Klausur zur Lehrveranstaltung „Maschinelles Lernen 1 – Grundverfahren“

(60 Minuten)

Nachname:	Vorname:
Matrikelnummer:	Studiengang & Semester:
Bekanntgabe-Code:	

Anmerkungen

- Legen Sie Ihren Studierendenausweis und ein gültiges Ausweisdokument gut sichtbar bereit.
- Tragen Sie Nachname, Vorname, Matrikelnummer, Studiengang & Semester und Bekanntgabecode deutlich lesbar ein und unterschreiben Sie das Klausurexemplar unten.
- Die folgenden 6 Aufgaben sind vollständig zu bearbeiten.
- Als Hilfsmittel ist nur ein nicht programmierbarer Taschenrechner zugelassen.
- Täuschungsversuche führen zum Ausschluss von der Klausur.
- Unleserliche oder mit Bleistift geschriebene Lösungen können von der Korrektur bzw. der Wertung ausgeschlossen werden.
- Beim Ausfüllen von Lücken gibt die Größe der Kästen keinen Aufschluss über die Länge des einzufügenden Inhaltes.
- Die Bearbeitungszeit beträgt 60 Minuten.

Ich bestätige, dass ich die Anmerkungen gelesen und mich von der Vollständigkeit dieses Klausurexemplars (Seite 1 - 13) überzeugt habe.

_____

Unterschrift

Nur für den Prüfer

Aufgabe	1	2	3	4	5	6	Gesamt	Note
Punkte	7	11	11	13	11	7	60	
Erreicht								

Aufgabe 1 (7 Punkte)

a) Was besagt das Rasiermesser-Prinzip (Occam's Razor)? (___/1P)

b) Warum erfüllt die strukturelle Risikominimierung (Structural Risk Minimization) das Rasiermesser-Prinzip (Occam's Razor)? (___/1P)

c) Beim überwachten Lernen teilt man große Mengen von Lernbeispielen häufig in drei disjunkte Teilmengen auf: Trainingsdaten, Validierungsdaten und Testdaten. Welche Funktionen haben die Teilmengen Validierungsdaten und Testdaten jeweils? (___/2P)

d) Betrachten wir ein binäres Klassifikationsproblem mit möglichen Ausgabewerten $y \in \{0,1\}$ und Eingabewerten $x \in \mathbb{R}^2$. Der Hypothesenraum sei gegeben durch die Menge $H = \{h_r | r \in \mathbb{R}^+\}$ mit $h_r(x) := \begin{cases} 1 & falls \ ||x||_2 \leq r \\ 0 & sonst \end{cases}$.

1. Wie ist die VC-Dimension bei einer Klassifikation allgemein definiert? (___/1P)

2. Was ist die VC-Dimension $VC(H)$ von H? Begründen Sie Ihre Antwort. (___/2P)

Aufgabe 2 (11 Punkte)

a) Gegeben zwei Hypothesen h_i und h_j. Unter welcher Annahme lässt sich die Maximum a-Posteriori Hypothese zur Maximum Likelihood Hypothese vereinfachen? (___/1P)

b) Unter welcher Annahme lässt sich der optimale Bayes-Klassifikator zum Naiven Bayes-Klassifikator vereinfachen? (___/1P)

c) Der folgende Datensatz beschreibt Beobachtungen des Computerkaufs in einem Geschäft gegeben der ebenfalls beobachteten Attribute:

$$\{Alter, Student, Einkommen\}$$

Zur Vereinfachung ist das Alter in drei Klassen diskretisiert:

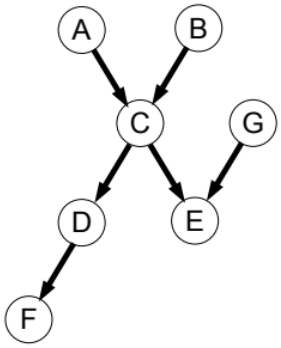
$$Alter(A) = \{A \leq 30, \quad A > 30 \vee A \leq 40, \quad A > 40\}$$

#	Alter(A)	Student(S)	Einkommen(E)	kauft Computer(C)
1	≤ 30	Ja	Mittel	Ja
2	$A > 30 \vee A \leq 40$	Nein	Niedrig	Nein
3	> 40	Nein	Hoch	Ja
4	≤ 30	Nein	Mittel	Nein
5	$A > 30 \vee A \leq 40$	Ja	Niedrig	Ja
6	≤ 30	Ja	Mittel	Ja
7	≤ 30	Nein	Hoch	Ja
8	> 40	Ja	Niedrig	Nein

1. Berechnen Sie die folgenden a-priori und bedingten Wahrscheinlichkeiten:
   $P(C = Ja)$, $P(C = Nein)$, $P(A \leq 30 \mid C = Ja)$ (___/1,5P)

2. Gegeben sei eine 24-jährige Person (Student) mit mittlerem Einkommen. Ein naiver Bayes-Klassifikator soll dazu dienen die Wahrscheinlichkeit zu bestimmen, dass diese Person sich einen Computer kauft. Begründen Sie Ihre Entscheidung mit Hilfe einer geeigneten Formel. (___/1,5P)

d) Bayes'sche Netze bieten eine effiziente Möglichkeit, die bedingte Wahrscheinlichkeit zwischen Variablen in einem DAG (gerichteter azyklischer Graph) zu kodieren.



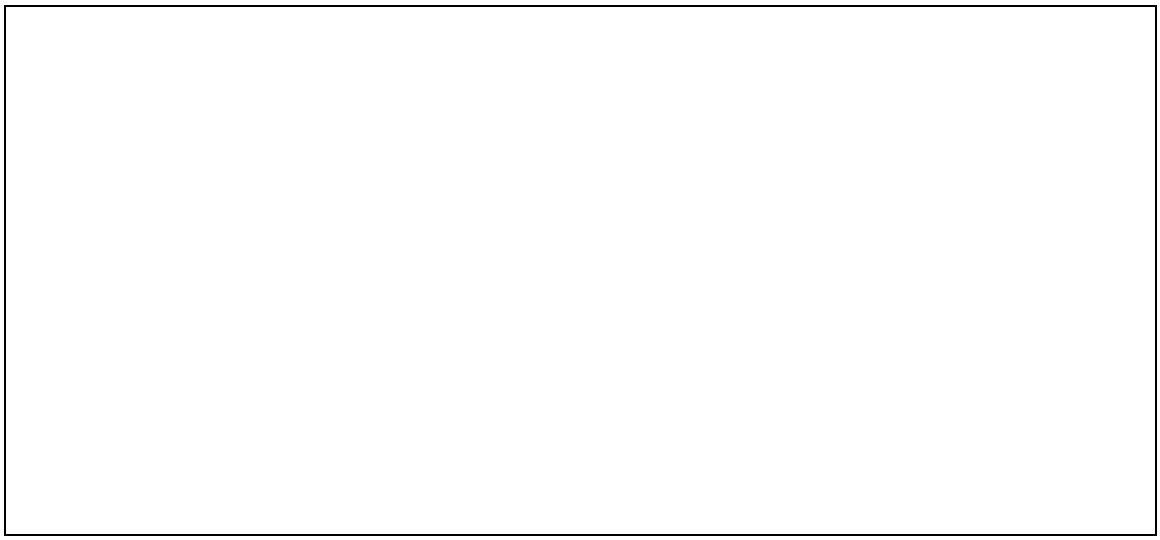
1. Definieren Sie die Gesamtwahrscheinlichkeit der Zufallsvariablen in fakturierter Form. (___/2P)

2. Welche Methode eignet sich zum Lernen Bayes‘scher Netze, wenn die Struktur eines Bayes’schen Netzes bekannt ist aber nur einige Zufallsvariablen beobachtbar sind. (___/1P)

e) Ein HMM (Hidden Markov Modell) ist definiert mit $\lambda = \{S - Zustände, V - Ausgabezeichen, A - Übergangswahrscheinlichkeiten, B - Emmisionswahrscheinlichkeiten, \Pi -$ Verteilung der Anfgangswahrscheinlichkeiten$\}$

1. Gegeben ist die Trainingssequenz O. Welche Methode eignet sich, um das Modell des Systems λ zu bestimmen? (___/1P)

2. Mit Hilfe des Vorwärts- und Rückwärts-Algorithmus kann $P(O|\lambda)$ berechnet werden. Im Vorwärtsalgorithmus wird die Wahrscheinlichkeit $\alpha_t\ (i)$ berechnet und im Rückwärtsalgorithmus die Wahrscheinlichkeit $\beta_t\ (i)$. Definieren Sie die beiden Wahrscheinlichkeiten. (___/2P)



Aufgabe 3 (11 Punkte)

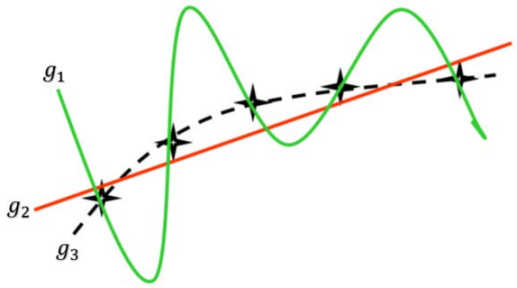
a) Was muss bei der Initialisierung der Gewichte von Neuronen eines neuronalen Netzwerks beachtet werden? Was wird dadurch vermieden? (___/1P)

b) Geben Sie die quadratische Fehlerfunktion E des Gradientenabstiegs an sowie die Formel der iterativen Gewichtsoptimierung $\Delta\vec{w}$ in Abhängigkeit von E. Benennen Sie die verwendeten Variablen. (___/2P)

c) Nennen Sie zwei Probleme, die bezüglich der Ausartung der Fehlerflächen beim Gradientenabstieg auftreten können. Geben Sie zwei Methoden an, mit denen diese Probleme jeweils vermieden werden können. (___/2P)

d) Wie unterscheidet sich Stochastic Gradient Descent (bzw. Pattern Learning) vom „echten“ Gradientenabstieg? Was sind die jeweiligen Vorteile der beiden Verfahren? (___/2P)

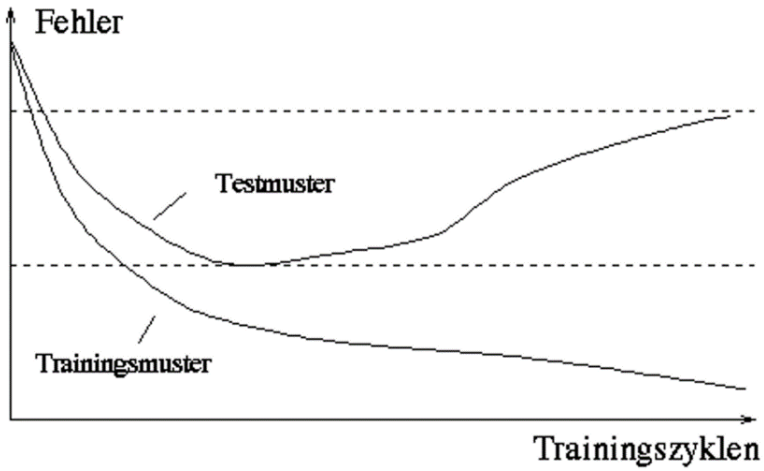
e) Die folgende Abbildung zeigt die Kurven g_1 und g_2, die von zwei verschiedenen neuronalen Netzen an die gleichen Trainingsdaten (Sterne) angepasst wurden. Die Kurve g_3 entspricht der zu erlernenden Funktion (Grundwahrheit).



1. Wie nennt man das Phänomen, das bei Kurve g_1 auftritt? Nennen Sie zwei Ansätze, die dieses Phänomen bei neuronalen Netzen verhindern können. (___/2P)

2. Welche Vermutung kann über die VC-Dimension des Netzes, das die Kurve g_2 als Approximation für die Trainingsdaten liefert, getroffen werden? (___/1P)

3. Betrachten Sie die folgende qualitative Skizze, die die Entwicklung der Fehlerfunktion für Trainings- und Testdaten im Trainingsverlauf eines Netzes zeigt. Zu welchem der zwei Netze passt diese Skizze am besten? Begründen Sie ihre Entscheidung. (___/1P)



Aufgabe 4 (13 Punkte)

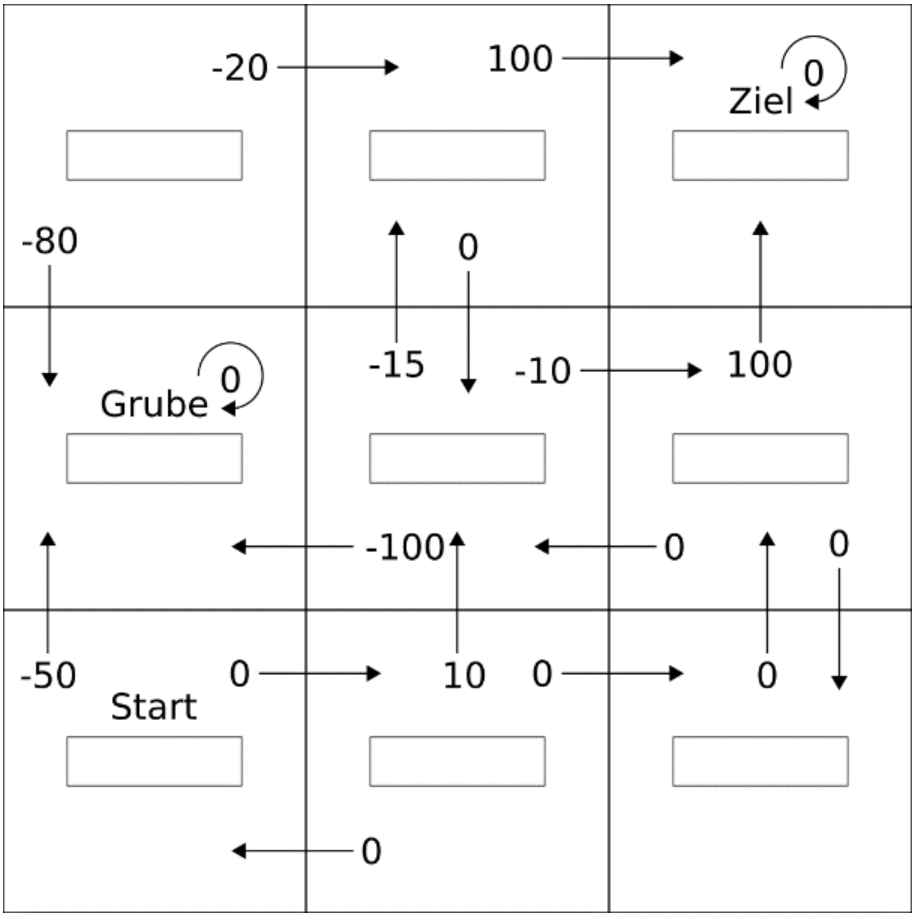
a) Durch welches Modell lässt sich die Problemstellung beim Reinforcement Learning formal darstellen? Welche vier Bestandteile werden für die Modellierung benötigt? (___/3P)

b) Was besagt die Markov-Bedingung? (___/1P)

c) Gegeben $V^{\pi}(s)$: Wie lässt sich die optimale Strategie formal bestimmen? Definieren Sie zusätzlich die rekursive Form der Bellmann Optimalitätsgleichung in Abhängigkeit von V. (___/2P)

d) Wie sollte man die Suchstrategie im Laufe des Lernprozesses anpassen und warum? Verwenden Sie die Begriffe *Exploitation* und *Exploration*. (___/2P)

e) Betrachten Sie die untenstehende Welt. Ein Agent kann sich mit den angezeigten Zustandsübergängen von Zelle zu Zelle bewegen. Die Belohnung für einen Übergang entspricht der Zahl an den Pfeilen. Nehmen Sie an, dass die optimale Strategie gelernt wurde. Tragen Sie die Zustandswerte dieser Strategie in die entsprechenden Kästchen ein (Diskontierungsfaktor = 0,9) und zeichnen Sie den Pfad der optimalen Strategie vom Start zum Ziel ein. Runden Sie ihre Ergebnisse auf ganze Zahlen. (___/5P)

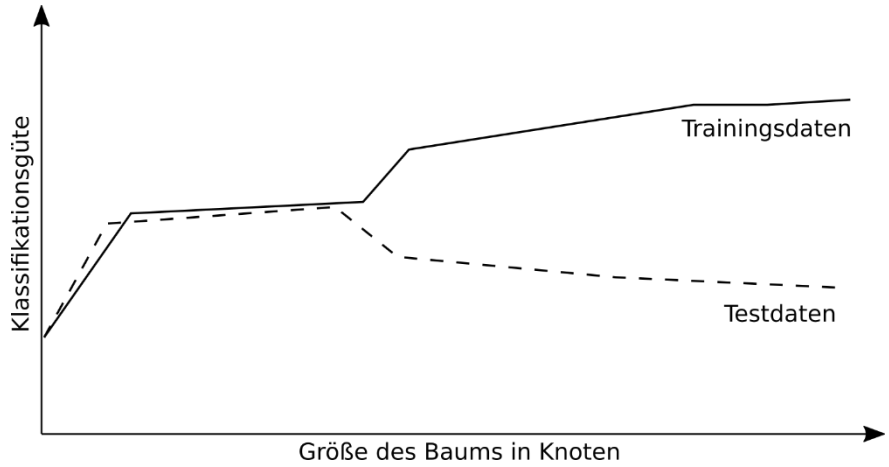


Aufgabe 5 (11 Punkte)

a) Ein Entscheidungsbaum (z.B. ID3) wird durch die Auswahl des jeweils besten Attributes konstruiert. Nennen Sie ein Maß für den Informationsgewinn durch Attribut A. Definieren Sie dieses Maß. (___/2P)

b) p ist der Anteil der positiven Beispiele in den Trainingsdaten S eines binären Klassifikationsproblems. Geben Sie die Formel der $Entropie(S)$ an. Skizzieren Sie den Verlauf der Entropie in Abhängigkeit von p. Markieren Sie die Punkte mit maximaler Trennungsschärfe der Klassen. (___/3P)

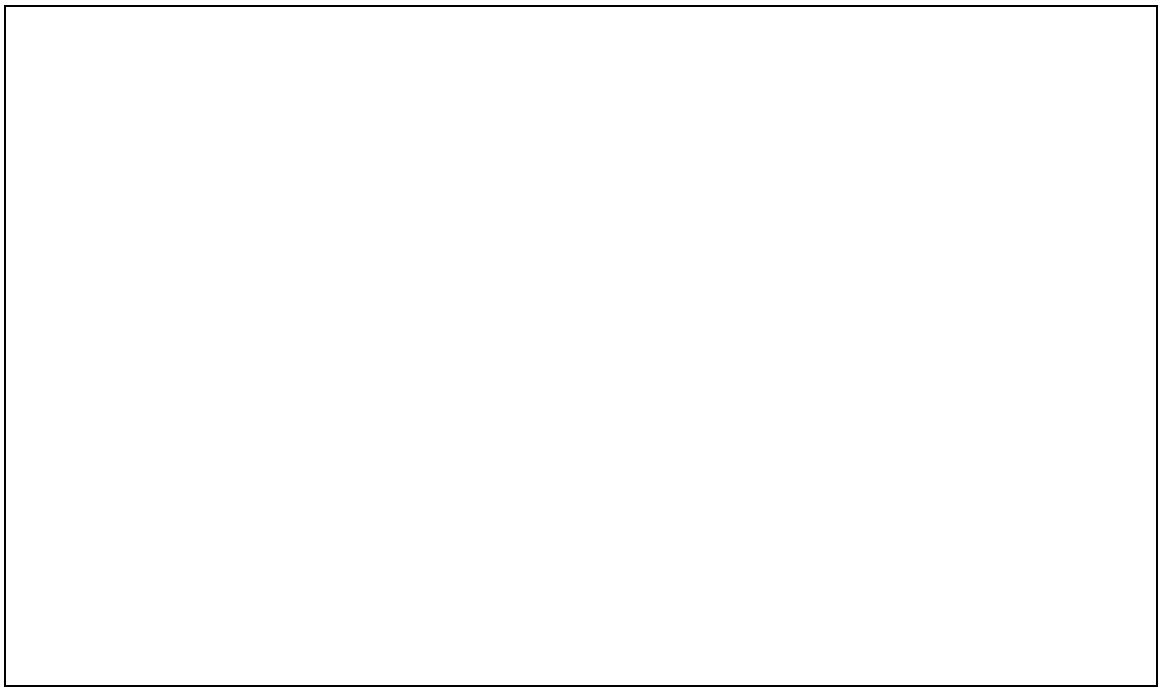
c) Markieren Sie ab welcher Knotenanzahl ein Overfitting beim Training eines Entscheidungsbaumes stattfindet. Begründen Sie warum. (___/1P)



d) Nennen Sie zwei Methoden um Overfitting bei Entscheidungsbäumen zu vermeiden. (___/1P)

e) Betrachten Sie die nachfolgende Tabelle über ausgetragene bzw. nicht ausgetragene Tennisspiele. Welches der Attribute eignet sich am besten als Entscheidungskriterium dafür, dass ein Tennisspiel stattfindet? Begründen Sie ihre Entscheidung. Skizzieren Sie basierend auf diesem Ergebnis den Entscheidungsbaum. (___/4P)

Nr.	Luftfeuchtigkeit	Wind	Tennis?
1	normal	schwach	nein
2	hoch	stark	nein
3	hoch	schwach	ja
4	normal	schwach	ja
5	normal	stark	nein
6	hoch	schwach	ja
7	hoch	stark	nein
8	normal	schwach	ja
9	hoch	stark	nein
10	normal	stark	ja
11	normal	schwach	nein
12	normal	stark	nein



Aufgabe 6 (7 Punkte)

a) Beschreiben Sie kurz die Grundidee, die der Methode der Support Vektor Klassifikation zugrunde liegt. Wie ist das Lernverfahren einzuordnen? (___/2P)

b) Geben Sie die Formeln für das Optimierungskriterium der optimalen Hyperebene und für die Randbedingung einer korrekten Klassifikation an (gegeben Trainingsbeispiele der Form $(\vec{x}, y)$). (___/2P)

c) Erklären Sie die Dualität zwischen Hypothesenraum und Merkmalsraum im Kontext des SVM Verfahrens (Version Space Duality). Wie ist die optimale Lösung im Hypothesenraum repräsentiert? (___/2P)

d) Welche Beobachtung erlaubt die Anwendung des „Kerneltricks“ zur Klassifikation von Beispielen in höherdimensionalen Räumen? (___/1P)