Übung 9 Bioinformatik

6.7.2022

1. Training Data

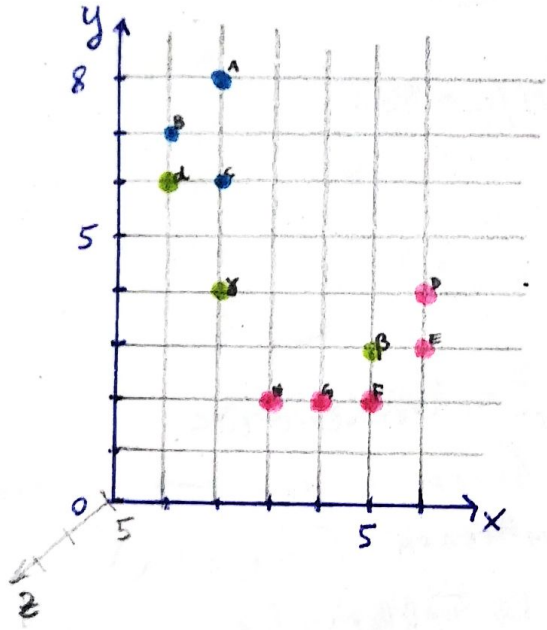
Tumorgewebe:

	A	B	C
x	2	1	2
y	8	7	6
z	5	5	5

gesundes Gewebe:

	D	E	F	G	H
x	6	6	5	4	3
y	4	3	2	2	2
z	5	5	5	5	5

dreidimensionale Werte/Punkte sind gegeben, dabei handelt es sich eigentlich um 2 dimensionale Werte, da der z-Wert nicht variiert → alle liegen in Ebene → 2D



Vorhersagen für

	α	β	γ
x	1	5	2
y	6	3	4
z	5	5	5

k nearest neighbours mit geringstem Abstand finden → die Eigenschaft/Merkmal das überwiegt gilt auch ~~für~~ d, β, γ

d

k = 1

(α) (n.n) nearest neighbours: B oder C (eindeutig) jeweils Abstand = 1
beide Tumor (Tu) → α = Tu

(β) n.n: E oder F (eindeutig erkennbar) $d_{\beta,B}/d_{\beta,C} = 1$
~~beide Tu → β = gG~~ beide gesundes Gewebe (gG) → β = gG

(γ) n.n: C oder H $d_{\gamma,C} = \sqrt{0^2+2^2} = 2$
$d_{\gamma,H} = \sqrt{1^2+2^2} = \sqrt{5}$
→ $d_{\gamma,C} < d_{\gamma,H}$ → γ = Tu

6.7.2022

Übung 9 Bioinformatik seite 2

k=2

(α) n.n = B & C (siehe k=1) → α = Tu

(β) n.n = E & F (") → β = gG

(γ) n.n = C & H $d_{\gamma,C} = \sqrt{4} = 2$ $d_{\gamma,H} = \sqrt{5}$ → γ = ?

k=3

(α) n.n = A, B, C → α = Tu

(β) n.n = E, F, D/G (alles noch sehr eindeutig) → β = gG

(γ) n.n = C & H + B oder G (?) $d_{\gamma,G} = \sqrt{2^2+2^2} = \sqrt{8}$ $d_{\gamma,B} = \sqrt{3^2+1^2} = \sqrt{10}$ $d_{\gamma,G} < d_{\gamma,B}$ → γ = gG

k=4

(α) n.n = A, B, C, H → α = Tu

(β) n.n = D, E, F, G → β = gG

(γ) n.n = C, H, B, G → γ = ?

k=5

(α) n.n = A, B, C, H & G oder D

$d_{\alpha,G} = \sqrt{3^2+4^2} = \sqrt{25}$

$d_{\alpha,D} = \sqrt{5^2+2^2} = \sqrt{29}$

$d_{\alpha,G} < d_{\alpha,D}$ → α = Tu

(β) n.n = D, E, F, G, H β → gG

(γ) n.n = C, H, B, G & A/D oder F

$d_{\gamma,D} = 4$

$d_{\gamma,F} = \sqrt{3^2+2^2} = \sqrt{13}$

$d_{\gamma,D} > d_{\gamma,F}$ → γ = gG

k=6

(α) n.n = A, B, C + 3 g.G → α = ?

(β) n.n. = D, E, F, G, H + irgendein Tu → β = gG

(γ) n.n = C, H, B, G, F, A/D → γ = gG oder ? (je nachdem, ob A/D)

k = 7

(α) n.n = A, B, C, 4 gG → α = gG

(β) n.n = D, E, F, G, H, C, B → β = gG

(γ) n.n = A, B, C, D, E, G, H → γ = gG

Bewertung:

- ~~ab~~ bei k = 6 kann es nur noch zu einem „Unentschieden" zwischen gG & Tu kommen oder zum Überwiegen von gG, da es nur 3 Trainingsdaten zu Tu gab
- ab k = 7 ~~kommt~~ kommt man ausschließlich zu dem Ergebnis, dass die nearest neighbours gesund ist, da diese Mehrheit der zahlenmäßig überwiegen

⇒ es kann bei zu hohen ks zu Fehlinterpretationen kommen (wenn diese über der ~~doppelten~~ zweifachen Größe der kleinsten Trainigsdatengruppe liegen)

⇒ ~~ks~~ k muss/sollte passend zur Anzahl der Trainingsdatenpunkte gewählt werden

⇒ zudem ungerade, da es bei geraden ks zu nicht aussagekräftigen Ergebnissen kommen kann (vgl. ~~β~~ (k=2), γ (k=4), γ (k=6), α (k=6)

2. $p(x) = \frac{1}{1 + e^{(-\beta_0 + \beta_1 x)}} = \frac{1}{1+e^{-z}}$ $\beta_0 = -5$ $\beta_1 = 1$

wurden mit maximum Likelyhood Methode berechnet

$$p(2) = \frac{1}{1+e^{(5+2)}} = 9{,}11 \cdot 10^{-4}$$

$$p(6) = \frac{1}{1+e^{(5+6)}} = 1{,}67 \cdot 10^{-5}$$

$$p(10) = \frac{1}{1+e^{(5+10)}} = 3{,}06 \cdot 10^{-7}$$

gleich Wahrscheinlich $\Rightarrow p = 0{,}5$

$\frac{1}{0{,}5} = 1 + e^{(5+x)} \quad |-1$

$1 = e^{(5+x)} \quad \ln$

$\ln 1 = \ln e^{(5+x)}$

$0 = 5 + x \quad |-5$

$-5 = x$

3. Funktion knn.cv

sie führt eine Kreuzvalidierung an dem gegebenen Datensatz durch

mit euklidischer Distanz