

1. Se dă mulțimea de antrenare, reprezentând înălțimea în cm a unei persoane și eticheta corespunzătoare:
 $[(160, F), (165, F), (155, F), (172, F), (175, B), (180, B), (177, B), (190, B)]$.
 Împărțind valorile continue (înălțimea) în 4 intervale (150-160, 161-170, 171-180, 181-190), calculați probabilitatea ca o persoană având 178 cm, să fie fată sau să fie băiat, folosind regula lui Bayes.

```
# NAIVE BAYES

#  $P(C|x) = ((P(x|C)) * P(C)) / P(x)$ 

#  $P(F) = 4/8 = 1/2$ 
#  $P(B) = 4/8 = 1/2$ 
#  $P(Cat1) = 2/8 = 1/4$ 
#  $P(F/Cat1) = 2/4 = 1/2$ 
#  $P(B/Cat4) = 1/4$ 

# Pentru (178, ?)
# Calculam  $P(Cat3/F)$  si  $P(Cat3/B) \Rightarrow$  MAXIMUL DINTRE ELE

#  $P(Cat3/F) = ((P(F|Cat3)) * P(Cat3)) / P(F) = (1/4 * 1/2) / 1/2 = 1/4$ 
#  $P(cat3/B) = (3/4 * 1/2) / 1/2 = 3/4$ 
```

```
# Exista ceva anumit MATRICEA DE CONFUZIE
# O matrice care are liniile TRAIN LABELS si coloanele TEST LABELS

#           F                               B
# F  nr de fete prezise corect    nr de fete prezise ca baieti

# B           --//--                     --//--

# Scopul meu e ca diagonala principala sa fie maxima (restul valorilor
# sunt misclassified)
```

```

[[51.  0.  0.  0.  0.  0.  1.  0.  1.  0.]
 [ 0. 48.  0.  0.  0.  0.  0.  0.  4.  0.]
 [ 2.  1. 50.  1.  1.  0.  1.  1.  1.  0.]
 [ 0.  0.  1. 49.  0.  0.  0.  0.  0.  3.]
 [ 0.  0.  0.  0. 33.  0.  0.  0.  2. 11.]
 [ 1.  0.  0.  9.  0. 34.  1.  0.  6.  1.]
 [ 1.  1.  0.  0.  1.  0. 43.  0.  2.  0.]
 [ 0.  1.  0.  0.  2.  0.  0. 41.  0.  6.]
 [ 0.  1.  3.  3.  1.  1.  1.  1. 34.  1.]
 [ 0.  0.  1.  1.  5.  0.  0.  0.  0. 35.]]

```

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

What proportion of positive identifications was actually correct?

Precision is defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

What proportion of actual positives was identified correctly?

Mathematically, recall is defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$L1(X, Y) = \sum_{i=1}^n |X_i - Y_i|$$

$$L2(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

L1 - MANHATTAN

L2 - EUCLID

- MSE: mean squared error: diferența dintre etichetele true și etichetele predicted la pătrat (scopul nostru e să obținem un mse și mae cât mai mic)
- MEA: mean absolute error: diferența dintre etichetele true și etichetele predicted supra numărului de date

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

$x_{scaled} = \frac{x - mean(x)}{\sigma}$, unde x_{mean} - media valorilor lui x
 σ - deviația standard

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

sample variance

$$s^2 = \frac{166}{5} = 33.2$$

sample standard deviation

$$s = 5.76$$

2. Care dintre următoarele asigura media 0 pentru fiecare din trasaturile din setul de date? X reprezintă setul de date, X_i reprezintă setul de trasaturi i al tuturor exemplilor din setul de date iar x reprezintă un exemplu din setul de date.

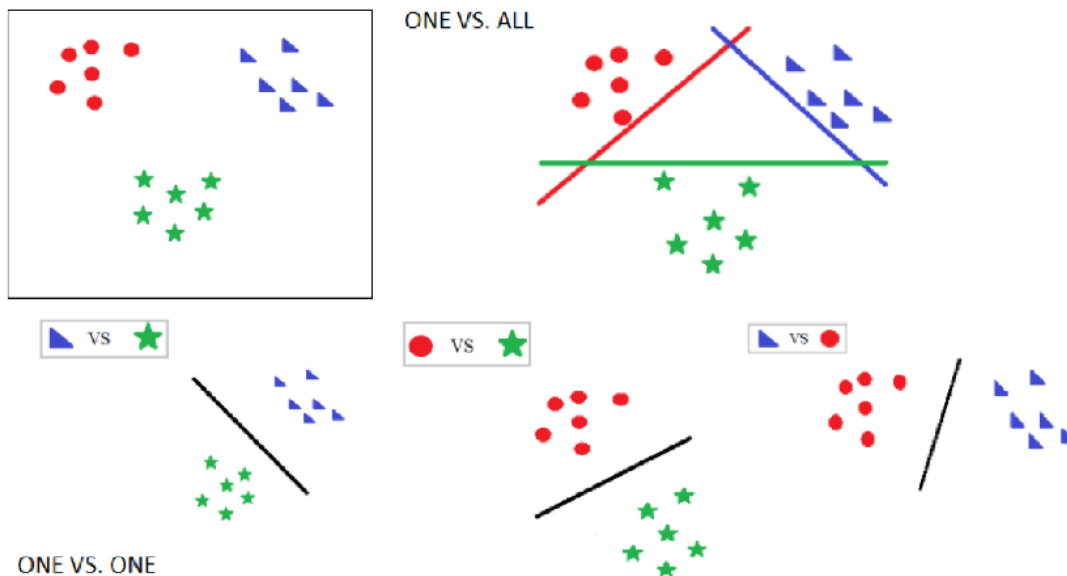
A. L1 Normalization ($x / \sum |x_i|$) pentru fiecare exemplu x)

B. Standard Normalization ($(X_i - \text{mean}(X_i)) / \text{std}(X_i)$) pentru fiecare trasatura i)

C. Min-Max Scaling ($(X_i - \min(X_i)) / (\max(X_i) - \min(X_i))$) pentru fiecare trasatura i)

D. L2 Normalization ($x / \sqrt{\sum x_i^2}$) pentru fiecare exemplu x)

Normalizarea standard aduce datele la medie 0 și deviație standard 1.



Există două abordări pentru a clasifica datele aparținând mai multor clase:

1. **ONE VS ALL:** Sunt antrenați $num_classes$ clasificatori, câte unul corespunzător fiecărei clase, care să o diferențieze pe aceasta de toate celelalte (toate celelalte exemple sunt privite ca aparținând aceleiași clase). Eticheta finală pentru un exemplu nou va fi dată de clasificatorul care a obținut scorul maxim.
2. **ONE VS ONE:** Sunt antrenați $\frac{num_classes * (num_classes - 1)}{2}$ clasificatori, câte unul corespunzător fiecărei perechi de câte două clase. Eticheta finală pentru un exemplu nou va fi cea care obține cele mai multe voturi pe baza acestor clasificatori.

- C MARE - OVERFITTING
- C MIC - UNDERFITTING

dreaptă este folosită abordarea soft margin. Variabila ξ_i sugerează cât de mult exemplul x_i are voie să depășească marginea.

$$\xi_i = \max(0, 1 - y_i(\langle x, w \rangle + b))$$