

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**(7 Punkte)**

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

Entities should not be multiplied beyond necessity

- 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)

*Validation data; use to monitor the performance of the model during the training process and to adjust hyperparameters
Test data; use for the final evaluation of the trained model after hyperparameter tuning and optimization
are completed.*

- 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 $h_r(x) := \begin{cases} 1 & \text{falls } \|x\|_2 \leq r \\ 0 & \text{sonst} \end{cases}$.

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

The VC Dimension $VC(H)$ of H is equal to the maximum number of data points which can be arbitrarily separated from H .

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

*3, $\alpha \in \mathbb{R}^2$ Hyperplane in \mathbb{R}^2
 $\Rightarrow 3$ separable values*

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)

$$p(h_i) = p(h_j)$$

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

example: G_i Conditionally independent

- 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:

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

#	Alter(A)	Student(S)	Einkommen(E)	kauf 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:

$$P(C = \text{Ja}), P(C = \text{Nein}), P(A \leq 30 | C = \text{Ja}) \quad (___ / 1,5P)$$

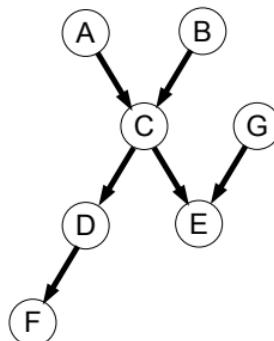
$$p(C = \text{Ja}) = \frac{5}{8} \quad p(C = \text{Nein}) = \frac{3}{8}$$

$$P(A \leq 30 | C = \text{Ja}) = \frac{p(A \leq 30, C = \text{Ja})}{p(C = \text{Ja})} = \frac{3}{8} / \frac{5}{8} = \frac{3}{5}$$

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)

$$\begin{aligned}
 & p(C=1) \cdot p(A \leq 30 | C=1) \cdot p(S=1 | C=1) \cdot p(E=M | C=1) \\
 & = \frac{5}{8} \times \frac{3}{6} \times \frac{3}{5} \times \frac{2}{6} = \frac{9}{120} \\
 & p(C=N) \cdot p(A \leq 30 | C=N) \cdot p(S=1 | C=N) \cdot p(E=M | C=N) \\
 & = \frac{3}{8} \times \frac{1}{3} \times \frac{1}{5} \times \frac{1}{3} = \frac{1}{120} \approx 0.0083 \quad \text{Ja}
 \end{aligned}$$

- 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)

$$\begin{aligned}
 p(A, B, C, D, E, F, G) = & p(A) \cdot p(B) \cdot p(C | A, B) \cdot p(D | C) \\
 & \cdot p(E | C, G) \cdot p(F | D, E) \cdot p(G)
 \end{aligned}$$

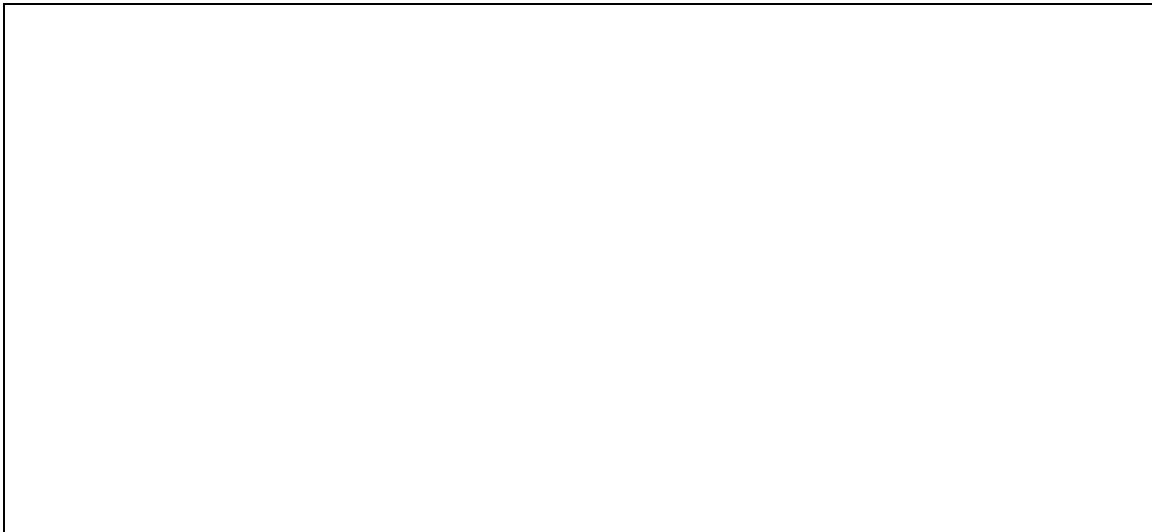
2. Welche Methode eignet sich zum Lernen Bayes'scher Netze, wenn die Struktur eines Bayes'schen Netzes bekannt ist aber nur einige Zufallsvariablen beobachtbar sind. (___ /1P)

Gradient ascent, expectation-maximization Algorithmus

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

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)

The initialisation of weights should be used in training dataset, it can prevent vanishing or exploding gradient.

- 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)

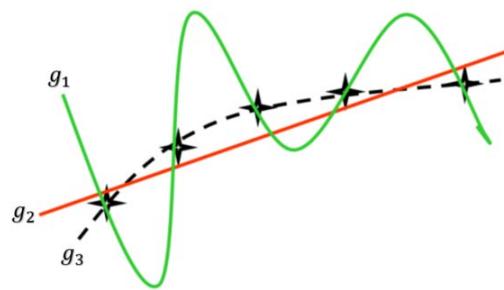
- c) Nennen Sie zwei Probleme, die bezüglich der Ausartung der Fehlerflächen beim Gradientenabstieg auftreten können. Geben Sie zwei Methoden an, mit denen diese Probleme jeweils vermieden werden können. (___/2P)

Exploding gradient: Solution: Gradient Clipping
Vanishing gradient: use activation function without saturation behavior

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

SGD uses a smaller subset B_{train} to approximate the gradient descent.
Gb: more stable, better hardware utilization
SGD: faster update cycles; computationally tractable with limited memory

- 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)

Overfitting

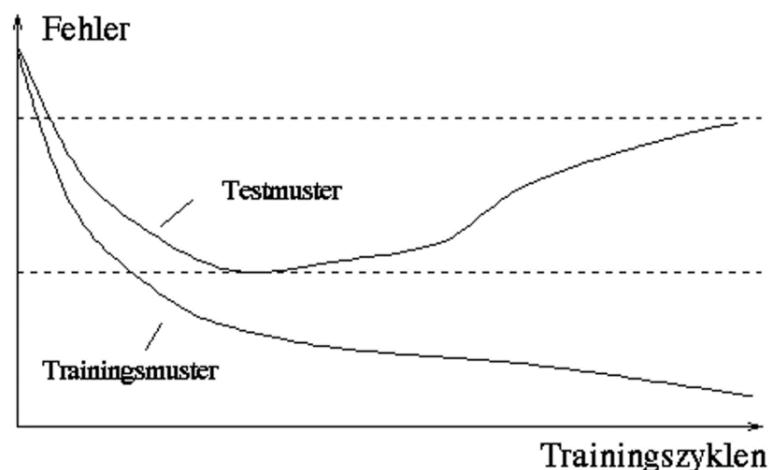
early stopping

decrease model capacity

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

Underfitting

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)



J1. The graph shows Overfitting and g_1 is also overfitting.

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)

Markov decision process
 State space (S) Action space (A)
 Reward function (R) Transition function (P)

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

$$P(S_{t+1} | S_t, a_t) = P(S_{t+1} | S_0, a_0, \dots, S_t, a_t)$$

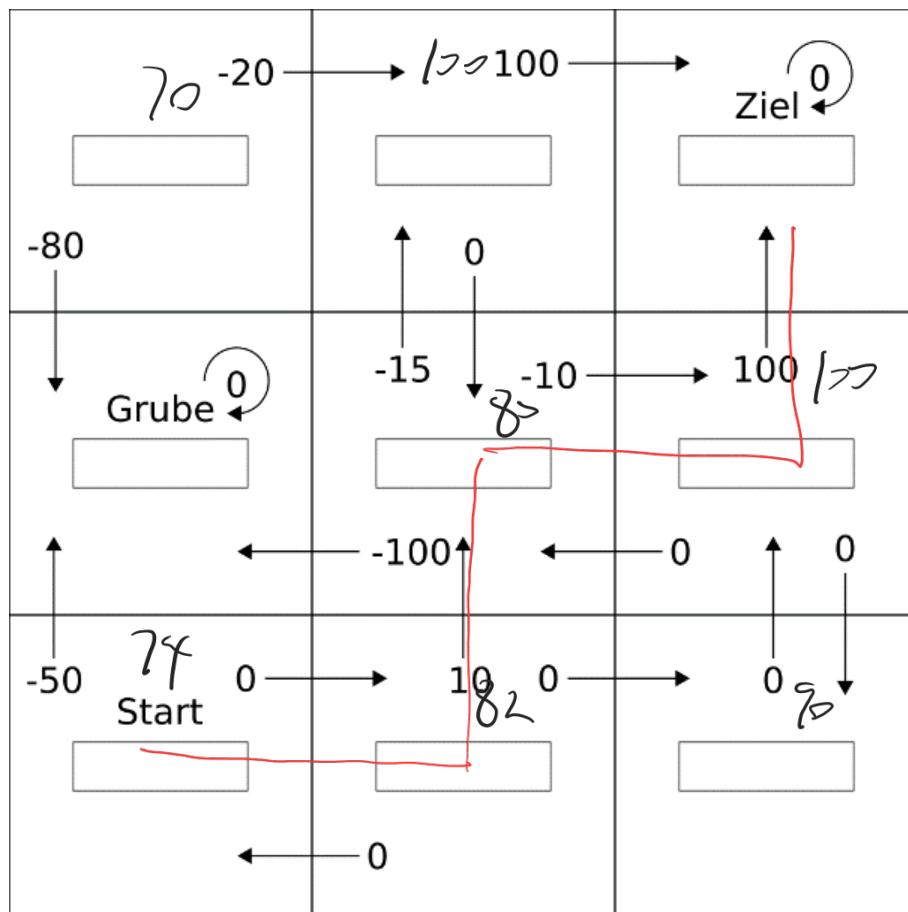
- c) Gegeben $V^\pi(s)$: Wie lässt sich die optimale Strategie formal bestimmen? Definieren Sie zusätzlich die rekursive Form der Bellmann Optimalitätsgleichung in Abhängigkeit von V . (____/2P)

$$\begin{aligned} \pi^*(s) &= \arg \max V^\pi(s) \\ \pi^*(s) &= \max_{a \in A} (r(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) V^\pi(s')) \end{aligned}$$

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

Agent must balance between exploration and exploitation so that all actions are sufficiently explored, but the return is also maximized.

- 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)



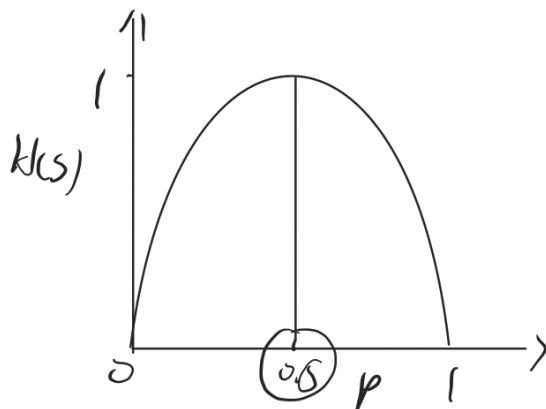
Aufgabe 5**(11 Punkte)**

- 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)

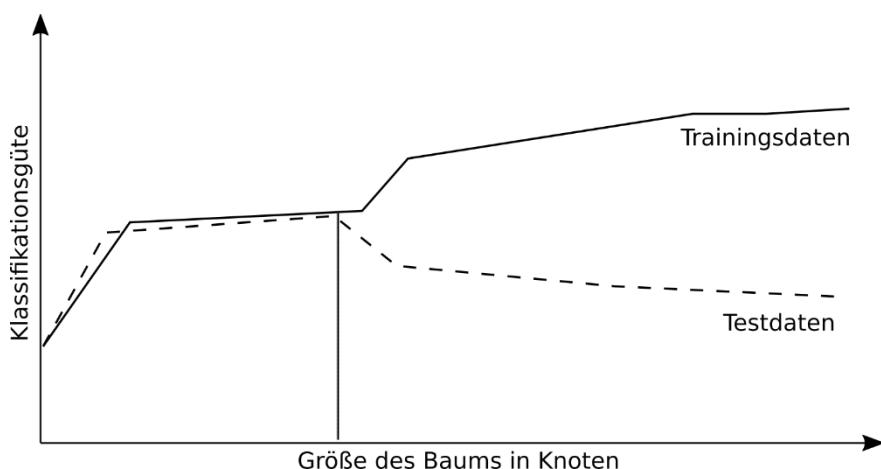
Greedy: Each time an attribute is selected, the attribute that maximizes the information gain is used to split the data.

- 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)

$$H(S) = -p \log_2 p - (1-p) \log_2 (1-p)$$



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



The training error keeps decreasing, while test error increases

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

Early stopping, Pruning

- 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

$$\text{Entscheidungsbaum: } H(S)_2 = -\frac{1}{12} \log_2 \frac{5}{12} - \frac{7}{12} \log_2 \frac{7}{12} = 0.98$$

$$H(S=h)_2 = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.98$$

$$H(S=l)_2 = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$IG(S, L) = 0.98 - \frac{1}{12} \times 0.98 - \frac{5}{12} \times 0.97 = 0.004$$

$$\text{Wind: } H(S=\text{schwach})_2 = -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0.92$$

$$H(S=\text{stark})_2 = -\frac{1}{6} \log_2 \frac{1}{6} - \frac{5}{6} \log_2 \frac{5}{6} = 0.65$$

$$IG(S, W) = 0.92 - \frac{1}{2} \times 0.92 - \frac{1}{2} \times 0.65 = 0.145 \approx 0.14$$

Wind ist besser

$S = [St, L]$

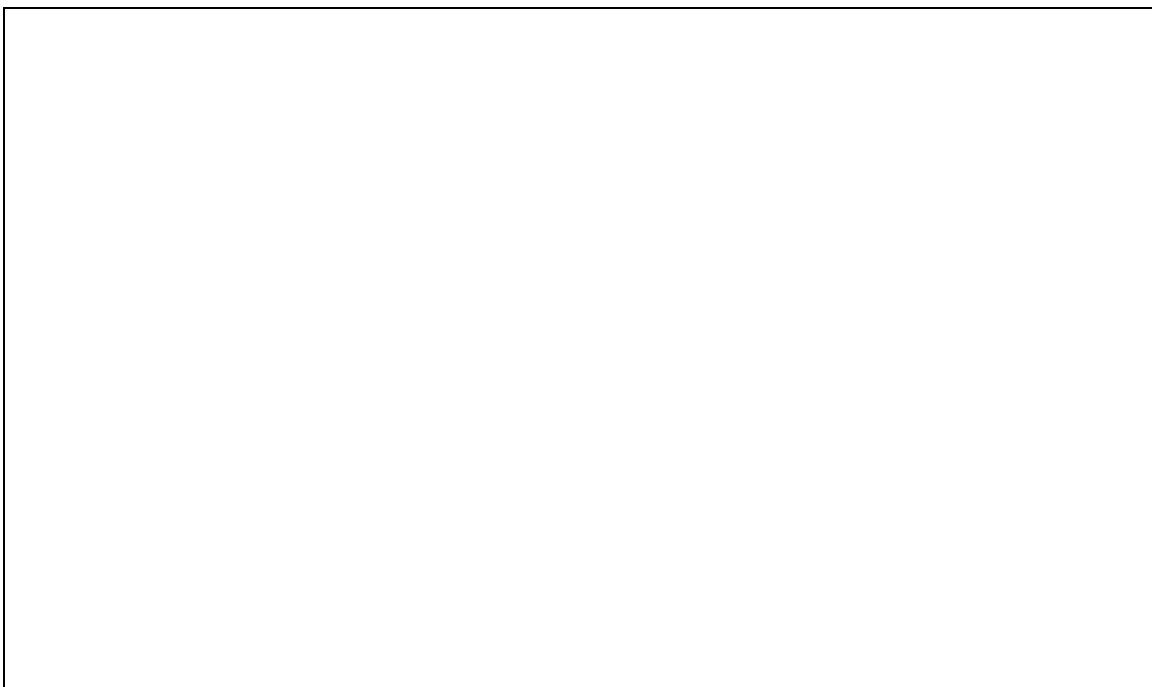


zurück

19t, 2-

zurück

17, 5-



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)

size of the margin determines the generalization capability
 find the best separating line/hyperplane with maximum margin
 for the class.

- 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)

$$\text{minimize } \|\vec{w}\|^2$$

$$y_i(\vec{w}^T \vec{x}_i + b) \geq 1, \quad i=1, \dots, n$$

- 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)

points in feature space correspond to hyperplanes in hypothesis space and vice versa.

The center of the largest hypersphere in hypothesis space

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

Transformation and dot product in high-dimensional spaces is computationally challenging and time-consuming.

