DHBW Ravensburg
Duale Hochschule Baden-Württemberg

# Prüfung

*Unsupervised Learning · Prüfung 00 Probeklausur*

| | |
|---|---|
| Studiengang | Data Science und Künstliche Intelligenz |
| Jahrgang | RV-WDS124 / RV-WDS224, Semester 4 |
| Modul | W4DSKI_201 Künstliche Intelligenz und Machine Learning |
| Veranstaltung | W4DSKI_201.1 Unsupervised Learning, Reinforcement Learning und weitere Konzepte |
| Prüfer | Prof. Dr. Mark Schutera |
| Datum | 15.06.2026 |
| Bearbeitungszeit | 30 Minuten |
| Max. Punktzahl | 25 Punkte |
| Hilfsmittel | Taschenrechner (nicht programmierfähig). |

## Anweisungen

Schreiben Sie Ihre Lösungen und Lösungswege auf die zusätzlich ausgeteilten Klausurbögen, nicht auf die Aufgabenblätter. Lösungen auf den Aufgabenblättern werden nicht gewertet. Die Heftung der Klausur bitte nicht entfernen. Ordnen Sie jede Lösung deutlich der zugehörigen Aufgabe zu. Begründen Sie Ihre Antworten kurz: eine reine Zahl ohne Rechenweg gibt keine volle Punktzahl. Schreiben Sie leserlich. Was nicht entzifferbar ist, wird nicht gewertet. Lesen Sie die Aufgabenstellung sorgfältig durch, bevor Sie mit der Bearbeitung beginnen. Lesen Sie ihre Lösungen am Ende noch einmal durch. Ich empfehle dafür, Zwischenlösungen einfach und Lösungen zweifach zu unterstreichen. Viel Erfolg!

## Achtung! Hinweis zum Umfang

Diese Probeklausur hat den halben Umfang einer regulären Klausur: Sie bearbeiten 25 Punkte in 30 Minuten anstelle der vollen 50 Punkte in 60 Minuten. Sie dient als Einstieg und deckt nur einen Teil des Prüfungsstoffs ab.

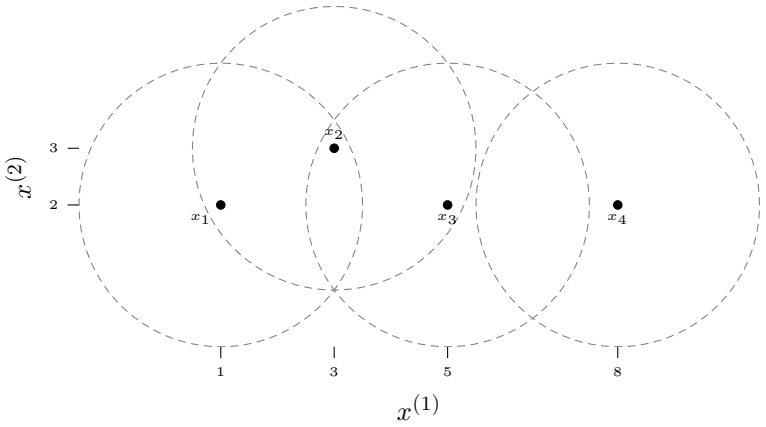
Durch Prüfer auszufüllen

| | 1 | 2 | Σ |
|---|---|---|---|
| Maximal | 15 | 10 | 25 |
| Erreicht | ____ | ____ | ____ |

## Aufgabe 1

15 Punkte

Gegeben sind vier Punkte $x_1 = (1, 2)$, $x_2 = (3, 3)$, $x_3 = (5, 2)$ und $x_4 = (8, 2)$, die Sie mit dichtebasiertem Clustern (density-based clustering) gruppieren. Verwenden Sie den festen Nachbarschaftsradius $\varepsilon = 2{,}5$ und die Dichteschwelle $n_{\min} = 2$. Die $\varepsilon$-Umgebung $N_\varepsilon(x_i)$ eines Punktes zählt hier den Punkt selbst nicht mit.




### (a)

*6 Punkte*

Bestimmen Sie für jeden der vier Punkte die Klasse. Beginnen Sie die Clusterbildung bei $x_2$ und geben Sie das resultierende Clustering an.

Koordinaten.

| | |
|---|---|
| $x_1$ | (1, 2) |
| $x_2$ | (3, 3) |
| $x_3$ | (5, 2) |
| $x_4$ | (8, 2) |

Im Algorithmus des dichtebasierten Clusterns greift ein Kernpunkt $p_c$ in seine Umgebung hinaus und nimmt dabei Punkte auf, die selbst keine Kernpunkte sind. Ein solcher Punkt $p_b$ wird zum Randpunkt, wenn

$$|N_\varepsilon(p_b)| < n_{\min} \quad \text{und} \quad p_b \in N_\varepsilon(p_c).$$

Benennen Sie das Phänomen, das aus dieser Vorschrift zwischen $p_c$ und $p_b$ folgt, und diskutieren Sie ausgehend von $x_1$, wie die Klassifizierung der Punkte nach der ersten Iteration aussähe, wenn Sie die Clusterbildung bei $x_1$ beginnen.

### (b)

*5 Punkte*

Welche geometrische Form nehmen die von dichtebasiertem Clustern gefundenen Cluster typischerweise an? Begründen Sie kurz, woran das liegt. Diskutieren Sie ihre Antwort im Kontrast zum zentroidenbasierten Clustering (Centroid-based Clusterin).

Was legt der Wert des Nachbarschaftsradius $\varepsilon$ fest? Beschreiben Sie, wie sich ein zu klein und ein zu groß gewählter Wert auf das entstehende Clustering auswirkt.

Wie bestimmen Sie aus den Daten selbst einen geeigneten Wert für $\varepsilon$? Benennen Sie die Methode und beschreiben Sie das Vorgehen anhand einer Skizze.

**(c)** *4 Punkte*

Der Datensatz wird nun um viele zusätzliche Merkmale erweitert und auf einen Einheitswürfel übertragen, sodass die Punkte in einem Raum der Dimension $D = 1000$ liegen; ihre Anzahl $N$ bleibt dabei unverändert. Benennen Sie das Phänomen, das mithilfe der nebenstehenden Beziehungen sichtbar wird, berechnen und erläutern Sie, was dies bei $D = 1000$ für die Wahl des Nachbarschaftsradius $\varepsilon$ bedeutet.

Benennen Sie eine Methode mit der gegen diese Problematik angegangen werden kann.

Hinweis. Für $N$ gleichverteilte Punkte im Einheitswürfel $[0, 1]^D$ skaliert der mittlere Abstand zum nächsten Nachbarn mit

$$d_{\mathrm{NN}} \propto N^{-1/D},$$

## Aufgabe 2

10 Punkte

Ein variationaler Autoencoder (variational autoencoder, VAE) trainiert den Encoder neben dem Rekonstruktionsverlust (reconstruction loss) mit einem KL-Term, der die kodierte Verteilung $q(z \mid x) = \mathcal{N}(\mu, \sigma^2)$ jedes Eingangs an den Prior $p(z) = \mathcal{N}(0, \mathbf{I})$ heranzieht.

### (a)

*4 Punkte*

Welche Struktur prägt das Heranziehen an den Prior $p(z) = \mathcal{N}(0, \mathbf{I})$ dem latenten Merkmalsraum auf? Arbeiten Sie qualitativ heraus, worin sich dieser Raum von einer Einbettung (embedding) unterscheidet, die allein durch den Rekonstruktionsverlust getrieben wird. Welches Distanzmaß ist für einen solchen Merkmalsraum prädestiniert? Begründen Sie kurz.

Warum ist der Rekonstruktionsloss dennoch relevant und welche Aufgabe übernimmt er?

Welche Grösse der Hauptkomponentenanalyse (principal component analysis, PCA) ist das Pendant zur Dimension des latenten Merkmalsraums? Begründen Sie kurz die Entsprechung.

### (b)

*3 Punkte*

Sie verwenden das obige VAE nun für few-shot learning. Wie passen Sie die obige Modellarchitektur dafür an?

Sie erhalten, nun die zu $x_1$ und $x_2$ gehörenden Label $\tilde{y}_1 = c_1$, $\tilde{y}_2 = c_2$ beschreiben Sie wie Sie nun ohne weiteres Training einen Klassifikator für die beiden Klassen $c$ umsetzen.

### (c)

*3 Punkte*

Ihr Unternehmen fertigt ein sicherheitsrelevantes Bauteil. Fehlteile (defects) treten selten auf, sodass Sie keine Fehler-Labels erzeugen können; aufgenommene Gutteile liegen dagegen reichlich vor. Sie trainieren dafür, wie oben, aber allein auf Gutteilen ein VAE.

Schlagen Sie einen Anomalie-Score $a(x)$ vor und geben Sie dafür eine Formel an.

Wie legen Sie die Schwelle $\tau$ für $a(x)$ datengetrieben fest? Geben Sie eine Regel an, die $\tau$ aus Mittelwert und Varianz (variance) der Gutteil-Scores $a(x)$ bildet, und geben Sie die Gleichung an.

Durch welche minimalistische Erweiterung können Siedie Abwägung zwischen Fehlalarmen und übersehenen Fehlteilen einstellen?