

**Klausur zur Lehrveranstaltung
„Maschinelles Lernen 1 – Grundverfahren“
(60 Minuten)**

Nachname:	Vorname:
Matrikelnummer:	Studiengang:

Anmerkungen

- Tragen Sie Nachname, Vorname, Matrikelnummer und Studiengang deutlich lesbar ein.
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 Rot- oder 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 - 14) überzeugt habe.

Unterschrift

Nur für den Prüfer

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

Aufgabe 1 - Lernen und Entscheidungsbäume**13 Punkte**

- a) Welche Ziele verfolgen Induktion und Deduktion? Wie erreichen Sie diese Ziele?

01. Pg (/2P)

Inductive reasoning is a method of reasoning, which a general principle is derived from a body of specific observations.

Deductive reasoning is a method of reasoning that starts with a general principle and examines the possibility to reach a

- b) Definieren Sie Konsistenz und Vollständigkeit im Hypothesenraum.

Sprechende, bsp (/1P)

Consistent: No negative examples are classified positive

Complete: All positive examples are classified positive,

- c) Ausgehend von der lerntheoretischen Abschätzung des realen Fehlers; von welchen drei Kriterien hängt der Lernerfolg einer Lernmaschine ab?

02 Pg (/1,5P)

Capacity of learning system

Optimization method

Learning instances in dataset

- d) Welche zwei entscheidenden Designentscheidungen müssen für den k-Means getroffen werden?

03 Pg (/2P)

Welche Probleme können durch falsche Entscheidungen entstehen?

GPy

Determine the number of clusters (k); Select initial cluster centers.

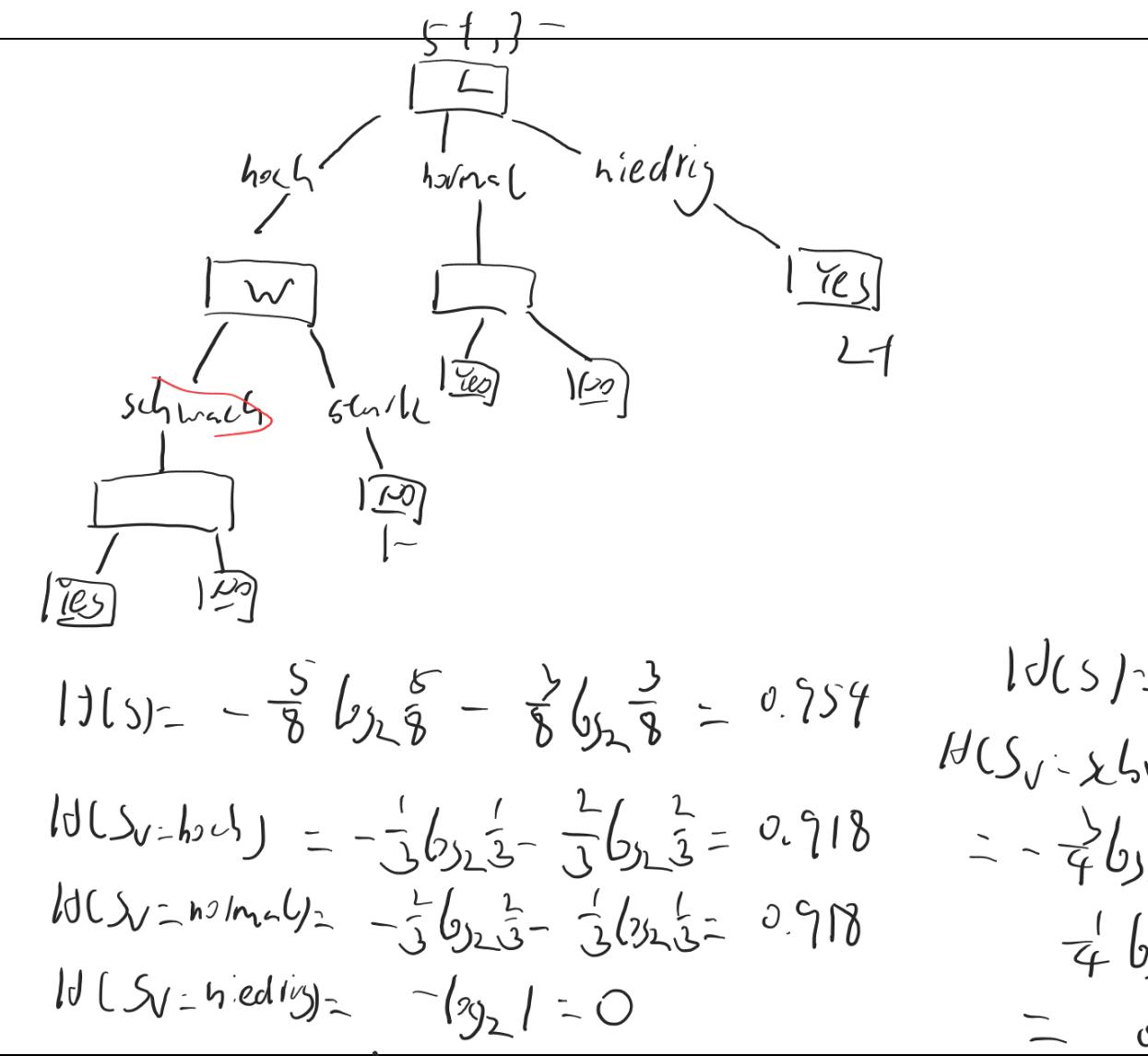
Problems:

- Suboptimal cluster formation (Wrong number of initial clusters)
- Overfitting or Underfitting

- e) Erstellen Sie für den gegebenen Datensatz mithilfe des ID3 Algorithmus einen Entscheidungsbaum und begründen Sie Ihre Entscheidungen rechnerisch.
- Hinweis:** Verwenden Sie den \log_2 und runden Sie auf

(___ / 6,5P)

	Luftfeuchtigkeit (L)	Wind (W)	Tennis? (T)
1	hoch	schwach	Nein
2	hoch	stark	Nein
3	hoch	schwach	Ja
4	normal	stark	Ja
5	niedrig	schwach	Ja
6	normal	stark	Nein
7	niedrig	schwach	Ja
8	normal	stark	Ja



$$H(S) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} = 0.954$$

$$H(S_{V=hoch}) = 0.954$$

$$H(S_{V=normal}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.918$$

$$= -\frac{1}{4} \log_2 \frac{1}{4} -$$

$$H(S_{V=niedrig}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.918$$

$$-\frac{1}{4} \log_2 \frac{1}{4}$$

$$H(S) - \frac{1}{3} \cdot H(S_{V=normal}) - \frac{2}{3} \cdot H(S_{V=niedrig}) = 0.265$$

$$= 0.711$$

$$0.954 - \frac{1}{3} \cdot 0.918 - \frac{2}{3} \cdot 0.918 = 0.265$$

$$H(S_{V=stark}) =$$

$$-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4}$$

$$0,984 - \frac{1}{8} \cdot 0,811 - \frac{1}{8} \cdot 1 = 0,0475$$

Aufgabe 2 – Neuronale Netze ζ is better

/ 13P

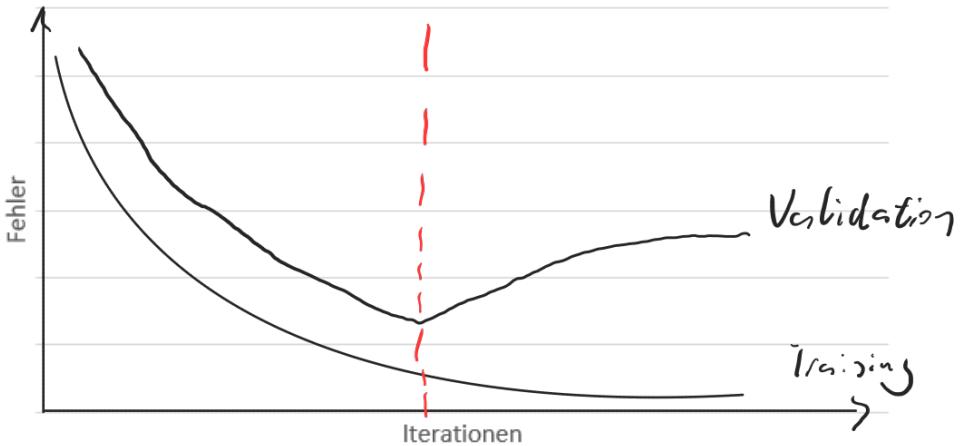
- a) Geben Sie die in der Vorlesung vorgestellte ReLU Funktion und deren Ableitung an. (___ / 1P)

$$f(x) = \max(0, x) \quad f'(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

- b) Während des Trainings bemerken Sie, dass die Genauigkeit auf dem Trainingsdatensatz ansteigt und die Genauigkeit auf dem Validierungsdatensatzes sinkt. (___ / 1,5P)
- Benennen Sie das Phänomen
 - Skizzieren Sie den typischen Verlauf des Trainingsfehlers und des Validierungsfehlers.
- OS, P34

Hinweis: Markieren Sie die beiden Fehlerkurven eindeutig mit Training und Validierung

Typischer Verlauf des Fehlers



Overfitting

- c) Gegeben ist ein Perzeptron mit Inputvektor \vec{x} , Gewichten \vec{w} und dem Bias b . (___/5P)
 Das Perceptron verwendet ReLU als Aktivierungsfunktion und gibt für die Eingabe \vec{x} , die bereits berechnete Ausgabe a aus.
 Führen Sie einen Backpropagation-Schritt mit der Fehlerfunktion L und Label \hat{y} durch und geben Sie die Gradienten der neuen Gewichte und des Bias an.
 Verwenden Sie anschließend diese Gradienten, um die Gewichte und den Bias mittels Gradientenabstieg zu verändern. Verwenden Sie dafür die Lernrate η .

$$\vec{x} = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \quad \vec{w} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad b = 1$$

$$\begin{aligned} z &= \sum_k w_k * x_k + b \\ a &= \underbrace{\text{ReLU}(z)}_{\text{ReLU}(2)} \\ L &= \frac{1}{2} * (\hat{y} - a)^2 \end{aligned}$$

$$\begin{aligned} w_1^{new} &= w_1^n + (-g) \frac{\partial L}{\partial w_1^n} \\ &= 1 - 0.2 = 0.8 \end{aligned}$$

$$\begin{aligned} \hat{y} &= 4, \quad a = 3, \quad \eta = 0.1, \quad b^{new} = b^n + (-g) \cdot \frac{\partial L}{\partial b^n} \\ &\quad \vdots \\ &= 1 + 0.1 - 1.1 \end{aligned}$$

$$z = 2x_1 - 2x_2 + 1 = 3$$

$$\frac{\partial L}{\partial a} = -(\hat{y} - a) = -1 \quad \frac{\partial a}{\partial z} = \begin{cases} 1 & z \ge 0 \\ 0 & z < 0 \end{cases} = 1$$

$$\frac{\partial L}{\partial w_2} = x_2 = 2 \quad \frac{\partial L}{\partial w_1} = x_1 = -2 \quad \frac{\partial L}{\partial b} = 1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_2} = -1 \times 1 \times 2 = -2$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_1} = -1 \times 1 \times (-2) = 2$$

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

$$w_2^{new} = w_2^n + (-g) \cdot \frac{\partial L}{\partial w_2^n} = 2 + 0.2 = 2.2$$

- d) Was gilt es im Zusammenhang mit dem Dropout zu beachten, wenn das Training abgeschlossen wurde und das Netzwerk in der Praxis eingesetzt werden soll? (___/1P)

Multiply the output of each neuron
with its dropout probability p

- e) Erweitern Sie die gegebene Fehlerfunktion mit einem geeigneten Regularisierungsterm, um das in Aufgabe c) benannte Phänomen zu verhindern. Benennen Sie die von Ihnen verwendeten Variablen.

$$L = \frac{1}{N} \sum_N (\hat{y}_n - a_n)^2$$

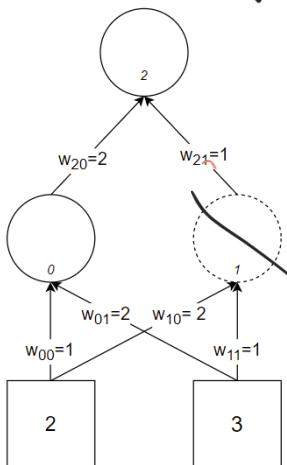
L = Loss, N = Anzahl an Samples im Minibatch, a = Output, \hat{y} = Target

$$L = \frac{1}{N} \sum_{n=1}^N (\hat{y}_n - a_n)^2 + \lambda \frac{1}{2} \|w\|_L^2$$

λ : Hyperparameter w : weight

- f) Bei der untenstehenden Abbildung eines KNN wird Dropout mit $p = 50\%$ verwendet. Berechnen Sie die Ausgaben der Neuronen 0, 1 und 2, wenn Neuron 1 in diesem Iterationsschritt ausgeschaltet wurde. Verwenden Sie als Aktivierungsfunktion für alle Neuronen ReLU und den Wert 0 für alle Biase.

$$f(x) = \max(0, x)$$



Neuron 0: Input: $2x1 + 3x2 = 8$

Output: $\max(0, 8) = 8$

Neuron 1: no output

Neuron 2: Input: $8x1 = 8$

Output: $\max(0, 8) = 8$

Aufgabe 3 – Convolutional Neural Networks**— / 5P**

- a) Die Größe eines RGB Eingabebildes eines CNNs beträgt 100x100 Pixel. In der ersten Schicht eines CNNs befinden sich 128 Convolutional Filter mit Kernelgröße 10x10, Stride = 2 und ohne Padding. In Schicht 2 folgt ein Max Pooling Layer mit Kernelsize 2x2 und in Schicht 3 ein Fully-Connected Layer mit 2 Neuronen (___/3,5P)
1. Berechnen Sie die Anzahl der trainierbaren Parameter der ersten Schicht. (1 Punkt)
 2. Welche Dimension hat die Feature Map nach der ersten Schicht? (1 Punkt) **06 P29?**
 3. Berechnen Sie die Anzahl der trainierbaren Parameter in Schicht 2 und Schicht 3. (1,5 Punkte) **06, P5F**

$$1, (10 \times 10 \times 3 + 1) \times 128 = 301 \times 128$$

$$2, \frac{100 - 10}{2} + 1 = 46$$

$$46 \times 46 \times 128$$

3, Schicht 2: **no learnable parameters**

Schicht 3: $25 \times 25 \times 128 \times 2 + 2$

- b) Wofür werden entweder Pooling Layer oder Strided Conv. Layer verwendet? (___/1,5P)
Nennen Sie für beide Verfahren jeweils ein Vorteil.

Reduce the Resolution of feature map

Pooling Layer: aggregation of small image areas in feature maps. local translation invariance

Strided Conv. Layer: decreases feature map resolution, Dimension Reduction

Aufgabe 4 – Support Vector Machines**___ / 7P**

- a) Formulieren Sie das grundlegende Optimierungsproblem für einen Support Vector Klassifikator. Geben Sie außerdem die Nebenbedingungen für die Optimierung an.

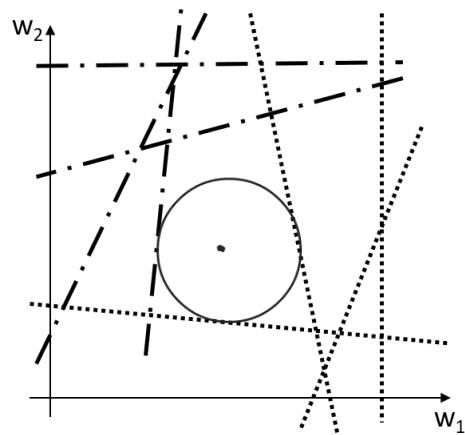
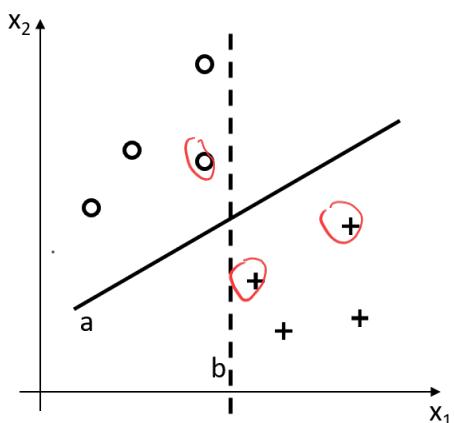
07. Pjz Pjz

Find the Hyperplane $\{x \in \mathbb{R}^2 | \vec{w}^T \vec{x} + b = 0, \vec{w}, b \in \mathbb{R}^2\}$, with maximum margin. minimize $\|\vec{w}\|^2$
 $y_i(\vec{w}^T \vec{x}_i + b) \geq 1, i=1 \dots n$

b)

(___ / 3P)

- Welche der beiden Hypothesen a oder b im linken Graph ist die optimale Hypothese?
- Markieren Sie die entsprechenden Stützvektoren (Support Vectors) im linken Graphen und zeichnen Sie die Hypothese in den rechten Graphen ein.



a,

- c) Erläutern Sie die Intention des Kernel Tricks in Bezug zu Support Vector Machines. Benennen Sie zusätzlich zwei aus der Vorlesung bekannte Kernel Funktionen. (___/2P)

07, p28

Instead of calculating the complex nonlinear transformation $\varphi(\cdot)$ and the scalar product, we use a kernel that does this implicitly, thus saving a lot of computational operations.

$$\text{dot product: } k(\bar{x}', \bar{y}') = \bar{x}' \cdot \bar{y}'$$

$$\text{Polynomial: } K(\bar{x}', \bar{y}') = (\bar{x}' \cdot \bar{y}') + c)^d$$

- d) Nennen Sie neben des Kernel Tricks eine weitere Methode, um nicht-linear separierbare Daten mit einer SVM klassifizieren zu können. (___/1P)

Soft Margin SVM

Aufgabe 5 --Reinforcement Learning / 11P

- a) Erklären Sie kurz was es bedeutet, wenn ein Ansatz modellfrei ist.

12. P17

(___/1P)

No explicit model necessarily
 without the true model of the environment,
 data is generated by interacting with
 environment.

- b) Erläutern Sie was der Vorteil von Reinforcement Learning mit
 Funktionsapproximation z.B: durch neuronale Netze gegenüber
 tabellarischem Reinforcement Learning ist.

(___/1P)

Reduce Computational complexity
 and can solve more complex
 problem

- c) Beschreiben Sie wie ein Target Network funktioniert und welches Problem es
 löst.

12. P49 49

The target network uses the parameters φ' ; Periodic update of the
 target network parameters; Works the correlation of Q-value and target
 changing target values

- d) Warum ist ein Strategiebasiertes Verfahren wie der REINFORCE Algorithmus
 ein On-Policy Algorithmus?

12. P32

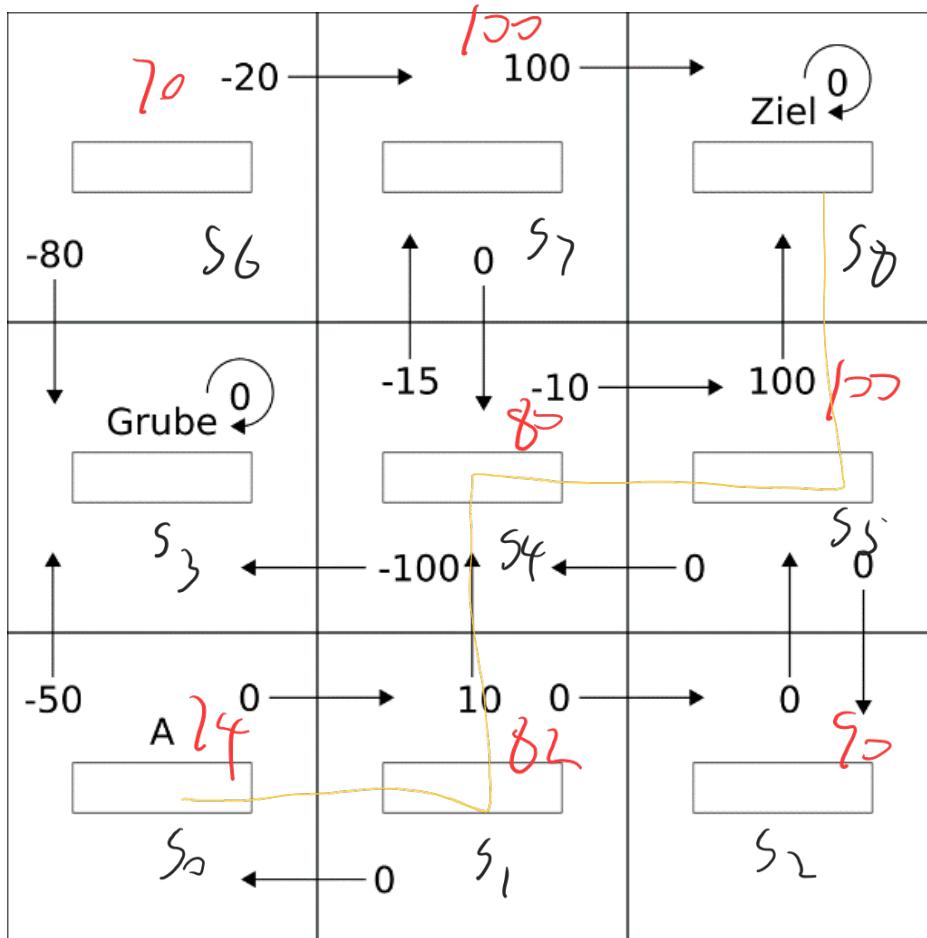
Because the target depends on the next
 action a' that must be chosen based on
 the current policy π_t .

- e) Nennen Sie einen Vorteil und einen Nachteil von strategiebasierten Verfahren. (___/1P)

*Flexibility in addressing complex problems
High expertise requirements*

- f) Betrachten Sie die untenstehende Umgebung. 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 ($V^*(s)$) dieser Strategie in die entsprechenden Kästen ein (Diskontierungsfaktor $\gamma = 0,9$). Runden Sie Ihre Ergebnisse auf ganze Zahlen.

Zeichnen Sie zusätzlich den Pfad der optimalen Strategie von Zustand A zum Zielzustand ein.



Aufgabe 6 –Bayes, HMM, SPN / 11P

a) Es sei

$$h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)$$

die Maximum a posteriori (MAP) Hypothese aller Hypothesen $h \in H$, gegeben
der beobachteten Daten D . Geben Sie die vereinfachte Variante von MAP an,
bei der Sie kein a priori Wissen über die Hypothesen haben?

P. P12 (____/1P)

$$h_{ML} = \arg \max_{h \in D} P(D|h)$$

- b) Erläutern Sie, was **Hidden** (deutsch. versteckt) und **Markov** in einem Hidden Markov Model (HMM) bedeuten. (____/2,5P)
 Wie können Sie ein HMM von einem Markov Entscheidungsprozess unterscheiden?

- c) Zur Vorhersage von Verspätungen im Schienenverkehr soll ein Naiver Bayes Klassifikator eingesetzt werden. Folgende Daten sind gegeben: (___ /3P)

Die vorauszusagende Verbindung fährt bei schlechtem Wetter ($W = \text{Schlecht}$) und in der Nacht ($T = \text{Nacht}$). Was ist die wahrscheinlichste Klassifikation (Verspätung ja/nein) gemäß des Naiven Bayes Ansatz? Geben Sie den Rechenweg an.

12.152

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

$$P(V=\text{ja}) = \frac{5}{10} = \frac{1}{2} \quad P(V=\text{nein}) = \frac{5}{10} = \frac{1}{2}$$

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

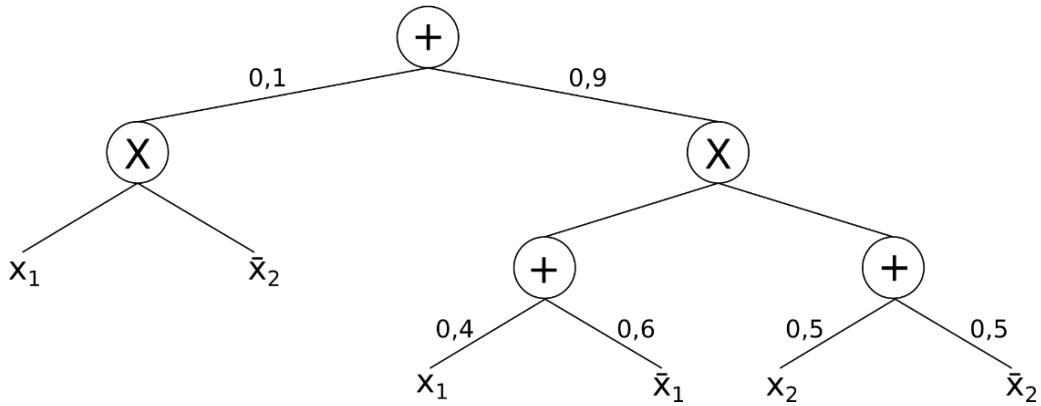
$$P(V=\text{nein}) P(W=\text{S} | V=\text{nein}) P(T=\text{N} | V=\text{nein}) = \frac{1}{2} \times \frac{1}{5} \times \frac{3}{5} = \frac{3}{25}$$

$$\frac{6}{25} > \frac{3}{25}$$

Classification: Verspätung = ja

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

- d) Berechnen Sie die Wahrscheinlichkeit $P(x_1 = 0, x_2 = 1)$ für das gegebene SPN, welches die Verteilung $P(X)$ repräsentiert. Geben Sie alle Zwischenergebnisse für alle Knoten an. Sie sollen dafür in die Abbildung schreiben. (___/2,5P)



- e) Gegeben sei das folgende SPN. Prüfen Sie, ob es valide, vollständig und konsistent ist. Geben Sie dabei die Scopes aller Knoten an. Sie dürfen hierzu in die Abbildung schreiben. (___/2P)

