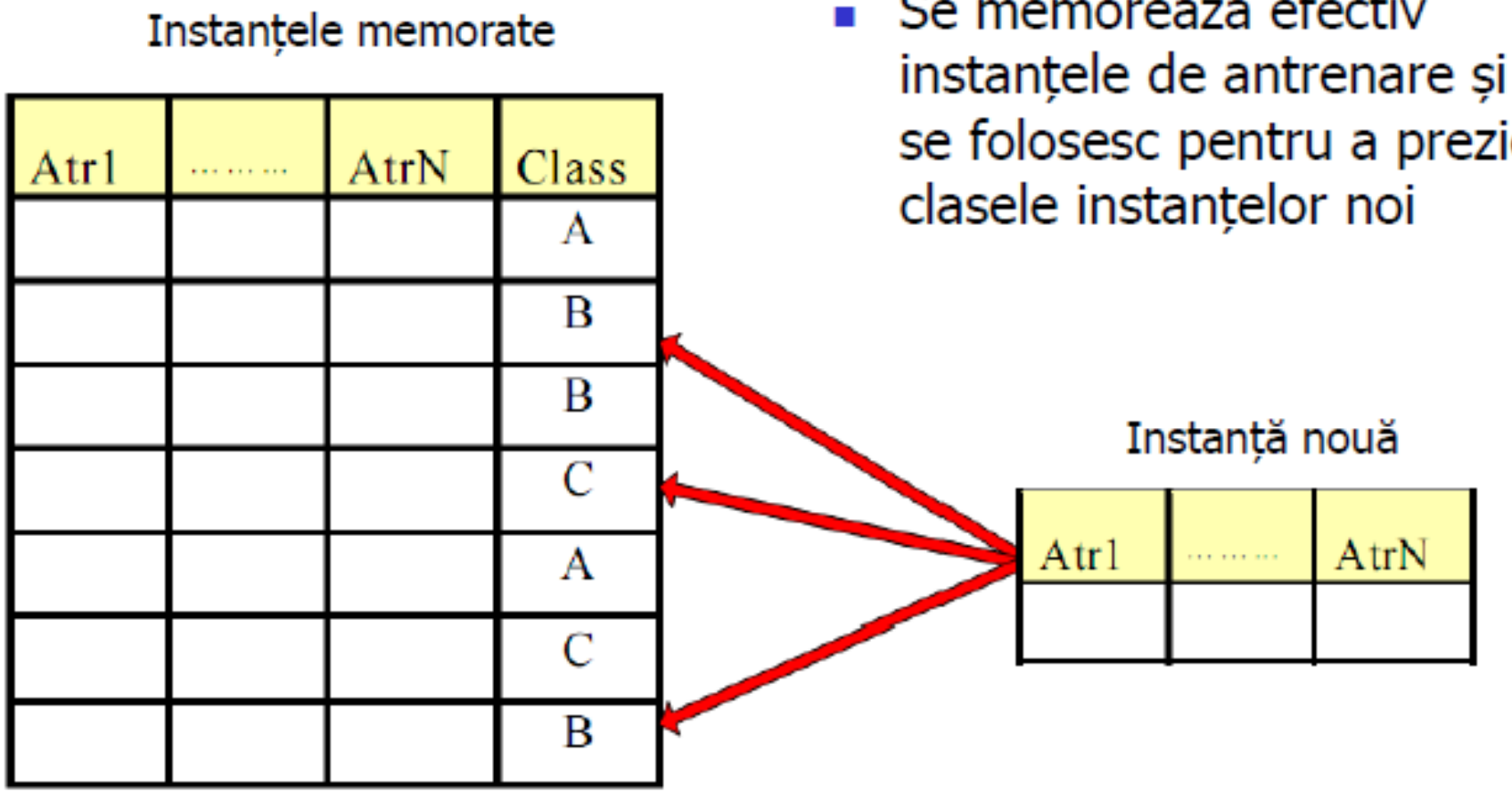
Algoritmi de clasificare

1. Clasificarea
   1. Arbori de decizie
   2. **Învățarea bazată pe instanțe**
      - Metoda celor mai apropiaţi k vecini

Clasificarea bazată pe instanțe
Lazy learners

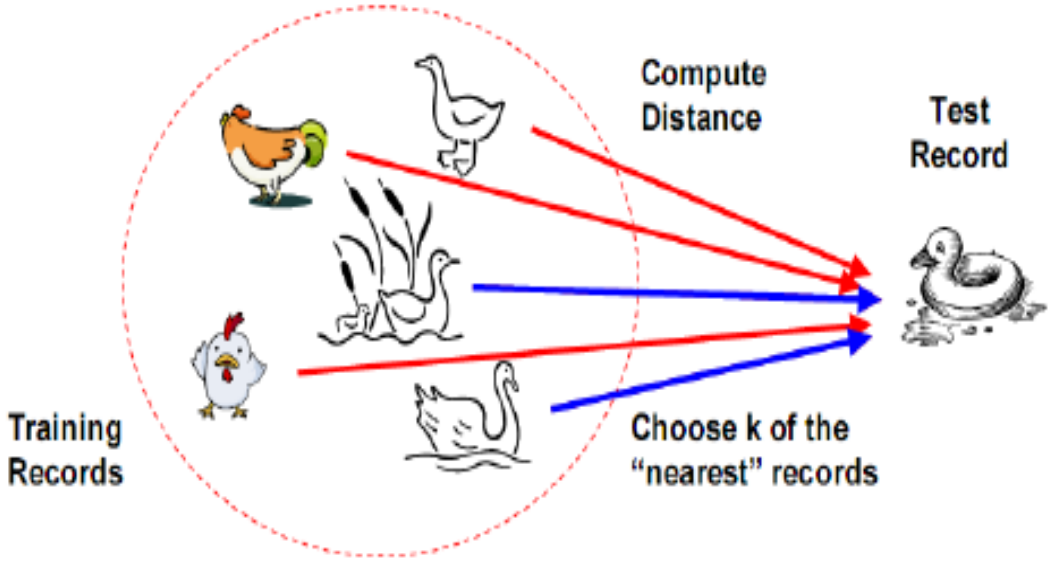
- Se memorează efectiv instanțele de antrenare și se folosesc pentru a prezice clasele instanțelor noi



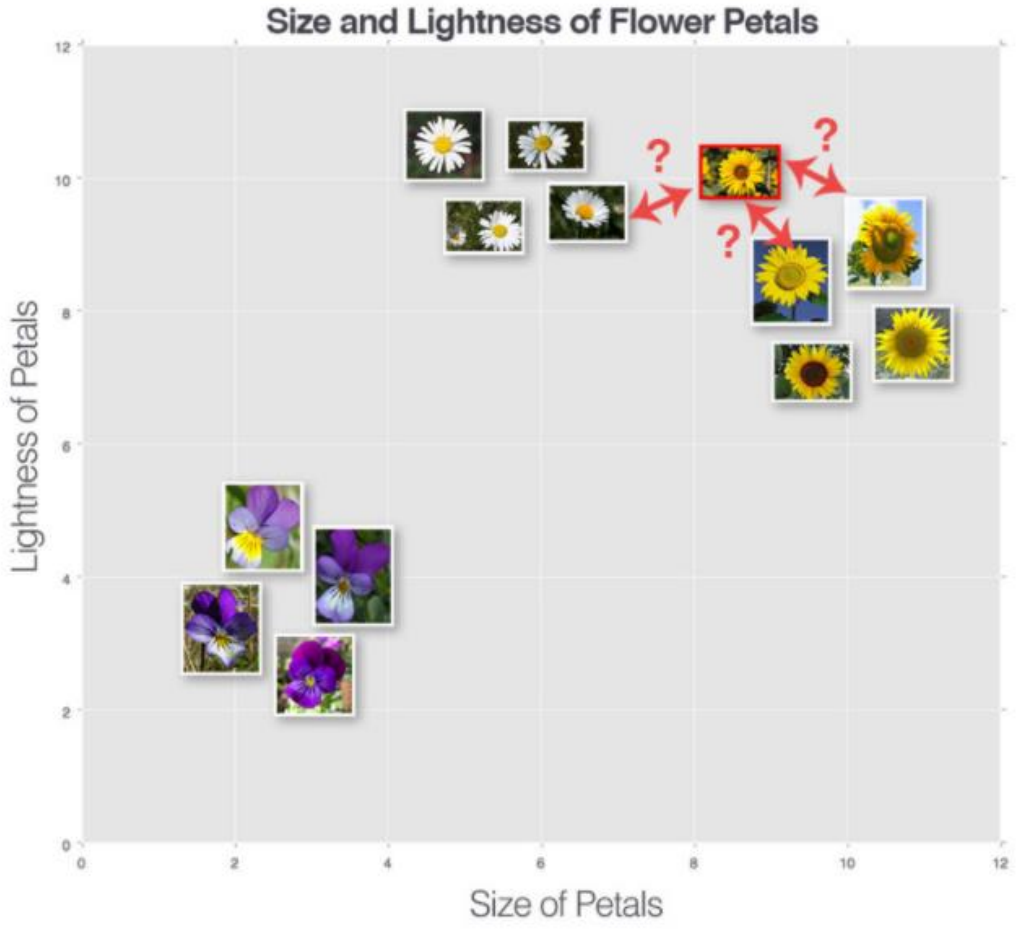
Metoda celor mai apropiați *k* vecini

- *k-Nearest Neighbors*, prescurtat: k-NN
  - Se folosesc cele mai apropiate k instanțe pentru a realiza clasificarea

1. Cum se reprezinta instanţele
2. Cum se determină similitudinile
3. Cum se alege numărul de vecini
4. Cum se realizează clasificarea

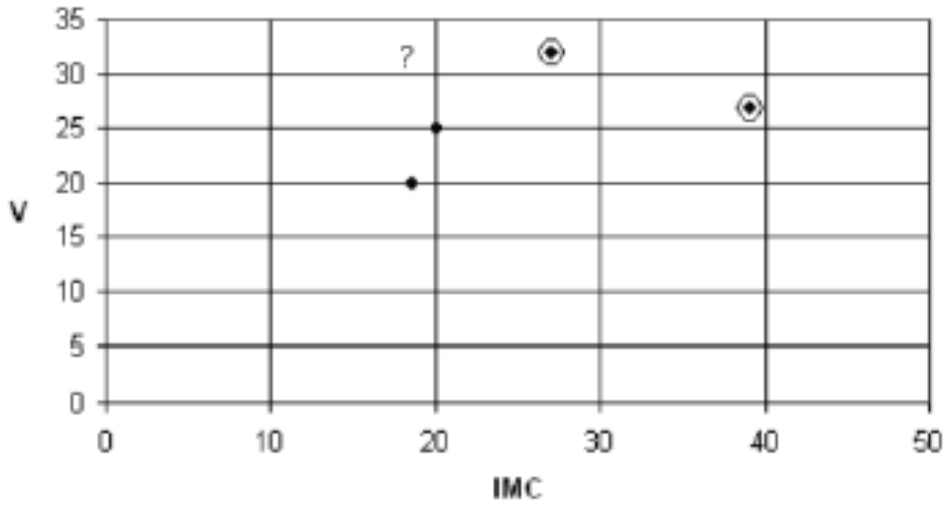


Metoda celor mai apropiați *k* vecini
Spune-mi cine iti sunt vecinii ca sa-ti spun cine esti



1. Reprezentarea instanţelor

- Pentru *n* atribute, instanțele pot fi văzute ca puncte într-un spațiu *n*-dimensional
- De exemplu, clasificarea riscului unor pacienți
- Atribute:
  - Indicele masei corporale *IMC* (= G / I^2)
  - Vârsta *V*
- Instanțe:
  - *IMC* = 18.5, *V* = 20
  - *IMC* = 27, *V* = 32
  - *IMC* = 39, *V* = 27
  - *IMC* = 20, *V* = 25

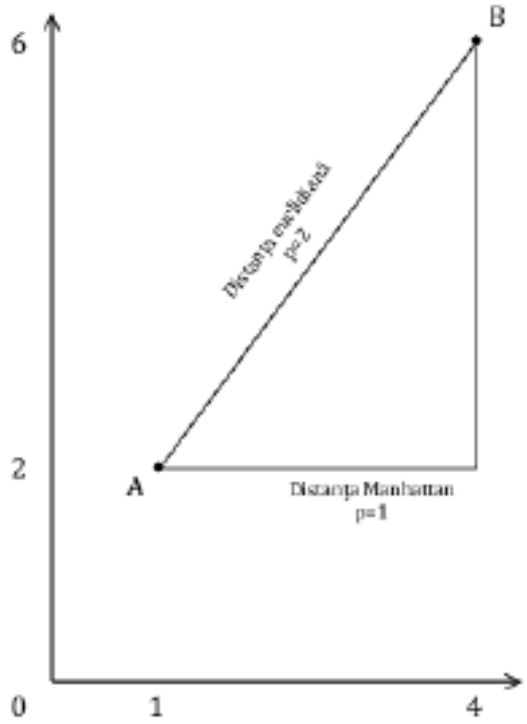


2. Calulcularea similitudinii/distantei Metrici pentru calcularea distanței

- Se folosesc în general particularizări ale distanței Minkowski

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

- Cele mai folosite metrici sunt:
  - Distanța euclidiană: $p = 2$
  - Distanța Manhattan: $p = 1$



2. Cum se calculeaza similitudinile Distanţa euclidiana

x = (x1, . . . , xn), y = (y1, . . . , yn),

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Definim *distanţa euclidiană dintre* $x, y \in \mathbb{R}^k$, astfel:

$$d(x, y) = ||x - y||(= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_k - y_k)^2}).$$

Observaţie. Pentru $k = 1$, $d(x, y) = \sqrt{(x - y)^2} = |x - y|$;
Pentru $k = 2$, $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$;
Pentru $k = 3$, $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$.

2. Calcularea similitudinii
Scalarea

Problemă legată de ordinul de mărime al datelor:

- avem două atribute: înălţimea şi greutatea unor persoane
- înălţimea e măsurată în metri; intervalul poate fi de ex [1.50 m, 2.00 m], deci cu o diferenţă de maxim 0.5
- greutatea se măsoară în kilograme; intervalul poate fi [50 kg, 200 kg]
- diferenţele de greutate domină pe cele în înălţime; o diferenţă de 1 kg este mai mare decât orice diferenţă de înălţime, contribuind deci prea mult la calculul distanţei
- Soluție: scalarea mărimilor

2. Calcularea similitudinii
Scalarea

- Se recomandă scalarea atributelor pentru a preveni dominarea măsurii de distanță de către un anumit atribut
- De exemplu:
  - Înălțimea unei persoane $\in$ [1.5, 2.1] m
  - Greutatea unei persoane $\in$ [50, 120] kg
  - Venitul unei persoane $\in$ [20000, 1000000] lei/an
- Valorile atributelor sunt normalizate:

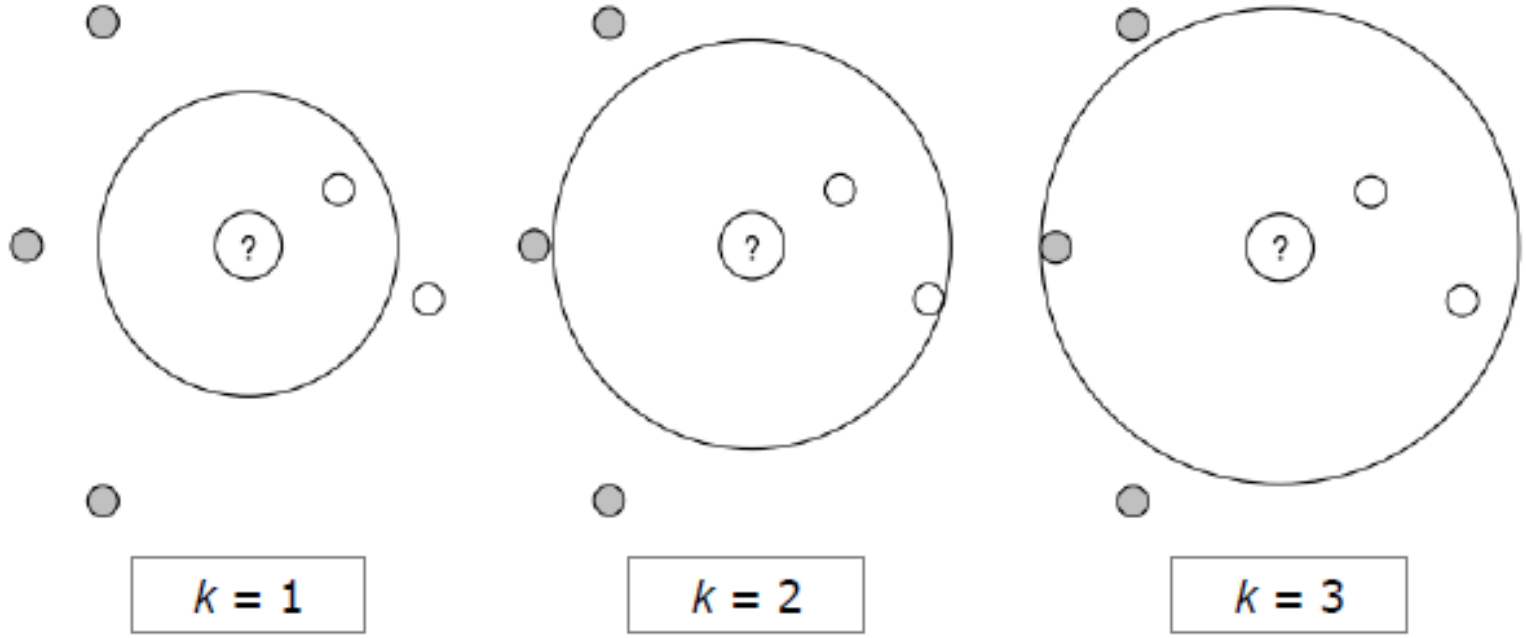
$$x_i' = \frac{(x_i - min_i)}{(max_i - min_i)} \in [0,1]$$

2. Calcularea similitudinii
Metrici pentru calcularea distanței.
Atribute nominale

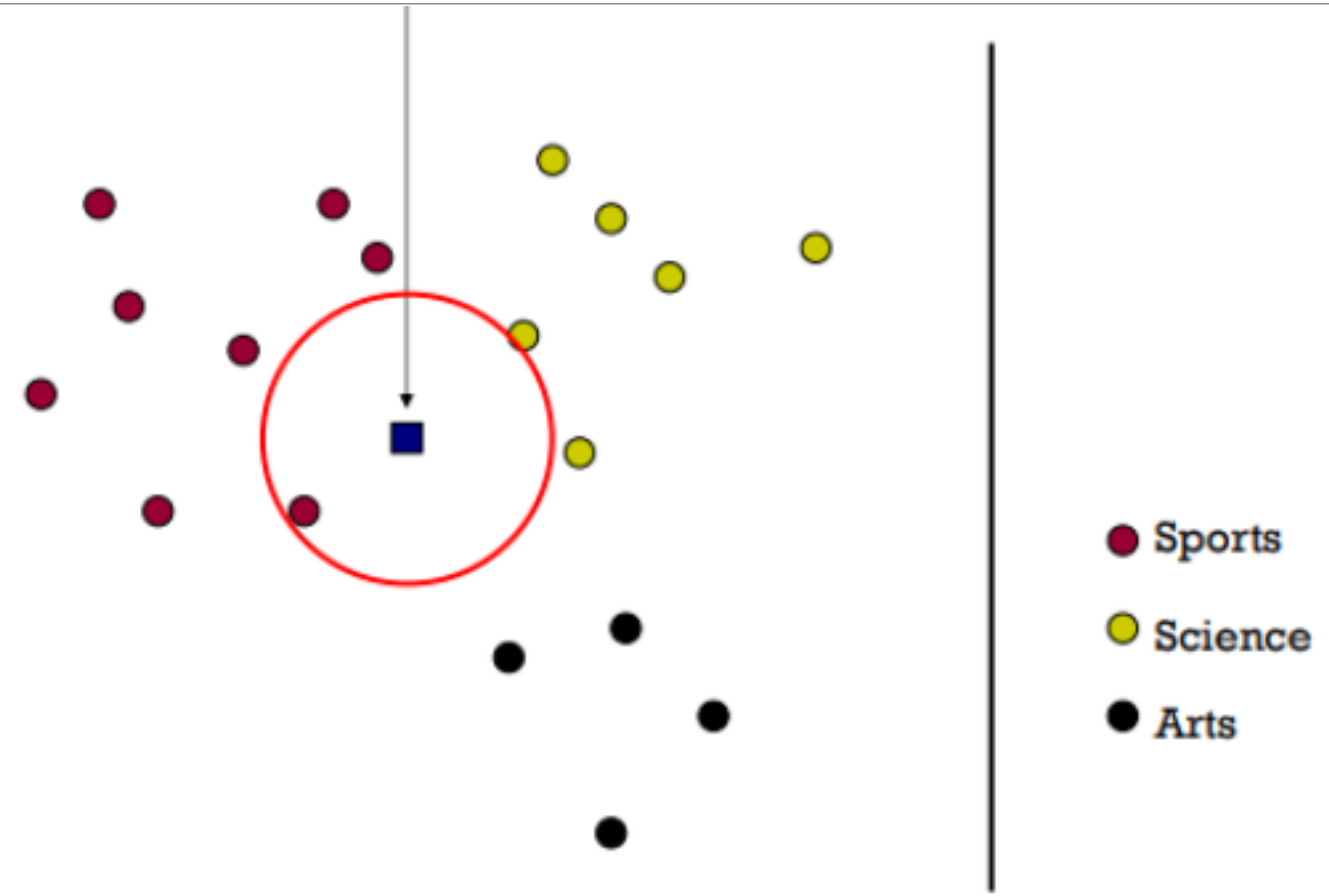
- Este necesară gasirea unei “distanţe” între valorile diferite ale atributelor nominale
  - Ex: distanţa dintre valorile: roşu, galben şi verde
- De obicei, se consideră distanţa zero pentru valori identice şi unu în caz contrar
  - Ex. având mai multe culori se poate utiliza o măsură metrică a nuanţelor din spaţiul culorilor, punând galbenul mai apropiat de portocaliu decât verde.
- Unele atribute au o importanţă diferită care ese reflectată în distanţa metrică cu ajutorul anumitor ponderi

3. Cum se alege numărul de vecini
Cei mai apropiați vecini

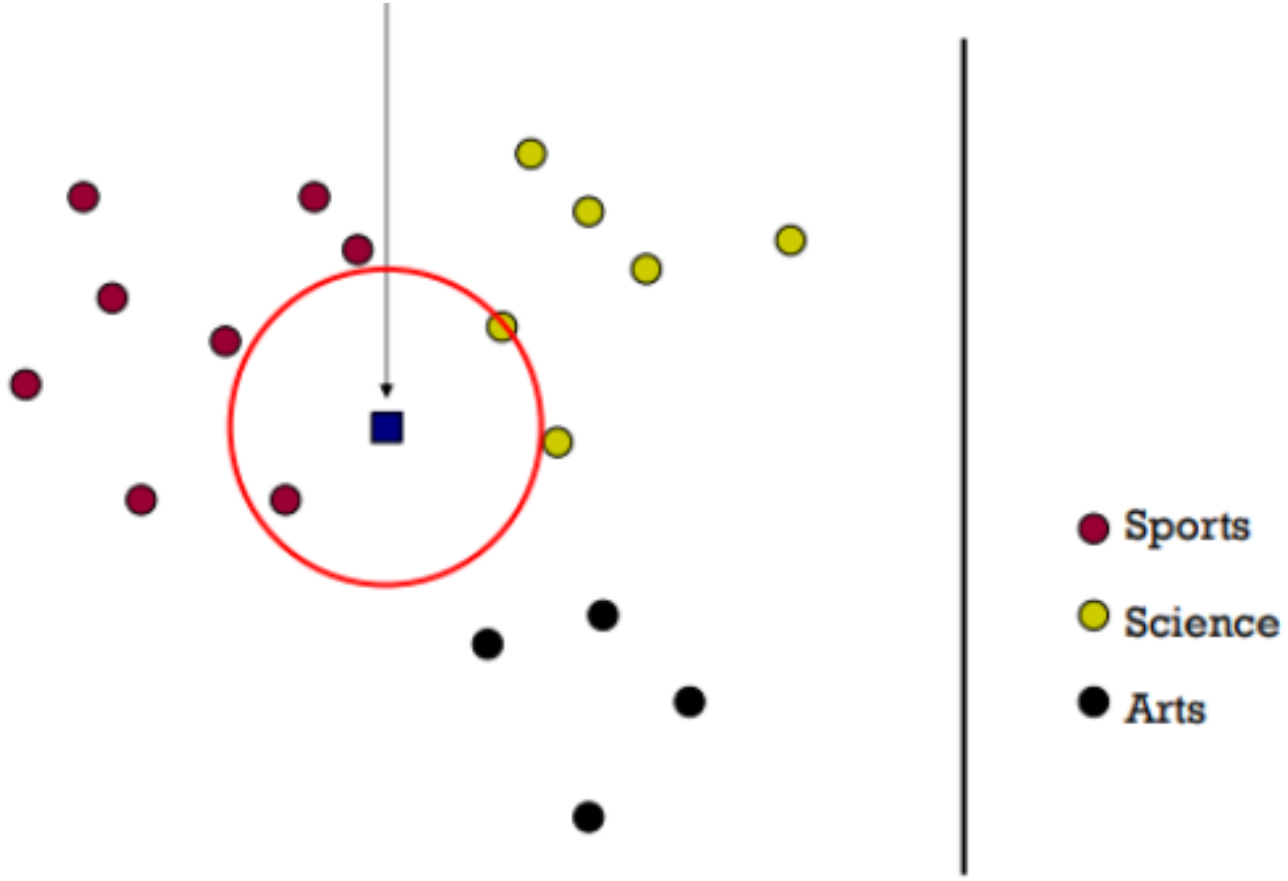
Cei mai apropiați *k* vecini ai unei instanțe *x* sunt punctele cu distanțele cele mai mici față de *x*



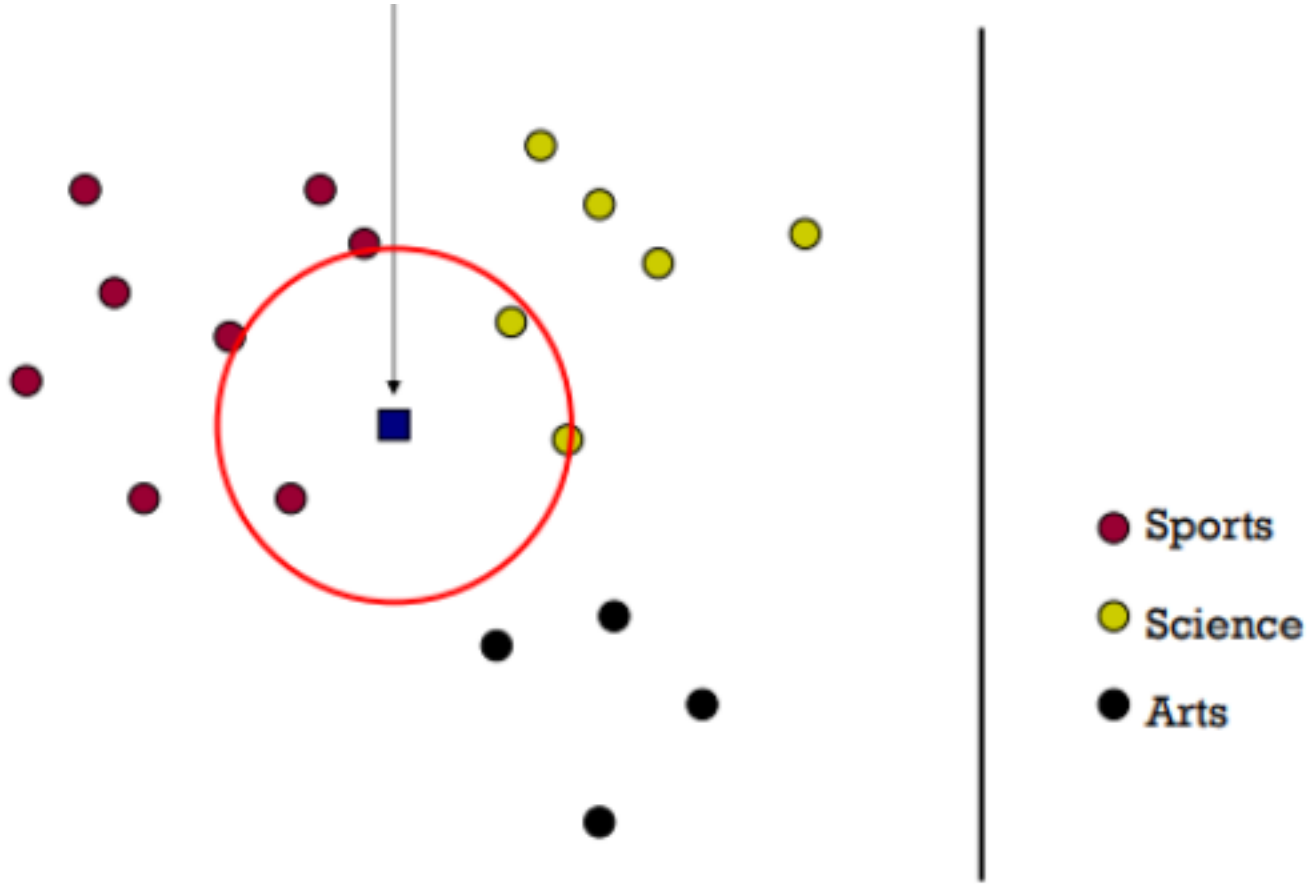
3. Cum se alege numărul de vecini
1-NN



3. Cum se alege numărul de vecini
2-NN



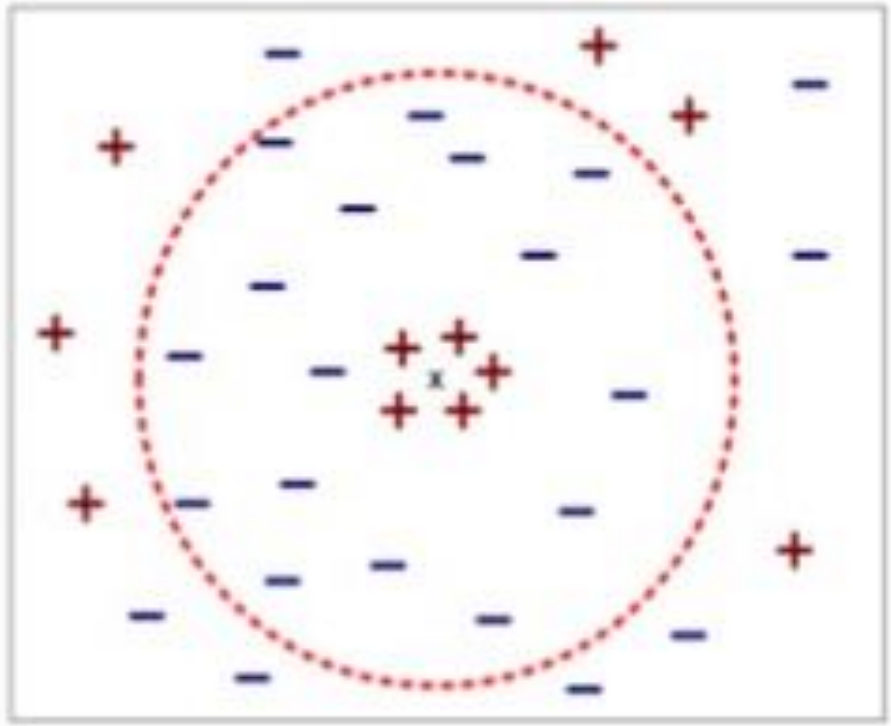
3. Cum se alege numărul de vecini
3-NN



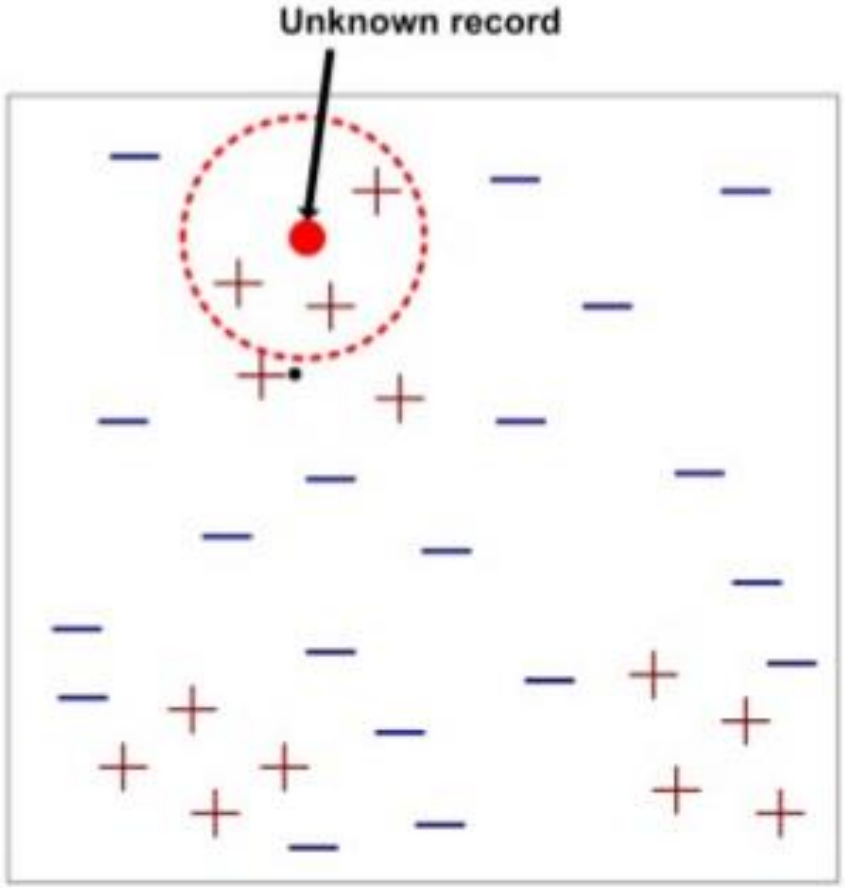
3. Cum se alege numărul de vecini
Numărul de vecini

Valoarea lui k este importantă:

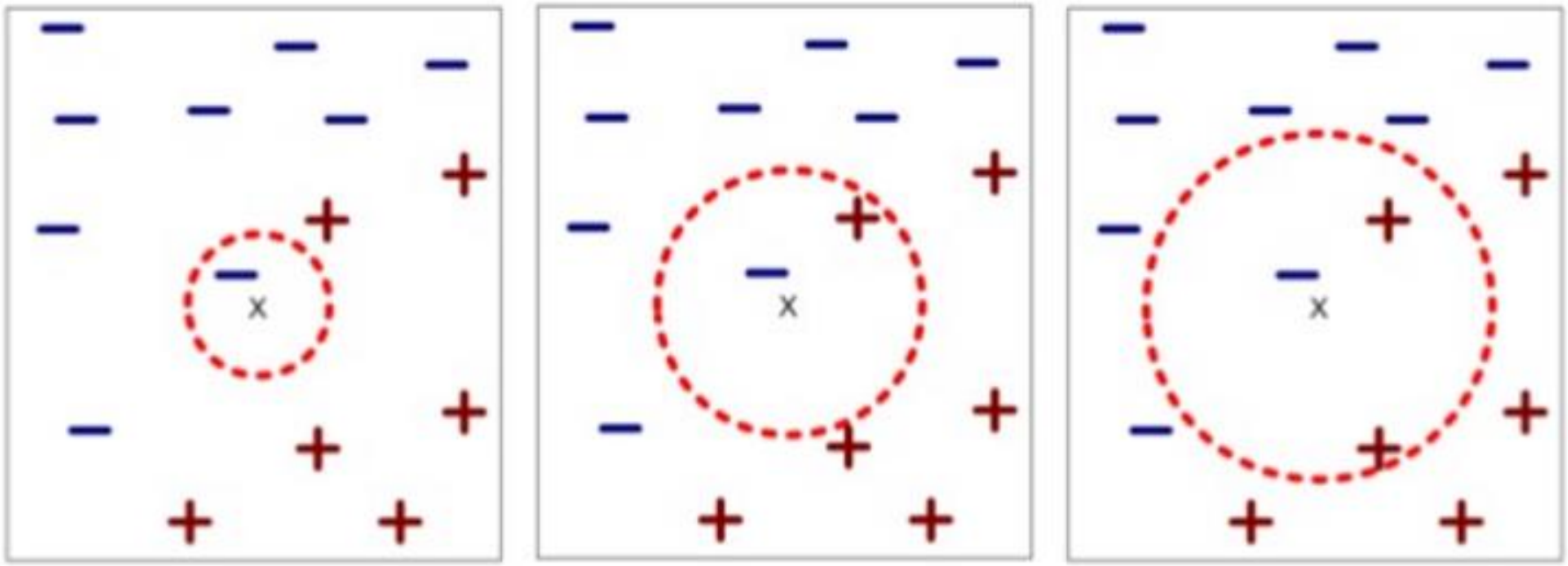
- dacă e prea mic, atunci clasificatorul poate fi suspectat de overfitting, pentru că devine prea senzitiv la zgomotul din datele de intrare (zgomot ⇒ date eronate); clasificarea poate fi afectată de zgomot
- dacă e prea mare, atunci s–ar putea ca prea mulţi dintre cei k vecini consideraţi să fie depărtaţi de punctul curent şi deci irelevanţi pentru clasificarea curentă; vecinătatea poate include puncte din alte clase



3. Cum se alege numărul de vecini
Exemplu: se consideră 3 vecini

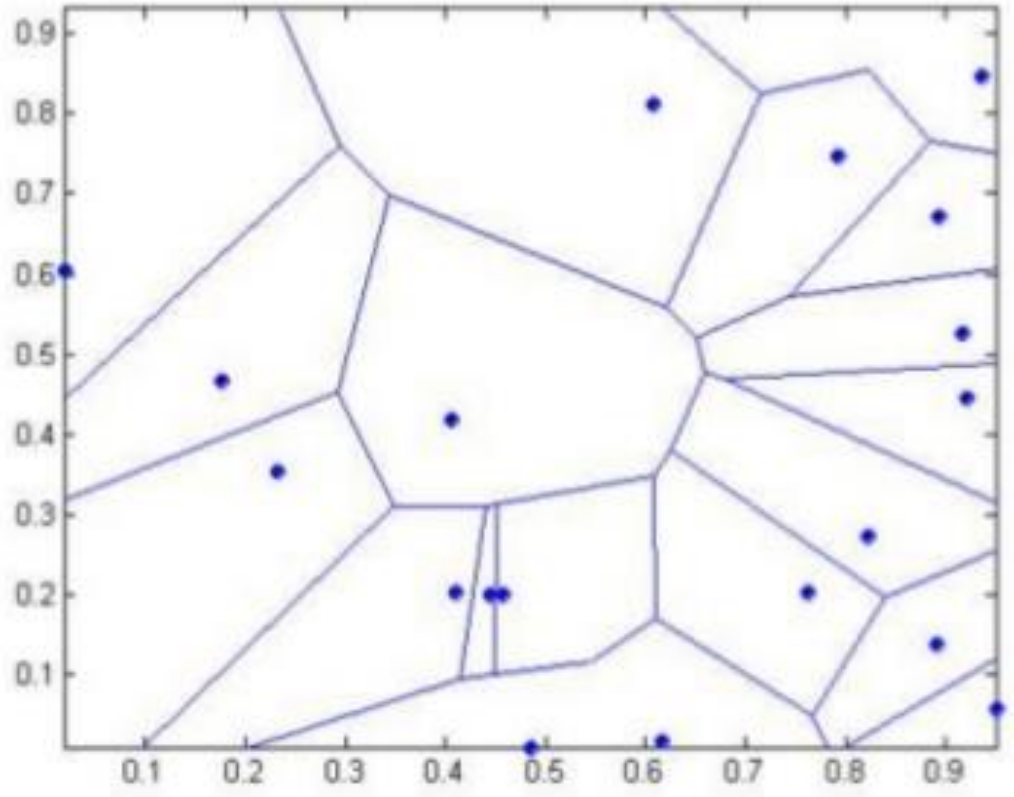


3. Cum se alege numărul de vecini
Metoda celor mai apropiați *k* vecini



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

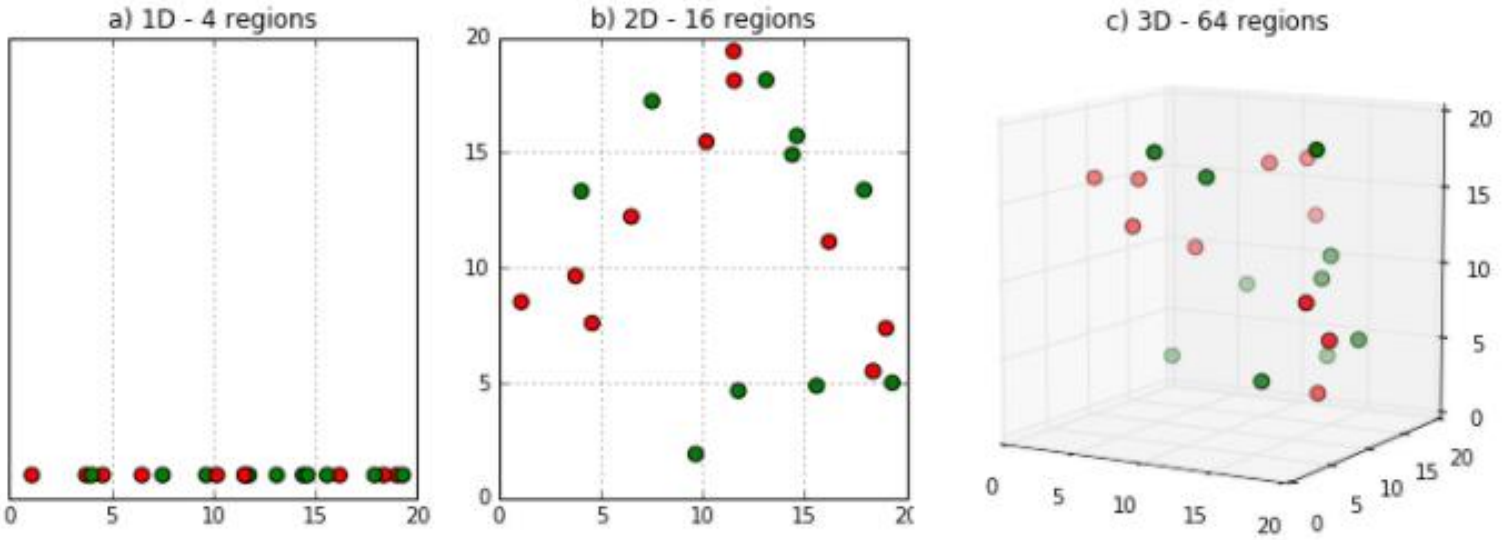
3. Cum se alege numărul de vecini 1-NN Diagrama Voronoi



- Pentru k=1 (Nearest neighbor) se obţine diagrama Voronoi.
- În interiorul unei zone delimitate, orice punct are aceeaşi clasă ca punctul marcat din acea zonă.

Dimensionalitatea datelor

- Blestemul dimensionalităţii engl. “curse of dimensionality”
- Datele devin mai rare în spațiile multi-dimensionale
- Dacă numărul de atribute este mare, este nevoie de mai multe instanțe de antrenare pentru a forma un model corect



Algoritm
Metoda celor mai apropiați *k* vecini

Fie k numărul de vecini consideraţi şi D setul de date de antrenare

Pentru fiecare exemplu de test x'

- Calculează d(x, x'), distanţa între exemplul de test şi datele (x, ·) din D
- Selectează Dz ⊆ D setul celor mai apropiaţi k vecini ai lui x'
- Calculează clasa asociată lui x'

$$y' = \arg\max_{v} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$$

Observaţii:

- funcţia I(·) este funcţia indicator, cu valoare 1 dacă argumentul are valoarea adevărat, 0 altfel.
- dacă există mai mulţi v care maximizează partea dreaptă a expresiei de mai sus, atunci se alege arbitrar unul din aceştia
- se foloseşte un sistem de votare în care fiecare vecin are acelaşi impact în determinarea clasei estimate

Weighted k-NN

Pentru a reduce senzitivitatea algoritmului k-NN faţă de alegerea lui k se poate folosi o ponderare a vecinilor

Toţi cei k vecini participă la decizia legată de clasa actuală, dar cu ponderi diferite:

- vecinii mai apropiaţi au pondere mai mare
- vecinii mai depărtaţi au pondere mai mică

Ponderea w poate fi descrescătoare cu distanţa faţă de punctul ce se vrea a fi clasificat

$$y' = \arg\max_v \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$$

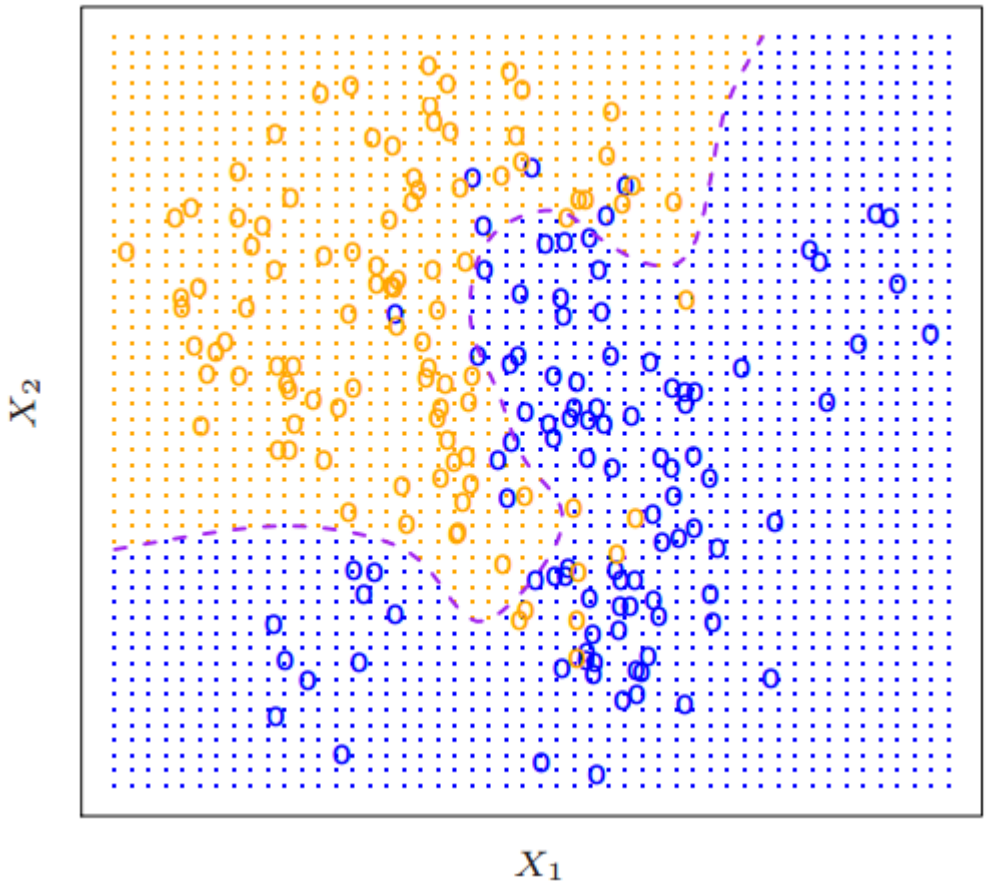
$$y' = \arg\max_v \sum_{(\mathbf{x}_i, y_i) \in D_z} w_i \times I(v = y_i)$$

Caracteristici ale metodei k-NN

- Un tip particular de “instance-based learning”
- Nu produce efectiv un model; timpul de învăţare este 0
- Clasificarea poate fi lentă, deoarece se face uz de tot corpusul de date din setul de instruire
- Clasificarea se face pe baza informaţiei locale (alți clasificatori folosesc un model global)
- k-NN poate produce suprafeţe de decizie arbitrar de complexe; suprafeţele pot avea variabilitate mare, puternic influenţate de setul de instruire
- Dacă nu se foloseşte preprocesare (scalarea diferitelor atribute ar trebui luată în considerare) sau o măsură de similaritate adecvată, valorile prezise pot fi greşite

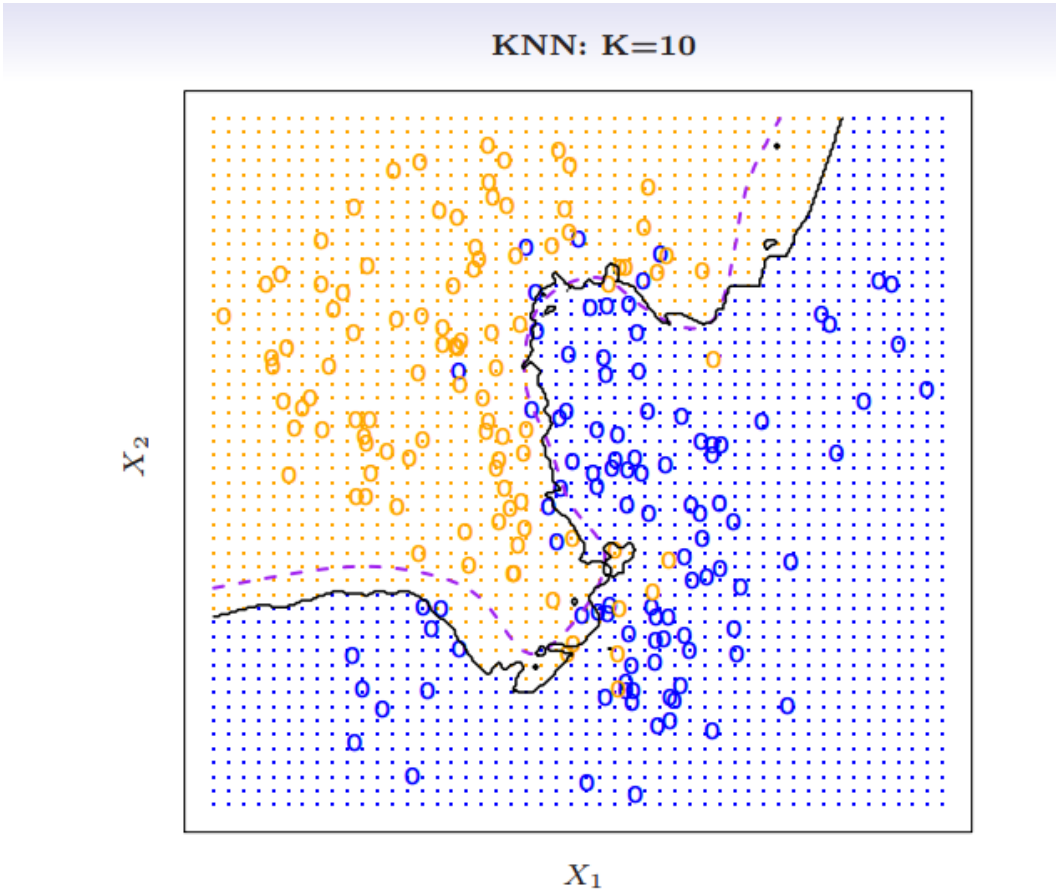
Exemplu

- 2 clase (galben si albastru)
- Decision boundary: linia punctata



Exemplu

Decizia optima: k=10



Exemplu