

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) Geben Sie die Formel für den empirischen Fehler einer Lernmaschine an. Es seien dafür die Daten $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$, die Hypothese h_θ , und die Verlustfunktion $L(h_\theta(x_i), y_i)$ gegeben. (____/1P)

OL, p11

$$E_D(h_\theta) = E_{(x,y)}[p[L(h_\theta(x), y)]] = \frac{1}{|D|} \sum_{(x,y) \in D} L(h_\theta(x), y)$$

/ Cost function / Risk: $L(h_\theta) = E_{(x,y)}[p[L(h_\theta(x), y)]] = \int L(h_\theta(x), y) p(x, y) dx dy$

- b) Warum wird in der Praxis der empirische Fehler anstelle des realen Fehlers berechnet? (____/1P)

OL, p10

Because the real risk $L(h_\theta)$ usually can't be calculated!

- c) Nennen Sie jeweils einen Vor- und einen Nachteil von Bagging. (____/1P)

p47

+: Higher quality of the resulting model (improved model accuracy)

-: Complexity and Computational Resources (more)

- d) Wie ist die Vapnik-Chervonenkis (VC) Dimension für lineare Klassifikation definiert? (____/1P)

OL, p59 p6?

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

- e) Im DBSCAN Algorithmus werden Datenpunkte in folgende drei Klassen eingeteilt: (___/3P)
1. Kernpunkte
 2. Erreichbare Punkte
 3. Rauschen

08. 120

Erklären Sie für jede Klasse wann ein Datenpunkt ihr zugeteilt wird.

Hinweis: Sie dürfen die Variablen ϵ und minPoints als gegeben ansehen.

Core Sample: A point is a core sample if at least minPoints points are within distance ϵ

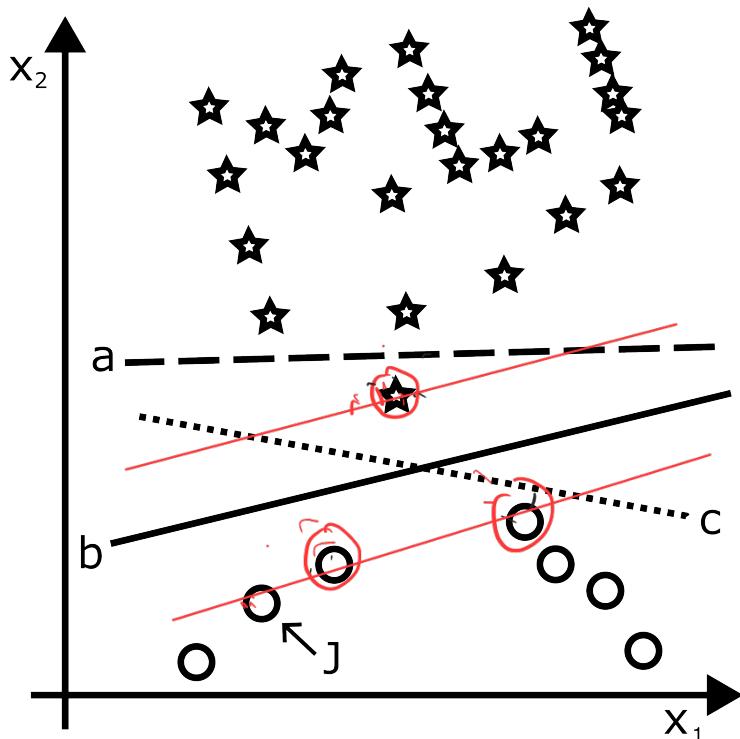
Neighbour: A point is a neighbor if there are not minPoints points within distance ϵ , but at least one point within distance ϵ is a core sample

Noise: There is no core sample within distance ϵ

- f) Ein Autoencoder besitzt einen Eingabetensor, einen Ausgabetensor und einen Tensor für die latente Repräsentation. Vergleichen Sie die Größe/Dimensionalität dieser drei Tensoren untereinander. (___/2P)

08. 134 32

The latent tensor is smaller than the input tensor as it compresses the data representation, while the output tensor has the same size as the input tensor since the goal of the autoencoder is to reconstruct the input data as accurately as possible

Aufgabe 2 Support Vector Machine /6 Punkte

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

07 18

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

b

- b) Markieren Sie die entsprechenden Stützvektoren (Support Vectors) in der oberen Abbildung durch Einkreisen. (Info: Falsch markierte Stützvektoren führen zu Punkteverlust. Streichen Sie falsche Ergebnisse durch.) (/1,5P)

07 19

- c) Wie verändert sich das Ergebnis aus b), wenn der Datenpunkt J aus dem Datensatz (/0,5P) entfernt wird?

no change

- d) Nennen Sie zwei Methoden, mit denen die SVM auch auf nicht linear separierbaren (___/1P)
Daten anwendbar ist.

07. P22

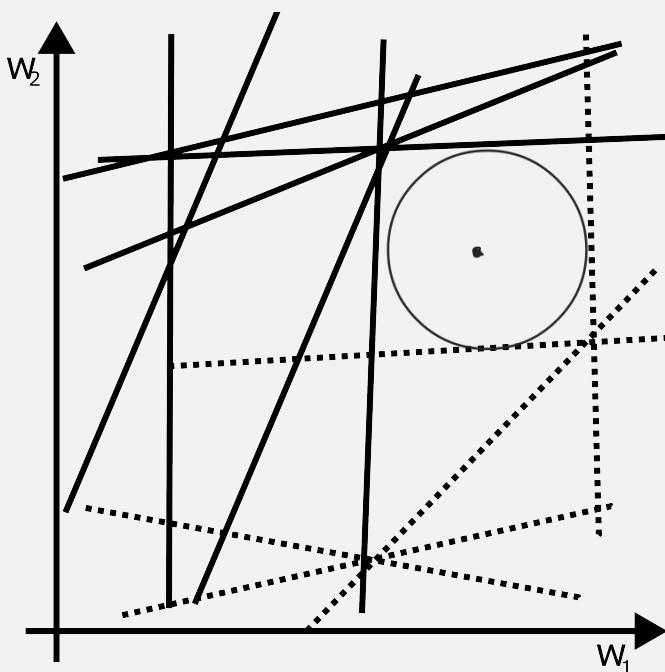
Soft Margin SVM
kernel-Trick

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

07. P28

Defines a dot product / similarity measure in the transformed space without explicitly calculating it.
Use a kernel that does this implicitly, thus saving a lot of computational operations.

- f) Sie haben in der Vorlesung über die Dualität des Merkmals- und Hypothesenraumes gelernt. Zeichnen Sie in den unten angegebenen Hypothesenraum die optimale Hypothese ein. Info: Durchgängige Linien gehören zu Klasse 1 und gepunktete Linien gehören zu Klasse 2. (___/1P)



Aufgabe 3 Neuronale Netze /14 Punkte

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

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

05 P13

Because second-order optimization methods need to calculate the Hessian matrix, which is quadratic and demands a substantial amount of memory.

Adam

- b) Welches Phänomen wird für Hypothese h in folgender Formel beschrieben, wobei D_L (___/2P) die Lerndaten, D_V die Validierungsdaten sind und $h, h' \in H$?

$$\hat{J}_{D_L}(h) < \hat{J}_{D_L}(h') \wedge \hat{J}_{D_V}(h) > \hat{J}_{D_V}(h')$$

02, P19

Nennen Sie außerdem zwei Verfahren um das Phänomen abzuschwächen.

P22

Overfitting

Decrease model capacity

Correct choice and search for optimal hypothesis h_0

- c) Wie können sich zu groß oder zu klein gewählte Lernraten auf den Trainingsverlauf des Modells auswirken? (1P) 05. P21 (___/2P)

Welchen Vorteil können dynamischen Lernraten, wie die häufig verwendete Cosinus-Reduktion (Cosine Annealing), bieten? (1P) Cpt

Too high learning rates can even diverge training.
Too small learning rates will get stuck in every local minimum and can't escape to find global minimum.

Better Convergence; Overcoming local minima;
Robustness to hyperparameter choices; improved generalization
adaptation to model complexity.

- (d) Ihr neuronales Netz verarbeitet Bilder, deren Pixel sie auf den Wertebereich [0,1] skalieren. Geben Sie eine Parameterinitialisierung der Netzparameter an, welche auf diesen Bildern zum Dying ReLU Problem führt. (1P) (___/1.5P)

Um dieses und ähnliche Probleme zu umgehen, wollen Sie eine andere, besser geeignete Initialisierungsmethode verwenden. Welche Komponente Ihres neuronalen Netzes sollte dabei Ihre Entscheidung am meisten beeinflussen? (0.5P) 05. P58

$$\text{Xg vgl. } w_{ij}^{(l)} \sim N(\mu=0, \sigma^2 = \sqrt{\frac{2}{n_{in} + n_{out}}}) \quad \text{l. Xavier}$$

ks imiz / He

$$w_{ij}^{(l)} \sim N(\mu=0, \sigma^2 = \sqrt{\frac{2}{n_{in}}})^2 \quad \text{1. Activation function}$$

dunction

$$\text{LeakyReLU}(x) = \begin{cases} x, & x > 0 \\ \alpha x, & x \leq 0 \end{cases}$$

- e) Gegeben ist ein Neuron mit Inputvektor \vec{x} , Gewichten \vec{w} und dem Bias b . Das Neuron verwendet eine LeakyReLU Aktivierungsfunktion mit $\alpha = \frac{1}{4}$ und gibt für die Eingabe \vec{x} , die Ausgabe a aus. Die Ausgabe a wurde für Sie bereits berechnet.

Führen Sie einen Backpropagation-Schritt mit der Fehlerfunktion L und Label \hat{y} durch. 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 $\vec{x} = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$, Gewicht $\vec{w} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ und Bias $b = 1$.

Ausgabe $a = -1$, Label $\hat{y} = 0$ und Lossfunktion $L = \frac{1}{2}(\hat{y} - a)^2$.

Als Hilfestellung geben wir Ihnen die Neuronenformel an:

$$z = \sum_k w_k \cdot x_k + b$$

$$a = \text{LeakyReLU}(z) = \begin{cases} z, & z > 0 \\ \alpha z, & z \leq 0 \end{cases}$$

$$\begin{aligned} z &= -2 \cdot 1 + 1 \cdot (-3) + 1 \\ &= -4 \end{aligned}$$

$$\frac{\partial L}{\partial a} = -(\hat{y} - a) = -1$$

$$\frac{\partial a}{\partial z} = \begin{cases} 1 & z > 0 \\ \frac{1}{4} & z \leq 0 \end{cases} = \frac{1}{4}$$

$$\frac{\partial z}{\partial w_0} = x_0 = 1$$

$$\frac{\partial z}{\partial w_1} = x_1 = -3$$

$$\frac{\partial z}{\partial b} = 1$$

$$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_0} = -1 \cdot \frac{1}{4} \cdot 1 = -\frac{1}{4}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_1} = -1 \cdot \frac{1}{4} \cdot (-3) = \frac{3}{4}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial b} = -1 \cdot \frac{1}{4} \cdot 1 = -\frac{1}{4}$$

- f) Das Neuron aus der vorherigen Aufgabe erhält nun eine andere Eingabe und Label. (___/1.5P)
 Dabei errechnete es im Backpropagation-Schritt folgende Gradienten:

$$\frac{\partial L}{\partial w_0} = 2$$

$$\frac{\partial L}{\partial w_1} = 1$$

$$\frac{\partial L}{\partial b} = -1$$

$$\Delta w^{(n)} = -\eta \cdot \frac{\partial L}{\partial w^{(n)}} \quad \alpha = \gamma$$

$$w^{(n+1)} = w^{(n)} + \Delta w^{(n)}$$

Führen Sie mit den vorgegebenen Gradienten, der Lernrate $\eta = 0.1$ und den Gewichten und Bias aus der vorhergegangenen Aufgabe einen Gewichts-Update-Schritt durch.
 Geben Sie $w^{(n+1)}$ und $b^{(n+1)}$ an.

$$w_0^{(n+1)} = w_0^{(n)} + (-\gamma \cdot \frac{\partial L}{\partial w_0}) = -L + (-0.1 \times 2) = -2.2$$

$$w_1^{(n+1)} = w_1^{(n)} + (-\gamma \cdot \frac{\partial L}{\partial w_1}) = 1 - 0.1 = 0.9$$

$$b^{(n+1)} = b^{(n)} + (-\gamma \cdot \frac{\partial L}{\partial b}) = 1 + 0.1 = 1.1$$

- (g) Inwiefern ändert sich das Gewichtsupdate wenn, statt einem Eingabevektor, ein Mini-batch mit mehreren Eingabekörpern für das Training verwendet wird? (___/1P)

$$w_k^{(n+1)} = w_k^{(n)} - \eta \frac{1}{\text{Batch Size}} \sum_{i=1}^{\text{Batch Size}} \frac{\partial L_i}{\partial w_k^{(n)}}$$

$$b^{(n+1)} = b^{(n)} - \eta \frac{1}{\text{Batch Size}} \sum_{i=1}^{\text{Batch Size}} \frac{\partial L_i}{\partial b^{(n)}}$$

faster update cycles

$$J_c = \frac{1}{B} \sum_{b=1}^B J_c(b, \hat{y}_b)$$

Aufgabe 4 Convolutional Neural Networks

_____ / 8 Punkte

- a) Wieso werden GPUs statt CPUs für das Training und Ausführen von CNNs verwendet? (____/1P)

✓ 6, P53

Parallelization on GPUs: GPUs are optimized for computer graphics which are mostly matrix operations. All mathematical functions of CNNs / NNs can be represented as matrix operations.

b)

- Heutzutage verwendet fast jedes neuronale Netz Skip-Connections. (____/3P)

1. In welcher Netzarchitektur wurden sie erstmalig vorgestellt?
2. Welches Problem wird durch die Verwendung von Skip-Connections umgangen?
3. Was ist die aktuelle Theorie weshalb Skip-Connections dieses Problem umgehen?

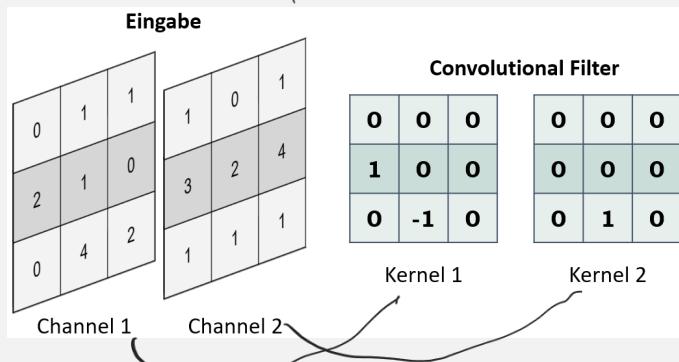
✓ 6, h55

h ResNet

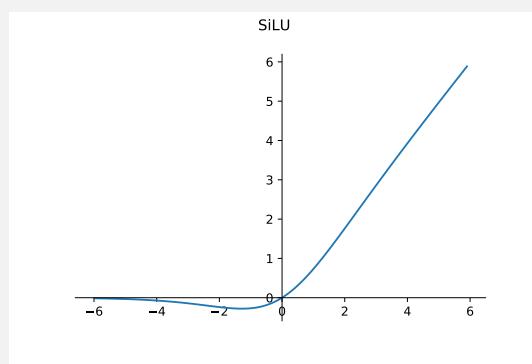
1, the problem of vanishing or exploding gradients

2, identity parallel to each residual block
add identity to output of block
Learn the change of existing feature maps.

- c) Gegeben ist ein 3×3 Eingabebild mit 2 Channels und ein 3×3 Convolutional Filter. (___/4P)



- Wie viele Bias-Terme würde dieser Convolutional Filter enthalten? (1P) 06. 24
- Nehmen sie an, dass alle Bias-Terme = 0 und kein Padding verwendet wird: Welche Ausgabe erzeugt dieser Filter? (1P)
- Die Ausgabe des Filters soll mit der SiLU Funktion aktiviert werden (siehe unten). Was ist die Ausgabe der Funktion? Runden Sie auf 2 Nachkommastellen. (1P)
- Nennen Sie einen Grund SiLU statt ReLU für ein CNN zu verwenden. (1P)



$$\text{SiLU} = x \cdot \frac{1}{1+e^{-x}}$$

l, l

$$2, 2^{-4} + 1 = -1$$

$$3, -1 \times \frac{1}{1+e^1} = -0.27$$

4, Smoother activation function, continuous derivative and potentially better handling of deeper

ReLU: $x < 0$ using ReLU

networks.

Aufgabe 5 Reinforcement Learning**_____ /11 Punkte**

- a) Nennen Sie zwei Gründe, warum die Berechnung der Zustandswertfunktion für die optimale Policy $V^*(s)$ rekursiv durchgeführt werden sollte, anstatt mit einer direkten Methode. (____/2P)

1. Reduce computational complexity

1. ~~Bellman equation~~

the true model of the environment is usually unknown

- b) Erläutern Sie, warum die Epsilon-Greedy-Policy für die Verbesserung der Policy in modellfreien RL-Ansätzen wichtig ist. (____/1P)

12, p14

Because with Epsilon-Greedy policy it can avoid to get stuck on a suboptimal action forever. There exists an "exploration-exploitation tradeoff"

- c) Begründen Sie, warum es sinnvoll ist, ein neuronales Netz zur Approximation der Zustandswertfunktion für komplexe Aufgaben, wie z.B. das autonome Fahren, zu verwenden. (____/1P)

f (fMRI)

Reduce computational complexity

Because RL is often applied in more complex problems, that is impossible to store all values in a table.

- d) Ein RL-Agent hat in einer Umgebung zwei Trajektorien gesammelt, die alle von Startzustand s_{start} ausgehen. Dabei wurde eine stochastische Policy π verfolgt. (___/2.5P)

- In der **ersten** Trajektorie hat der Agent die folgenden Belohnungen erhalten:
 $\{r_{start} = 0, r_1 = 2, r_2 = -1, r_3 = 3, r_4 = 0\}$
- In der **zweiten** Trajektorie hat der Agent die folgenden Belohnungen erhalten:
 $\{r_{start} = 0, r_1 = 1, r_2 = -1, r_3 = 2, r_4 = 2\}$

Berechnen Sie den Gewinn für jede dieser Trajektorien $\{G_1, G_2\}$ unter der Voraussetzung, dass $\gamma = 0,9$ ist. Evaluieren Sie anschließend die Policy π im Zustand s_{start} mithilfe der Zustandswertfunktion $\hat{V}_\pi(s_{start})$ anhand der gesammelten Trajektorien.

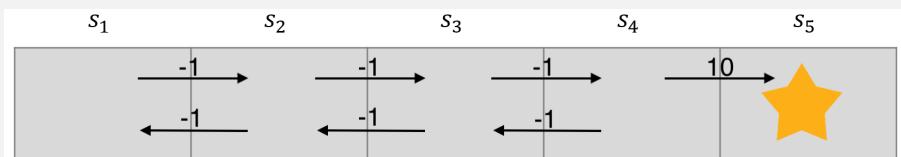
$$G_1 = r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \gamma^4 r_4 = 0 + 0.9 \times 2 + 0.9^2 \times (-1) + 0.9^3 \times 3 = 3.17$$

$$G_2 = r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \gamma^4 r_4 \\ = 0 + 0.9 \times 1 + 0.9^2 \times (-1) + 0.9^3 \times 2 + 0.9^4 \times 2 = 2.862$$

$$\hat{V}_{\pi}(s_{start}) = \frac{G_1 + G_2}{2} = 2.966$$

- e) Ein Roboter soll in einem Labyrinth schnell seinen Weg zum Terminalzustand s_5 finden. (___/4,5P)
 Die **deterministische** Umgebung wird als Markov'scher Entscheidungsprozess (MDP) mit den folgenden Eigenschaften modelliert:

- Aktionsraum $\mathcal{A} = \{\text{links, rechts}\}$
- Zustandsraum $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5\}$
- Terminalzustand s_5
- Belohnungsfunktion: Siehe Pfeile mit Belohnungen in der unteren Grafik.
- Discount Factor $\gamma = 0,9$



Es gibt zwei Policies im Hypothesenraum $\Pi = \{\pi_1, \pi_2\}$

- π_1 ist **deterministisch**:
 - $\pi_1(\text{rechts}|s) = 1, \quad \forall s \in \{s_1, s_2, s_3, s_4\}$
- π_2 ist **stochastisch**:
 - $\pi_2(\text{rechts}|s_1) = 1$
 - $\pi_2(\text{links}|s_i) = \pi_2(\text{rechts}|s_i) = 0,5, \quad \forall s \in \{s_2, s_3, s_4\}$

Berechnen Sie die Zustandswertfunktion für beide Policies $V_{\pi_1}(s)$ und $V_{\pi_2}(s)$ für alle Zustände. Erläutern Sie, welche Policy die Bessere ist.

Wichtig: Berechnen Sie $V_{\pi_2}(s)$ nur bis zu Iteration ($K = 2$) und nehmen Sie an, dass $V_{\pi_1}(s)$ und $V_{\pi_2}(s)$ mit 0 initialisiert sind.

$$\begin{aligned} \pi_1: \quad V_{\pi_1}(s_4) &= 1 \times (-1 + 0.9 \times 0) = -1 \quad V_{\pi_1}(s_3) = 1 \times (-1 + 0.9 \times -1) = -8 \\ V_{\pi_1}(s_2) &= 1 \times (-1 + 0.9 \times 0) = 0.9 \quad V_{\pi_1}(s_1) = 1 \times (-1 + 0.9 \times 0.9) = 0.41 \end{aligned}$$

$$\pi_2: \quad K=0 \quad \boxed{0} \quad \boxed{0} \quad \boxed{0} \quad \boxed{0} \quad \boxed{0}$$

$$k=1 \quad v(s_1) = -1$$

$$v(s_2) = 0.5 \times (-1) + 0.5 \times (-1) = -1$$

$$v(s_3) = -1$$

$$v(s_4) = -0.5 + 0.5 \times 0 = 0.5$$

$$\boxed{-1 \quad -1 \quad -1 \quad 0.5 \quad 0}$$

$$k=2 \quad v(s_1) = 1 \times [-1 + 0.9 \times (-1)] = -1.9$$

$$v(s_2) = 2 \times 0.5 \times [-1 + 0.9 \times (-1)] = -1.9$$

$$v(s_3) = 0.5 \times [-1 + 0.9 \times (-1)]$$

$$+ 0.5 \times [-1 + 0.9 \times 0.5] = 0.575$$

$$v(s_4) = 0.5 \times [-1 + 0.9 \times (-1)] + 0.5 \times 0 \\ = 0.05$$

$$\boxed{-1.9 \quad -1.9 \quad 0.575 \quad 0.05 \quad 0}$$

π_1 ist besser

Aufgabe 6 HMM, Bayes, Entscheidungsbäume _____/12 Punkte

- a) Worin unterscheidet sich ein Hidden Markov Model (HMM) von einem diskreten Markov Prozess? (___/1P)

- b) Nennen Sie die 3 grundlegenden Probleme aus der Vorlesung, die sich mit Hidden Markov Modellen lösen lassen können. Beschreiben Sie darüber hinaus **eines** dieser 3 Probleme mit einem Stichpunkt. (___/2P)

- c) Zur Vorhersage von Verspätungen im Nahverkehr soll ein Naive Bayes Klassifikator eingesetzt werden. Die vorauszusagende Verbindung fährt bei schlechtem Wetter (W = Schlecht) und am Tag (T = Tag). Welche Klassifikation ist gemäß des Naive Bayes Ansatzes am wahrscheinlichsten? Geben Sie den Rechenweg sowie Ihre Schlussfolgerung an.

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	Gut	Nacht	Nein
5	Gut	Tag	Nein
6	Schlecht	Tag	Ja
7	Schlecht	Nacht	Ja
8	Gut	Nacht	Ja
9	Schlecht	Tag	Ja
10	Schlecht	Tag	Ja

10, P32

Classification of new examples: (Schlecht, Tag)

$$P(V=Ja) = \frac{5}{10} = \frac{1}{2} \quad P(V=Nein) = \frac{5}{10} = \frac{1}{2}$$

$$P(V=Ja)P(W=S|Ja)P(T|Ja) = \frac{1}{2} \times \frac{4}{5} \times \frac{3}{5} = \frac{6}{25}$$

$$P(V=Nein)P(W=S|Nein)P(T|Nein) = \frac{1}{2} \times \frac{1}{5} \times \frac{2}{5} = \frac{2}{25}$$

-> Classification: Verspätung = Ja

Normalized probability: $\frac{\frac{6}{25}}{\frac{6}{25} + \frac{2}{25}} = \frac{3}{4}$

- d) Geben Sie ein Beispiel in dem Entscheidungsbäume besser geeignet sind als neuronale Netze. (___/1P)

03.11.26

Decision trees may be more suitable than neural networks in scenarios requiring interpretable decisions and simple decision boundaries. For instance, in a creditworthiness assessment where clear and easily understandable decision criteria are essential, decision trees might be preferred.

- e) Welches Problem lösen Random Forests im Vergleich zu Bagging, wenn Entscheidungsbäume als Modell verwendet werden? Durch welche Modifikation des Bagging Algorithmus wird dies erreicht? (___/2P)

03.11.26

Problem: Models are highly correlated

Modifikation: Bootstrap \rightarrow Training \rightarrow Aggregation

Before each split, sample a subset
of ~~all~~ attributes as possible
candidates for splitting

f)

(___ /3P)

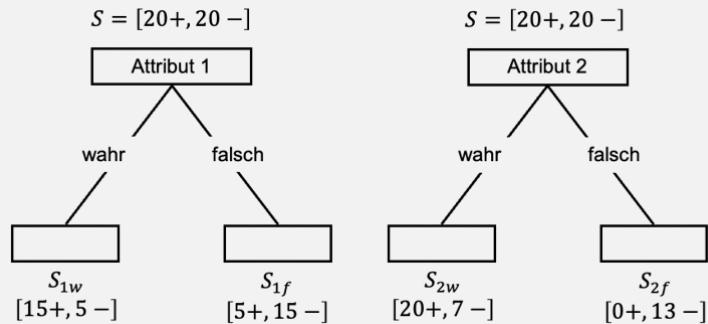


Abbildung 1: Schreibweise: [Anzahl positive Bsp. (+), Anzahl negative Bsp. (-)]

In der oben dargestellten Abbildung finden Sie den Vergleich zweier Attribute A_1 (links) und A_2 (rechts), die als Testattribute eines Entscheidungsbaumes 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.

$$IG(S, A) = H(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} H(S_v)$$

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

Attribut 1: $IG(S_1) = - \frac{15}{20} \log_2 \frac{15}{20} - \frac{5}{20} \log_2 \frac{5}{20} = 1$

$ID(S_{1w}) = - \frac{15}{20} \log_2 \frac{15}{20} - \frac{5}{20} \log_2 \frac{5}{20} = 0.81$

$ID(S_{1f}) = - \frac{5}{20} \log_2 \frac{5}{20} - \frac{15}{20} \log_2 \frac{15}{20} = 0.81$

$IG(S, \text{Attribut 1}) = 1 - \frac{15}{20} \times 0.81 - \frac{5}{20} \times 0.81 = 0.17$

Attribut 2: $IG(S_2) = - \frac{20}{20} \log_2 \frac{20}{20} - \frac{7}{20} \log_2 \frac{7}{20} = 1$

$ID(S_{2w}) = - \frac{20}{27} \log_2 \frac{20}{27} - \frac{7}{27} \log_2 \frac{7}{27} = 0.73$

$ID(S_{2f}) = - \frac{7}{27} \log_2 \frac{7}{27} - 1 \log_2 1 = 0$

$IG(S, \text{Attribut 2}) = 1 - \frac{20}{27} \times 0.73 - \frac{7}{27} \times 0 = 0.44$

Attribut besser