Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB)
Angewandte Technisch-Kognitive Systeme (ATKS)
Prof. Dr. J. Marius Zöllner

Karlsruher Institut für Technologie

Klausur über den Stoff der Lehrveranstaltung: „Maschinelles Lernen 1 - Grundverfahren“

(60 Minuten)

Name:	Vorname:
Matrikelnr.:	Studiengang:

Anmerkungen:

- Legen Sie Ihren Studierendenausweis gut sichtbar bereit.
- Tragen Sie Nachname, Vorname, Matrikelnummer und Studiengang deutlich lesbar ein und unterschreiben Sie das Klausurexemplar unten.
- Die folgenden 6 Aufgaben sind vollständig zu bearbeiten. Jede Antwort muss entweder in deutscher oder englischer Sprache formuliert sein.
- Als Hilfsmittel sind ausschließlich folgende zugelassen:
  - ein nicht programmierbarer Taschenrechner
  - ein nicht beschriftetes Wörterbuch
- Täuschungsversuche führen zum Ausschluss von der Klausur.
- Unleserliche oder mit Bleistift geschriebene Lösungen können von der Korrektur bzw. der Wertung ausgeschlossen werden.
- Die Bearbeitungszeit beträgt 60 Minuten.

Ich bestätige, dass ich die Anmerkungen gelesen und mich von der Vollständigkeit dieses Klausurexemplars (Seite 1 - 19) überzeugt habe.

_____

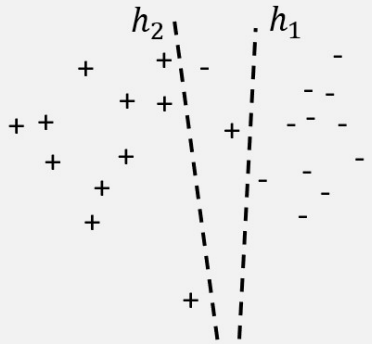
Unterschrift

Nur für den Prüfenden:

Aufgabe	1	2	3	4	5	6	Gesamt
Punkte	9	6	14	8	11	12	60
Erreicht							

Aufgabe 1 Lerntheorie & Unüberwachtes Lernen ____/9 Punkte

a) **(___/4P)**



Die Hypothesen h_1 und h_2 aus der Gleichung sind nicht Teil eines Versionsraums, da sie nicht konsistent und vollständig sind.

Welche von ihnen ist **konsistent** aber nicht **vollständig**?
Welche ist **vollständig** aber nicht **konsistent**?

Zeichnen sie eine vollständige und konsistente Hypothese in das Schaubild ein.

Welche Eigenschaft ergibt sich für den Versionsraum eines linearen Modells mit linearem Hypthesenraum H_{lin} für dieses Schaubild?

Hypothese ist **konsistent** und **nicht vollständig**:

Hypothese ist **nicht konsistent** und **vollständig**:

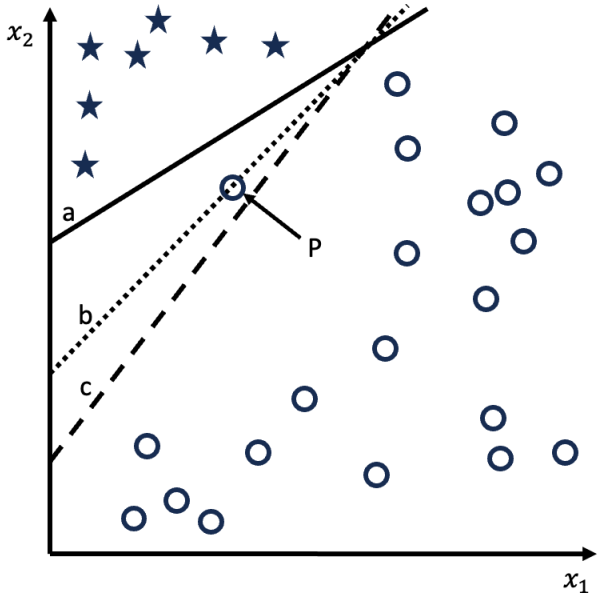
Eigenschaft H_{lin}:

b) Sie haben einen gelabelten Datensatz. Sie teilen ihn auf in einen Test- und Trainingsdatensatz. **(___/2P)**
Sie trainieren sehr viele unterschiedliche Modelle auf dem Trainingdatensatz und evaluieren auf dem Testdatensatz. Sie verwenden immer die besten Modelle des Testdatensatzes um sie weiter zu verbessern. Außerdem verwenden sie den Testdatensatz um Early-Stopping zu implementieren.
Wenn sie die Modelle in der echten Welt einsetzten, fällt ihnen auf, dass das beste Modell des Testdatensatzes nicht die beste Genauigkeit liefert sondern ein Modell, das im Testdatensatz schlechter abgeschnitten hatte. Was könnten die Gründe dafür sein?

c) Sie verwenden K-means um ihre Daten zu clustern, aber Ihnen fällt auf, dass die Cluster nur sehr schlecht die zugrundeliegenden Klassen abbilden. Nennen Sie drei Fehler- bzw. Problemquellen die dazu führen könnten **(___/3P)**

Aufgabe 2 Support Vector Machine

____/6 Punkte



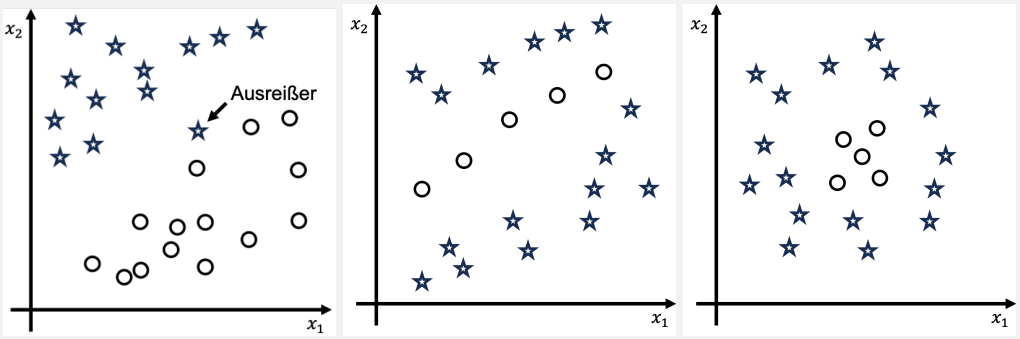
Gegeben ist ein Datensatz mit den zwei Klassen Stern und Kreis. Dieser soll mit einer linearen SVM korrekt klassifiziert werden. In der oberen Abbildung finden Sie eine grafische Darstellung der Datenpunkte und verschiedenen Hypothesen im Musterraum/Merkmalsraum.

a) Geben Sie an welche der Hypothesen (`a, b, c`) das optimale Ergebnis des SVM Algorithmus auf diesem Datensatz ist. **(___/1P)**

b) Markieren Sie die 4 Stützvektoren (Support Vectors) in der oberen Abbildung durch Einkreisen. (Info: Falsch markierte Stützvektoren führen zu Punkteverlust. Streichen Sie falsche Ergebnisse durch.) **(___/2P)**

c) Wie verändert sich das Ergebnis aus a), wenn der Datenpunkt `P` aus dem Datensatz entfernt wird? **(___/0,5P)**

d) **(___/1.5P)**



Datensatz a) Datensatz b) Datensatz c)

Zu sehen sind drei verschiedene Datensätze (a, b, c). Geben Sie für jeden Datensatz an, welche SVM Methode (lineare SVM, soft margin SVM, nicht-lineare SVM) sich für diesen am besten eignet.

e) Erläutern Sie kurz den Kernel-Trick und welcher Vorteil sich daraus ergibt. **(___/1P)**

Aufgabe 3 Neuronale Netze

____/14 Punkte

a) Optimierungsmethoden zweiter Ordnung sollten den Loss besser minimieren. Dennoch werden sie bei neuronalen Netzen nur sehr selten verwendet. Wieso? **(___/2P)**

Nennen Sie zusätzlich eine Optimierungsmethode erster Ordnung, die eine Optimierungsmethode zweiter Ordnung approximiert.

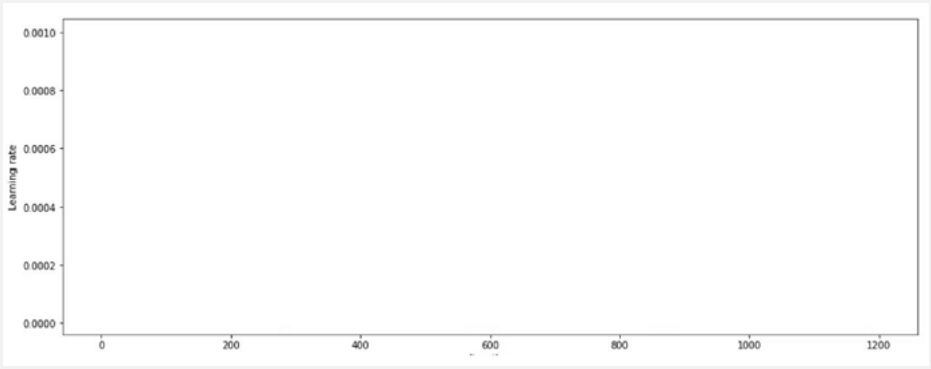
b) Geben Sie die in der Vorlesung vorgestellte ReLU Funktion und deren Ableitung an. **(___/1P)**

c) Was ist der Vorteil von ReLU gegenüber einer Sigmoid- oder Tanh-Aktivierungsfunktion? **(___/1.5P)**
Die genannten Aktivierungsfunktionen sind nicht-linear, was würde passieren wenn Sie nur lineare Aktivierungsfunktionen für ein neuronales Netz verwenden würden?

d) Wozu wird Datenaugmentierung durchgeführt und was ist dabei zu beachten? Nennen Sie drei Beispiele wie man Bilddateien augmentieren kann. **(___/2P)**

e) **(___/2P)**

1. Warum werden meist dynamische Lernraten statt statischen Lernraten verwendet?
2. Welchen Vorteil können zyklische Lernraten gegenüber monoton absinkenden Reduktionsverfahren haben?
3. Skizzieren Sie den Verlauf einer zyklischen Lernrate mit Cosinus-Reduktion (cosine-annealing)



f) Gegeben ist ein Neuron mit Inputvektor $\vec{x}$, Gewichten $\vec{w}$ und dem Bias b. Das Neuron verwendet eine Sigmoid Aktivierungsfunktion $\sigma()$ und gibt für die Eingabe $\vec{x}$, die Ausgabe a aus. Führen Sie einen Backpropagation-Schritt mit der Fehlerfunktion L und Label $\hat{y}$ durch. **(___/4P)**

Errechnen Sie die Gradienten der Gewichte und des Bias $\frac{\partial L}{\partial w_0}$, $\frac{\partial L}{\partial w_1}$ und $\frac{\partial L}{\partial b}$.

Geben Sie zusätzlich die Zwischenergebnisse $\frac{\partial L}{\partial a}$, $\frac{\partial a}{\partial z}$, $\frac{\partial z}{\partial w_0}$, $\frac{\partial z}{\partial w_1}$ und $\frac{\partial z}{\partial b}$ an.

Eingabevektor $\vec{x} = [-3, 2]$, Gewicht $\vec{w} = \left[\frac{1}{5}, \frac{4}{5}\right]$ und Bias $b = \frac{2}{5}$ und Label $\hat{y} = 1$.

Als Hilfestellung geben wir Ihnen die Neuronenformel an und die jeweiligen Ausgaben a, und z. Weiter geben wir die Ableitung der Sigmoid Funktion an, sodass keine explizite Sigmoid-Berechnung notwendig ist.

$$z = \sum_k w_k \cdot x_k + b$$
$$a = \sigma(z)$$
$$L = \tfrac{1}{2}(\hat{y} - a)^2$$
$$z = \frac{7}{5}, \quad \sigma(z) = a = 0.8, \qquad \sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$$

$\frac{\partial L}{\partial a} =$

$\frac{\partial a}{\partial z} =$

$\frac{\partial z}{\partial w_0} =$

$\frac{\partial z}{\partial w_1} =$

$\frac{\partial z}{\partial b} =$

$\frac{\partial L}{\partial w_0} =$

$\frac{\partial L}{\partial w_1} =$

$\frac{\partial L}{\partial b} =$

g) Dasselbe Neuron aus der vorherigen Aufgabe (Gewicht $\vec{w} = \left[\frac{1}{5}, \frac{4}{5}\right]$ und Bias $b = \frac{2}{5}$) erhält nun eine andere Eingabe und eine andere Bezeichnung. Dabei errechnete es im Backpropagation-Schritt folgende Gradienten: **(___/1.5P)**

$$\frac{\partial L}{\partial w_0} = -1$$
$$\frac{\partial L}{\partial w_1} = 2$$
$$\frac{\partial L}{\partial b} = 1$$

Führen Sie mit den vorgegebenen Gradienten, der Lernrate $\eta = \frac{1}{5}$ und den Gewichten und Bias aus der vorhergegangenen Aufgabe einen Gewichts-Update-Schritt durch. Geben Sie $\vec{w^{n+1}}$ und b^{n+1} an.

$w_0^{n+1} =$

$w_1^{n+1} =$

$b^{n+1} =$

Aufgabe 4 Convolutional Neural Networks ____/8 Punkte

a) Gegeben ist ein Input-Bild mit den Dimensionen $32 \times 24 \times 3$ $(H \times W \times C)$. **(___/3P)**
Gewünscht ist eine Output-Featuremap mit der Dimension $24 \times 22 \times 4$.
Geben Sie die Dimensionen $H_k \times W_k \times C_{in} \times C_{out}$ eines Convolutional Layers an, das in Kombination mit einem Padding von je 1px auf allen Seiten, und einem Stride von 1px die gewünschten Output-Dimensionen erzeugt.

b) Berechnen Sie die folgende Convolution **mit Stride=2** und ohne Padding. **(___/2P)**

Input (4x4x1)

1	1	0	2
2	0	1	2
-3	1	2	1
1	1	0	0

*

Kernel (2x2x1x1)

1	-1
-1	1

=

c) Bei CNNs ist die Größe der Kernel ein wichtiger Hyperparameter. **(___/3P)**

1. Welche Vorteile haben 3×3 Kernel gegenüber größeren z.B. 7×7 Kernel hinsichtlich des rezeptiven Felds?
2. Wie verhält sich das rezeptive Feld bei CNNs die nur 1×1 Kernels verwenden?
3. Die Ausgabe eines 3×3 Kernels hat eine kleinere Auflösung als seine Eingabe. Erklären Sie eine Art von Padding und eine dazu passende Größe, um dies zu verhindern.

Aufgabe 5 Reinforcement Learning ____/11 Punkte

a) Was ist das primäre Ziel von Reinforcement Learning? Geben Sie eine kurze Beschreibung an. **(___/1P)**

b) Was ist die Hauptbedingung, die ein Prozess erfüllen muss, um als Markov-Entscheidungsprozess bezeichnet zu werden? **(___/1P)**

c) Definieren Sie kurz den Begriff „modellfrei“ im Kontext von Reinforcement Learning. **(___/1P)**

d) Warum werden Funktionsapproximatoren (wie z.B. neuronale Netze) in RL benötigt? Geben Sie dazu auch ein kurzes Beispiel an, in dem ein Fuktionsapproximator benutzt werden muss. **(___/1P)**

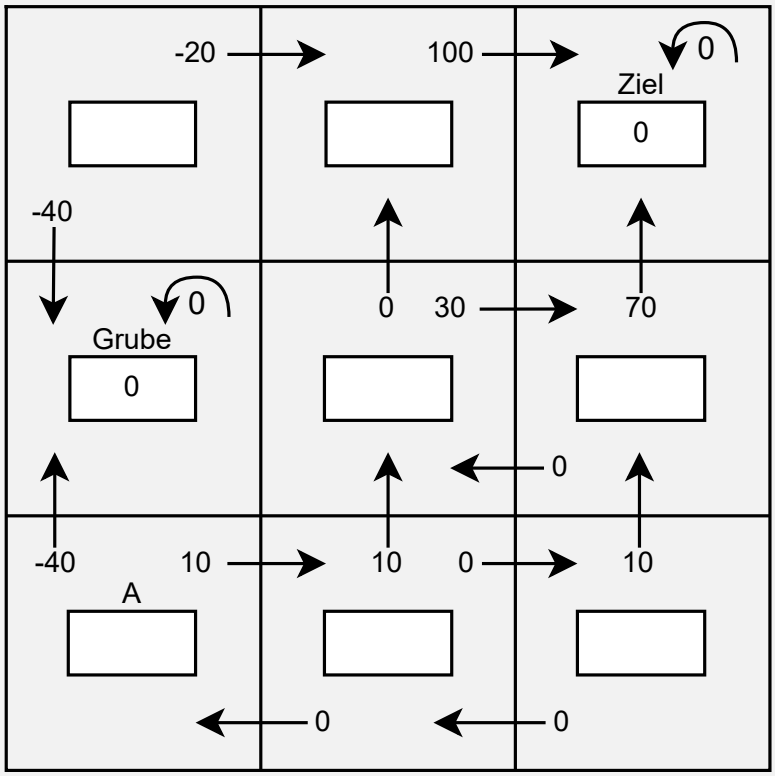
e) Ein RL-Agent hat in einer neuen Umgebung zwei Trajektorien gesammelt, die alle von einem Anfangszustand s_0 ausgehen. Dabei wurde eine Policy π' verwendet. **(___/2.5P)**

- In der **ersten** Trajektorie hat der Agent folgende Belohnungen erhalten: $\{r_0 = 0, r_1 = 3, r_2 = -2, r_3 = 4, r_4 = 1\}$.
- In der **zweiten** Trajektorie hat der Agent folgende Belohnungen erhalten: $\{r_0 = 0, r_1 = 2, r_2 = -2, r_3 = 3, r_4 = 3\}$

Berechnen Sie den diskontierten Gewinn für jede dieser Trajektorien $\{G_1, G_2\}$ unter der Annahme, dass $\gamma = 0,8$ ist. Evaluieren Sie anschließend die Policy π' im Zustand s_0 mithilfe der Zustandswertfunktion $\hat{V}_{\pi'}(s_0)$ anhand der gesammelten Trajektorien.

Hinweis: Richtige Gleichungen reichen für die volle Punktzahl.

f) Betrachten Sie die untenstehende Welt. Ein Agent kann sich mit den angezeigten Aktionen (Pfeilen) von Zelle zu Zelle bewegen. Die Belohnung für eine Aktion entspricht der Zahl an dem entsprechenden Pfeil. Nehmen Sie an, dass die optimale Strategie gelernt wurde. **(___/4,5P)**



Tragen Sie die Zustandswerte $V^*(s)$ dieser Strategie in die entsprechenden Kästen ein (Diskontierungsfaktor $\gamma = \mathbf{0,8}$). Runden Sie Ihre Ergebnisse auf ganze Zahlen. Zeichnen Sie den Pfad der optimalen Strategie von Zelle A zum Ziel ein.

Aufgabe 6 HMM, Bayes, Entscheidungsbäume ____/12 Punkte

a) Nennen Sie zwei Beispielsanwendungen aus der Vorlesung, die mithilfe eines Hidden Markov Models gelöst werden können. **(___/1P)**

b) Welche der folgenden Eigenschaften treffen auf HMMs zu? **(___/2P)**

1. Vollständig beobachtbare Zustände
2. Nicht vollständig beobachtbare Zustände
3. Zeitinvariantes Modell
4. Zeitvariantes Modell
5. Beschränkter zeitlicher Horizont
6. Unendlicher zeitlicher Horizont
7. Doppelt-Stochastischer Prozess

c) Zur Vorhersage von Verspätungen im Fernverkehr soll ein Naive Bayes Klassifikator eingesetzt werden. Die vorauszusagende Verbindung fährt bei gutem Wetter (W = Gut) und am Tag (T = Tag). Welche Klassifikation ist gemäß des Naive Bayes Ansatzes am wahrscheinlichsten? Geben Sie die Formel, den Rechenweg, sowie Ihre Schlussfolgerung an. **(___/3P)**

Folgende Daten sind gegeben. Grau hinterlegte Einträge dienen zur Vereinfachung beim Ablesen und markieren die eingetretenen Verspätungen (V = Ja).

Nr.	Wetter (W)	Tageszeit (T)	Verspätung (V)
1	Schlecht	Nacht	Nein
2	Schlecht	Nacht	Nein
3	Gut	Tag	Nein
4	Schlecht	Tag	Nein
5	Gut	Nacht	Nein
6	Gut	Tag	Ja
7	Schlecht	Tag	Ja
8	Schlecht	Nacht	Ja
9	Gut	Nacht	Ja
10	Schlecht	Tag	Ja

d) **(___/4P)**

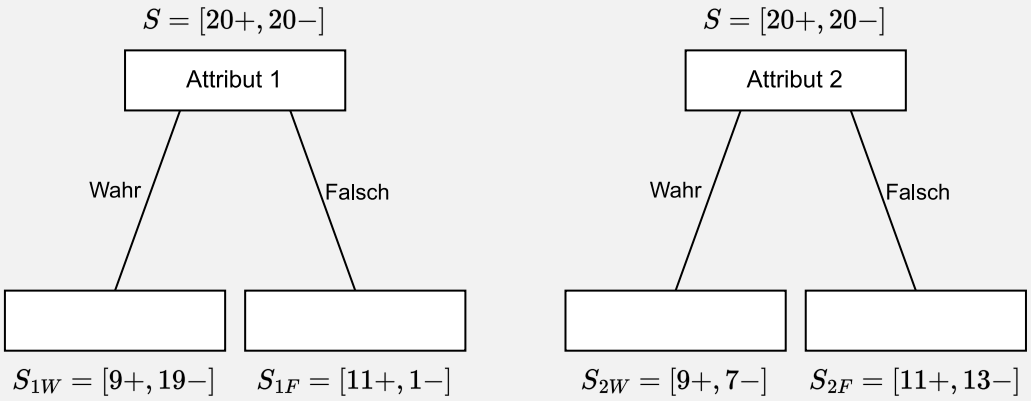


Abbildung 1: **Schreibweise**: [Anzahl positive Bsp. (+), Anzahl negative Bsp. (-)]

In der oben dargestellten Abbildung finden Sie den Vergleich zweier Attribute A_1 (links) und A_2 (rechts), die als Testattribute eines Entschteidungsbaumes ausgewählt werden können. Welches der Attribute eignet sich ausgehend vom Informationsgewinn/Informationgain besser als Entscheidungskriterium? Begründen Sie Ihre Antwort rechnerisch und runden Sie auf zwei Nachkommastellen.
Hilfestellung: Formel für den Informationsgewinn und die Entropie.

$$\mathrm{IG(S, A)} = \mathrm{H}(S) - \Sigma_{v \in V(A)} \frac{|S_v|}{|S|} \mathrm{H}\left(S_v\right)$$

$$H(S) = -\sum_{i=1}^{K} p\left(y_i\right) \log_2 p\left(y_i\right)$$

e) Nennen Sie **zwei** Vorteile von Entscheidungsbäumen gegenüber neuronalen Netzen. **(___/1P)**

f) Gegeben ist Bayes' Theorem. Geben Sie die Namen der Terme $P(A)$ und $P(A|B)$ an. **(___/1P)**

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$