

Naive Bayes

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

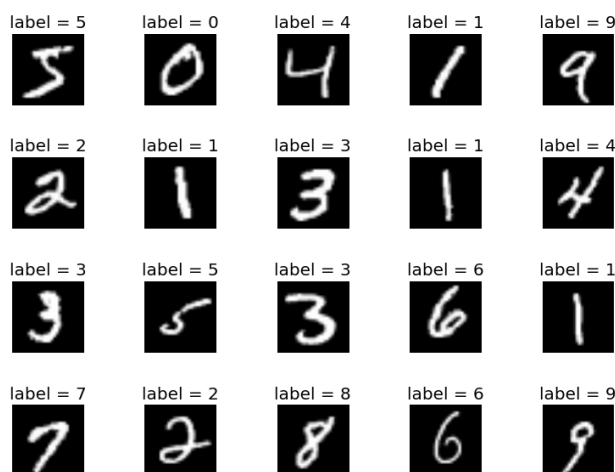
Likelihood
Class Prior Probability
Posterior Probability
Predictor Prior Probability

Regula Bayes

În acest laborator vom clasifica cifrele scrise de mână din subsetul **MNIST** folosind Naive Bayes.

MNIST¹ este o bază de date cu cifre scrise de mână (0-9), conținând 60.000 de imagini pentru antrenare și 10.000 pentru testare. Imaginile sunt alb-negru având dimensiunea de 28x28 pixeli. În cadrul laboratorului vom lucra pe un subset, împărțit astfel:

- În 'train_images.txt' sunt 1.000 de imagini din mulțimea de antrenare, fiecare fiind stocată pe câte o linie a matricei de dimensiune 1000 x 784 (28 x 28 = 784).
- În 'test_images.txt' sunt 500 de imagini din setul de testare.
- Fișierele 'train_labels.txt' și 'test_labels.txt' conțin etichetele imaginilor.



Exemple de imagini din setul de date MNIST.

Descărcați arhiva care conține datele de antrenare și testare [de aici](#).

¹ <http://yann.lecun.com/exdb/mnist/>

Pentru vizualizarea unei imagini din mulțimea de antrenare trebuie să redimensionăm vectorul de 1 x 784 la 28 x 28.

```
from skimage import io # pentru afisarea imaginii
import numpy as np

train_images = np.loadtxt('train_images.txt') # incarcam imaginile
train_labels = np.loadtxt('train_labels.txt', 'int') # incarcam etichetele avand
                                                    # tipul de date int

image = train_images[0, :] # prima imagine
image = np.reshape(image, (28, 28))
io.imshow(image.astype(np.uint8))
io.show()
```

Deoarece datele noastre (valorile pixelilor) sunt valori continue, va trebui sa le transformăm în valori discrete cu ajutorul unei histogramme. Vom stabili numărul de intervale la care vom împărți lungimea intervalului valorilor continue, apoi vom asigna fiecărei valori continue indicele intervalul corespunzător.

```
bins = np.linspace(start=0, stop=255, num=num_bins) # returneaza intervalele
x_to_bins = np.digitize(x, bins) # returneaza pentru fiecare element intervalul
                                # corespunzator
                                # Atentie! Indexarea primul interval se face de la 1
```

1. Antrenarea clasificatorului (fit)

O imagine $X = \{x_1, x_2, \dots, x_{784}\}$ din mulțimea de antrenare are dimensiunea de 1x784. Conform presupunerii clasificatorului Naive Bayes, vom considera fiecare pixel ca fiind un atribut **independent** în calcularea probabilității apartenenței lui X la clasa c .

$$P(c|X) \propto p(c) \prod_{i=1}^{784} P(x_i|c) \quad | \text{ aplicăm logaritm}$$

$$\log(P(c|X)) = \log(P(c)) + \sum_{i=1}^{784} \log(P(x_i|c))$$

Pentru aplicarea regulii Naive Bayes avem nevoie de:

1. $P(c) = \frac{\text{numărul exemplelor din clasa } c}{\text{numărul total de exemple}}$, probabilitatea ca un exemplu să se afle în clasa c
2. $P(x_i|c) = \frac{\text{numărul exemplelor din clasa } c \text{ care au atributul } i \text{ egal cu o anumită valoare}}{\text{numărul exemplelor din clasa } c}$,

Pentru implementarea regulii Naive Bayes avem nevoie de:

1. Un vector de dimensiune (num_classes, 1) în care să stocăm $P(c)$ (probabilitatea fiecărei clase).

- Un vector de dimensiune (num_features, num_bins, num_classes), în care stocăm $P(x_i | c)$ (probabilitatea ca atributul x_i să aibă o anumită valoare (într-un interval dat de num_bins) dacă face parte din clasa c).

2. Prezicerea etichetelor pe baza clasificatorului (predict)

$$P(c|X) \propto p(c) \prod_{i=1}^n P(x_i|c), \text{ unde } X = \{x_1, x_2, \dots, x_n\} \text{ cu } x_1, \dots, x_n \text{ attribute independente.}$$

Probabilitatea ca exemplul $X = \{x_1, x_2, \dots, x_{784}\}$ să provină din clasa c , se obține prin înmulțirea (sau adunarea logaritmilor) probabilităților individuale ale atributelor acestuia condiționate de clasa c . Vom calcula $P(c|X)$ pentru fiecare clasă c ($c \in [1, \text{num_classes}]$), iar eticheta finală este dată de clasa cu probabilitatea cea mai mare.

Pentru implementarea predicției folosind clasificatorul Naive Bayes avem nevoie de:

- Cei 2 vectori ($P(c)$, $P(x|c)$) calculați anterior.
- Un vector *probs* de dimensiune (num_samples, num_classes) în care să stocăm $P(c|X)$ (probabilitatea ca X să se afle în clasa c) pentru fiecare exemplu de test.
- Eticheta prezisă este indicele probabilității maxime de pe fiecare linie din vectorul *probs*.

Exerciții

- Creați clasa **NaiveBayes**, având constructorul `__init__(self, num_bins, max_value)` în care se setează și se calculează granițele (extremitățile) intervalelor în care trebuie împărțit spațiul de valori. Definiți metoda *values_to_bins* care primește o matrice de dimensiune (n_samples, n_features), iar pentru fiecare exemplu și fiecare atribut calculează indexul intervalului corespunzător.
- Definiți metoda *fit* care primește datele de antrenare (*train_images* și *train_labels*) și calculează și stochează probabilitățile necesare aplicării regulii Naive Bayes (după formulele de mai sus).

OBS. Pentru stabilitate numerică adunați $1e-10$ la $P(x|c)$.

- Definiți metoda *predict* care primește argumentul *test_images* și returnează etichetele prezise folosind regula Naive Bayes (logaritmând probabilitățile).
- Definiți metoda *score* care primește argumentele *test_images* și *test_labels* și returnează acuratețea medie.

5. Testați clasificatorul Naive Bayes pe subsetul MNIST folosind $num_bins \in \{3, 5, 10\}$.

OBS. Acuratețea pe care trebuie să o obțineți pentru $num_bins = 5$ este de 79.0%.

6. Calculați matricea de confuzie $C = (C_{ij})$ unde C_{ij} reprezintă numărul de exemple de testare din clasa i clasificate în clasa j . Plotați câteva exemple de testare clasificate greșit.

Funcții numpy:

```
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
```