Hausaufgabe 9

① Trainingsdaten:

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

krank

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

gesund

Beachte, dass $z=5$ für alle Werte. Damit vereinfacht sich das Problem auf ein zweidimensionales:

$\alpha = (1,6)^T$: Distanzen $d_{\alpha A} = \sqrt{5}$, $d_{\alpha B} = 1$, $d_{\alpha C} = 1$, $d_{\alpha D} = \sqrt{29}$, $d_{\alpha E} = \sqrt{34}$, $d_{\alpha F} = \sqrt{32}$, $d_{\alpha G} = 5$, $d_{\alpha H} = \sqrt{20}$

1. kNN$(\alpha, k=1) = \{B\} = \{C\}$ krank
2. kNN$(\alpha, k=3) = \{A, B, C\}$ krank
3. kNN$(\alpha, k=5) = \{A, B, C, H, G\}$ krank

α stammt mit hoher Wahrscheinlichkeit aus krankem Gewebe

Für $k=7$ ist die Entscheidung immer gesund, da nur 3 Datensätze krank gelabeled sind. Deswegen hat die Berechnung dieser Menge keinen Sinn.

$\beta = (5,3)^T$: $d_{\beta A} = \sqrt{34}$, $d_{\beta B} = \sqrt{32}$, $d_{\beta C} = \sqrt{18}$, $d_{\beta D} = \sqrt{2}$, $d_{\beta E} = 1$, $d_{\beta F} = 1$, $d_{\beta G} = \sqrt{2}$, $d_{\beta H} = \sqrt{5}$

1. kNN$(\beta, k=1) = \{E\} = \{F\}$ gesund
2. kNN$(\beta, k=3) = \{E, F, D(G)\}$ gesund
3. kNN$(\beta, k=5) = \{D, E, F, G, H\}$ gesund

β stammt mit hoher Wahrscheinlichkeit aus gesundem Gewebe

$\gamma = (2,4)^T$: $d_{\gamma A} = 4$, $d_{\gamma B} = \sqrt{10}$, $d_{\gamma C} = 2$, $d_{\gamma D} = 4$, $d_{\gamma E} = \sqrt{17}$, $d_{\gamma F} = \sqrt{13}$, $d_{\gamma G} = \sqrt{8}$, $d_{\gamma H} = \sqrt{5}$

1. kNN$(\gamma, k=1) = \{C\}$ krank
2. kNN$(\gamma, k=3) = \{C, H, G\}$ gesund
3. kNN$(\gamma, k=5) = \{C, H, G, B, F\}$ gesund

Ein Zuordnung von γ kann nur unter Vorbehalt vorgenommen werden, da die kNN Methode für unterschiedliche k andere Ergebnisse produziert.

Aufgabe 9.2 Logistische Regression (30 %)

Sie betrachten die Wahrscheinlichkeiten $p(x)$, dass ein Patient erkrankt ist. Hierfür haben Sie für verschiedene Patienten die Expression x eines bestimmten Genes ermittelt. Eine logistische Regression der funktionalen Form

$$p(x) = \frac{1}{1+\exp(-\beta_0+\beta_1 x)} \quad (4)$$

hat die Parameter $\beta_0 = -5$ und $\beta_1 = 1$ ergeben.

decision boundary bei $p(x_{db}) = 0{,}5$. Umstellen nach x:

$$\frac{1}{p} = 1 + \exp(-\beta_0 + \beta_0 x)$$

$$\Leftrightarrow \exp(-\beta_0+\beta_1 x) = \frac{1}{p} - 1$$

$$\Leftrightarrow \beta_1 x = \ln\left(\frac{1}{p}-1\right) + \beta_0$$

$$x = \frac{\ln\left(\frac{1}{p}-1\right) + \beta_0}{\beta_1}$$ und mit $p = 0{,}5$; $\beta_0 = -5$ und $\beta_1 = 1$ ist

$$x_{db} = -5$$

Weiterhin ist $p(x=2) = 0{,}911\ 10^{-3} \approx 0{,}9\ \%$

$p(x=6) = 0{,}017 \cdot 10^{-3} \approx 0{,}02\ \%$

$p(x=10) = 0{,}0003 \cdot 10^{-3} \approx 0{,}0003\ \%$