HOWEST TOEGEPASTE INFORMATICA, 2022-2023, © BRIAN BAERT

DATA ANALYTICS

HOOFDSTUK 6 (DEEL 3)
-KNN
-PERFORMANTIE VAN CLASSIFIERS

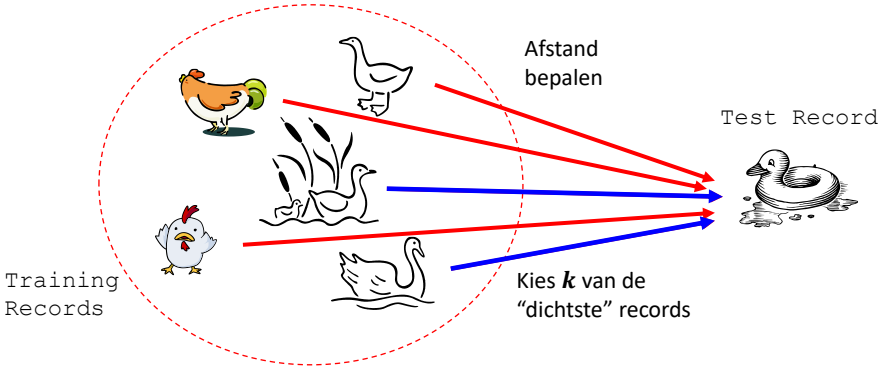
howest.be

1

NEAREST NEIGHBOR CLASSIFIER

Zie video

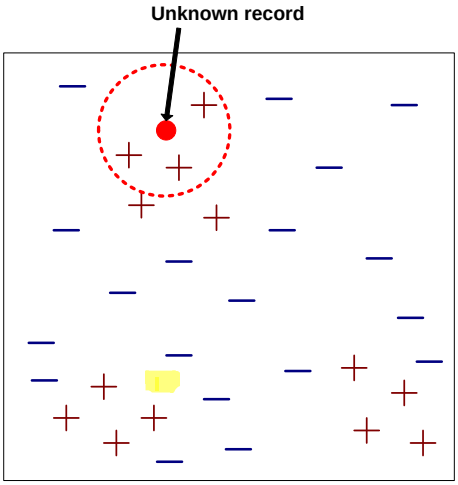
- **Basisidee**
  "If it walks like a duck, quacks like a duck, then it's probably a duck



2

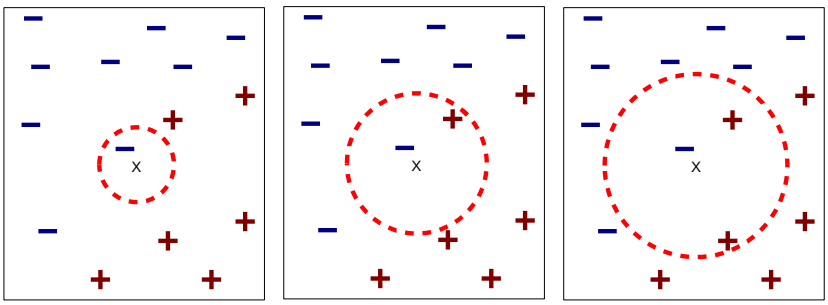
NEAREST NEIGHBOR CLASSIFIER

- Vereist drie dingen
  - De verzameling van **opgeslagen records**
  - **Afstandsmaat** om de afstand tussen records te berekenen
  - De waarde van ***k***, het aantal dichtste buren
- Om een nieuw record (unknown record) te classificeren:
  - Bereken afstand tot andere records
  - Identificeer de *k* dichtste buren
  - Maak gebruik van de class labels van de dichtste buren om de class te bepalen van het nieuwe record (b.v. door majority voting)



3

DEFINITIE "NEAREST NEIGHBOR CLASSIFIER"



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

k dichtste buren van een record $\vec{x}$ zijn punten die de ***k*** kleinste afstand hebben tot $\vec{x}$.

4

NEAREST NEIGHBOR CLASSIFIER

- Bereken de afstand tussen twee punten (records)
  → Euclidische afstand

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

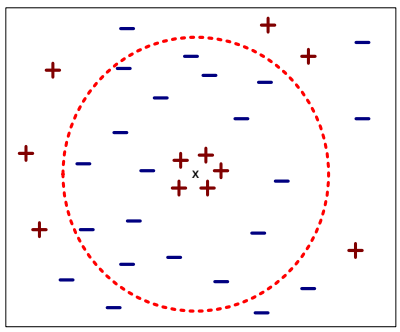
- Bepaal vervolgens de class van de dichtste burenlijst
  → Neem de *meerderheidsstemming* van class labels tussen de dichtste buren (majority voting)
  → Weeg de stemming t.o.v. de afstand met wegingsfactor

$$w = \frac{1}{d^2}$$

5

NEAREST NEIGHBOR CLASSIFIER

- Kiezen van de juiste waarde voor *k*
  → Als *k* te klein is, gevoelig aan ruis
  → Als *k* te groot is, de omgeving kan dan punten van andere klassen bevatten.



6

NEAREST NEIGHBOR CLASSIFIER

- Schaalbaarheidsissues
  - → Attributen kunnen geschaald zijn om te voorkomen dat afstandsmaten gedomineerd worden door één of meerdere attributen.
  - → Voorbeelden:
    - → Grootte van een persoon kan variëren tussen 1.5m en 1.9m
    - → Gewicht van een persoon kan variëren tussen 50kg en 150kg
    - → Inkomen van een persoon kan variëren tussen $10K en $1M

7

PROBLEEM MET EUCLIDISCHE AFSTANDSMAAT

- Bij hoge dimensionaliteit
- Tegenstrijdige uitkomsten

1 1 1 1 1 1 1 1 1 1 1 0		1 0 0 0 0 0 0 0 0 0 0 0	
0 1 1 1 1 1 1 1 1 1 1 1	VS	0 0 0 0 0 0 0 0 0 0 0 1	
d = 1.4142		d = 1.4142	

Oplossing: Normaliseren (z-score berekenen) van de vectoren tot eenheidslengte

8

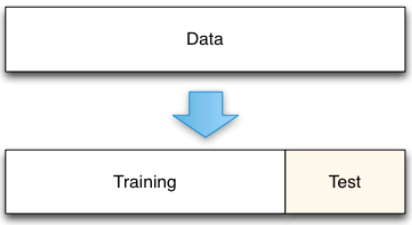
OEFENINGENREEKS

- 6.5.5
  1 en 2

9

EVALUATIEMETHODEN

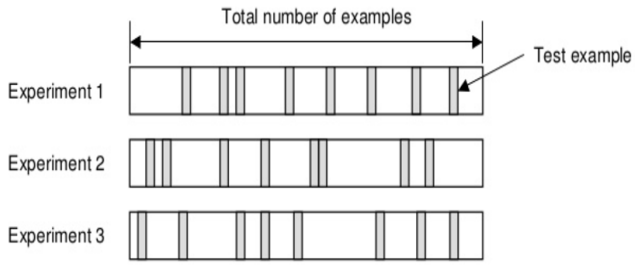
- Holdout
  dataset wordt opgesplitst in twee disjuncte verzamelingen



10

EVALUATIEMETHODEN

- Random subsampling
  holdout meerdere keren na elkaar

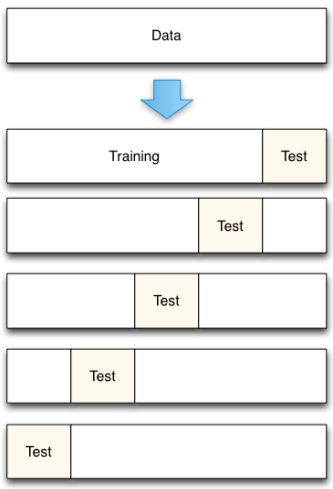


11

EVALUATIEMETHODEN

- Cross-validation
  Elk record wordt evenveel keer gebruikt voor training en precies één keer voor test set.

  5-fold cross-validation →



12

EVALUATIEMETHODEN

Zie video op Leho!

- Confusion matrix

		Voorspelde klasse	
		Class=Yes	Class=No
Werkelijke klasse	Class=Yes	(TP)	(FN)
	Class=No	(FP)	(TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \qquad \text{F}_1 = \frac{2 * TP}{2 * TP + FP + FN} \qquad \text{recall} = \frac{TP}{TP + FN}$$

$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{precision} = \frac{TP}{TP + FP} \qquad \text{FPR} = \frac{FP}{FP + TN}$$

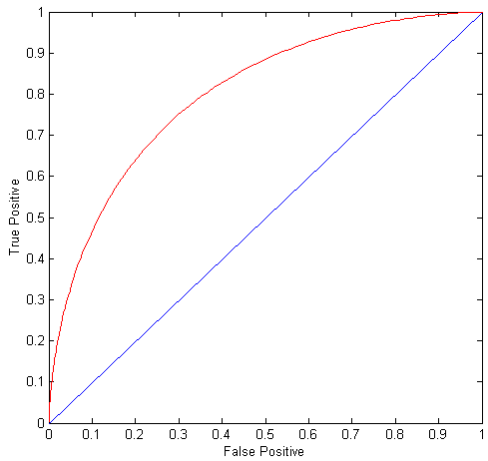
13

ROC CURVE

(TP, FP)

- (0,0): alles behoort tot de negatieve klasse
- (1,1): alles behoort tot de positieve klasse
- (0,1): ideale wereld

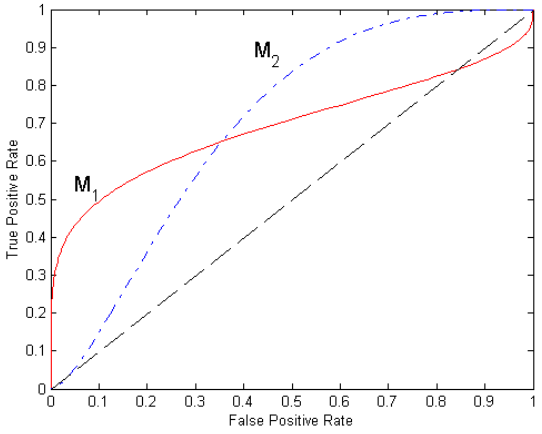
- Diagonaal
  → Willekeurig gokken
  → Onder de diagonaal
  → voorspelling is tegengestelde van de positieve klasse



14

ROC CURVE GEBRUIKEN - VOORBEELD

- Geen enkel model is consequent beter dan de andere
  - M_1 is beter voor kleine FPR
  - M_2 is beter voor grote FPR
- Oppervlakte onder de ROC-curve
  - Ideaal:
    - Opp = 1
  - Willekeurig gokken:
    - Opp = 0.5



15

OEFENINGENREEKS

- 6.5.5
  3 en 4

16