Oplossingen 6.5.5

Oefening 1

Gegeven volgende één-dimensionale data set

x	0,5	3,0	4,5	4,6	4,9	5,2	5,3	5,5	7,0	9,5
y	-	-	+	+	+	-	-	+	-	-

Classificeer het datapunt $x = 5{,}0$ op basis van de 1-, 3-, 5- en 9-nearest neighbors (maak gebruik van majority voting).

Oplossing oefening 1

x	**y**	**D(x, x=5,0)**
0,5	-	|5,0-0,5| = 4,5
3,0	-	|5,0-3,0| = 2,0
4,5	+	|5,0-4,5| = 0,5
4,6	+	|5,0-4,6| = 0,4
4,9	+	|5,0-4,9| = 0,1
5,2	-	|5,0-5,2| = 0,2
5,3	-	|5,0-5,3| = 0,3
5,5	+	|5,0-5,5| = 0,5
7,0	-	|5,0-7,0| = 2,0
9,5	-	|5,0-9,5| = 4,5

1-NN: y=+

3-NN: y=-

5-NN: y=+

9-NN: y=-

Zie Excel bestand voor detailuitwerking.

Oefening 2

Maak gebruik van onderstaande dataset en de nearest neighbour classifier met k=1 om voor onderstaande testrecords de waarde te bepalen van de klasse "GOOD SURF" te bepalen

ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

Veronderstel dat het model gebruik maakt van de Euclidische afstand om de dichtste buren te vinden, welke voorspelling geeft het model dan voor volgende nieuwe records?

ID	WAVE SIZE (FT)	WAVE PERIOD (SECS)	WIND SPEED (MPH)	GOOD SURF
Q1	8	15	2	?
Q2	8	2	18	?
Q3	6	11	4	?

Oplossing oefening 2

ID	Wave size (ft)	Wave period (secs)	Wind speed (mph)	Good surf	distance to Q1	distance to Q2	distance to Q3
1	6	15	5	yes	3,61	18,49	4,12
2	1	6	9	no	13,38	12,08	8,66
3	7	10	4	yes	5,48	16,16	**1,41**
4	7	12	3	yes	**3,32**	18,06	1,73
5	2	2	10	no	16,40	10,00	11,53
6	10	2	20	no	22,29	**2,83**	18,79

Q1	8	15	2	**yes**
Q2	8	2	18	**no**
Q3	6	11	4	**yes**

Zie Excel bestand voor detailuitwerking.

Oefening 3

In volgende tabel vind je voorspellingen terug (`Prediction`) voor een bepaalde categorie (`Target`) op basis van een model. Bepaal de gevraagde performantiematen.

ID	Target	Prediction	ID	Target	Prediction	ID	Target	Prediction
1	false	false	8	true	true	15	false	false
2	false	false	9	false	false	16	false	false
3	false	false	10	false	false	17	true	false
4	false	false	11	false	false	18	true	true
5	true	true	12	true	true	19	true	true
6	false	false	13	false	false	20	true	true
7	true	true	14	true	true			

a. Confusion matrix en misclassificatieratio (aantal verkeerd geclassificeerde t.o.v. totaal aantal geclassificeerde records).
b. Precision, recall en f_1-maat.

Oplossing oefening 3

a)

		Prediction	
		True	**False**
Target	**True**	8	1
	False	0	11

De misclassificatieratio bedraagt dan:

$$\frac{FP + FN}{TP + TN + FP + FN} = \frac{0 + 1}{8 + 11 + 0 + 1} = 0{,}05$$

b)

$$\text{precision} = \frac{TP}{TP + FP} = \frac{8}{8 + 0} = 1$$

$$\text{recall} = \frac{TP}{TP + FN} = \frac{8}{8 + 1} = 0{,}889$$

$$F_1 = 2 * \frac{p * r}{p + r} = 2 * \frac{1 * 0{,}889}{1 + 0{,}889} = 0{,}941$$

Oefening 4

Veronderstel dat je werkt op een spam detectiesysteem. Het probleem heb je geformuleerd als een classificatietaak waar “Spam” de positieve klasse is en “not-Spam” de negatieve klasse. Je trainingdataset bevat 1000 e-mails, 99% van deze worden als “not-Spam” geclassificeerd en 1% als “Spam”.

a. Wat is de nauwkeurigheid (accuracy) van een classifier die altijd “not-Spam” geeft?
b. De fractie van spam e-mails die correct als “Spam” geclassificeerd worden, wordt gemeten door de “recall” waarde. Bereken de recall van de classifier uit (a).

Oplossing oefening 4

a)

$$\text{accuracy} = \frac{990}{1000} = 99\%$$

b)

$$\text{recall} = \frac{0}{0 + 10} = 0$$