la poziția j din imaginea X. Pe baza analizei intensității pixelilor clasificatorul clasifică imaginea X într-una din cele 10 clase.

Regula de clasificare pentru clasificatorul Naïve Bayes este de a alege clasa care maximizează probabilitatea a-posteriori. Pe baza calculelor (revedeți materialul de la curs) ajungem la formulele:

$$c^* = \underset{i=0,1,\ldots,9}{\operatorname{argmax}}\left(\sum_{j=1}^{n=784} \log(P(X_j = x_j \,/\, c = i)) + \log(P(c = i))\right)$$

unde:

- $P(X_j = x_j|c = i)$ reprezintă probabilitatea ca, pentru o imagine din clasa i, intensitatea X_j a pixelului de la poziția j să ia valoarea x_j;
- P(c=i) reprezintă probabilitatea a-priori (fără a observa niciun pixel al imaginii X) ca imaginea X să conțină cifra i.

Întrucât estimarea probabilității $P(X_j = x_j|c = i)$ nu poate fi realizată robust din datele de antrenare (din cauza lipsei unui număr mare de date de antrenare) vom aproxima această probabilitate cu probabilitatea ca intensitatea X_j să ia o valoare dintr-un anumit interval $[a_k, b_k]$ din mulțimea de valori {0, 1, …, 255}, notată cu $P(a_k \leq X_j \leq b_k|c = i)$. O posibilitate este de a partiționa {0, 1, …, ,255} într-un număr de 4 părți egale, obținând intervalele [0, 63], [64, 127], [128, 191], [192, 255]. Astfel probabilitatea $P(X_j = 25|c = i)$ va fi aproximată cu $P(0 \leq X_j \leq 63|c = i)$ întrucât 25 se află în intervalul [0, 63] iar probabilitatea $P(X_j = 154|c = i)$ va fi aproximată cu $P(128 \leq X_j \leq 191|c = i)$ întrucât 154 se află în intervalul [128, 191]. Practic mulțimea de 256 de valori posibile {0, 1, 2 ,…, 255} este redusă numai la 4 valori posibile, corespunzătoare celor 4 intervale.

Calculul probabilităților $P(a_k \leq X_j \leq b_k|c = i)$ pentru $i \in \{0, 1, \ldots, 9\}$, $j \in \{1, \ldots, 784\}$, }, $k \in \{1, \ldots, 4\}$ înseamnă de fapt calculul unui matrice M cu elemente de forma $M_{ijk} = P(a_k \leq X_j \leq b_k|c = i)$. Matricea M are trei dimensiuni: indicele i ce se referă la clasa i reprezintă numărul liniei i (în total sunt 10 linii = 10 clase), indicele j ce se referă la componenta (valoarea pixelului) j a imaginii X reprezintă numărul coloanei j (în total sunt 784 de coloane = 784 de pixeli în imaginea X), indicele k ce se referă la intervalul $[a_k, b_k]$ reprezintă adâncimea pe dimensiunea a treia a matricei (în total sunt 4 intervale). Pentru fiecare linie i și coloană j avem că: $M_{ij1}+ M_{ij2}+ M_{ij3}+ M_{ij4}=1$.

Codul de la care porniți exemplifică pentru clasa i = 0 și poziția j = 370 calculul vectorului de 4 probabilitati $M_{ij} = (M_{ij1}, M_{ij2}, M_{ij3}, M_{ij4})$. Acest lucru trebuie să-l realizați și voi pentru toate perechile (i, j) în vederea calculării matricei M. Apoi aveți de implementat regula de clasificare de mai sus

pentru clasificatorul Naïve Bayes prin care alegeți clasa cu maximul probabilității a-posteriori. Pe baza acestei reguli de clasificare veți clasifica exemplele de testare şi calcula apoi matricea de confuzie.

Realizați următoarele:

1. Scrieți o funcție care calculează matricea M din datele de antrenare. Puteți să vă verificați că ați calculat corect pe baza exemplului pentru i=0, j=170 si probabilitățile calculate. Aveți grija să adunați o constantă mica 0.00000001 = 10e-9 la elementele matricei M astfel încât să nu aveți probabilități egale cu 0 în matricea M.
2. Scrieti o funcție care implementează regula de clasificare pentru clasificatorul naive Bayes pentru un exemplu de testare.
3. Folosiți regula de clasificare implementată a clasificatorului naive Bayes și clasificați exemplele de testare. Calculati matricea de confuzie.
4. Care sunt perechile de exemple de testare cu cele mai multe misclasificari in matricea de confuzie. Plotati toate aceste exemple misclasificate.

Funcții numpy utile în rezolvarea laboratorului:

```
x = np.array([1, 2, 3, 4, 3, 4])
np.argmin(x)  # returneaza pozitia elementului minim
np.argmax(x)  # returneaza pozitia elementului maxim
np.where(x == 3) # returneaza indecsi care satisfac conditia
```