

**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 - 18) überzeugt habe.

Unterschrift

Nur für den Prüfer

Aufgabe	1	2	3	4	5	6	Gesamt	Note
Punkte	12	10	10	5	11	12	60	
Erreicht								

Aufgabe 1 – Lerntheorie und Unsupervised Learning**___ / 12P**

- a) Verfahren des maschinellen Lernens lassen sich in unterschiedliche Kriterien einordnen. Nennen Sie zwei davon und geben Sie deren Ausprägungen oder Abstufungen an

08. P3

Supervised learning; Data with labels y is given
goal: function approximation $x \rightarrow y$

Unsupervised learning; Unlabeled data is given
Goal: Structure approximation $x \approx x'$

- b) Beschreiben Sie die Problematik des „Overfitting“. Wie kann dies verhindert werden

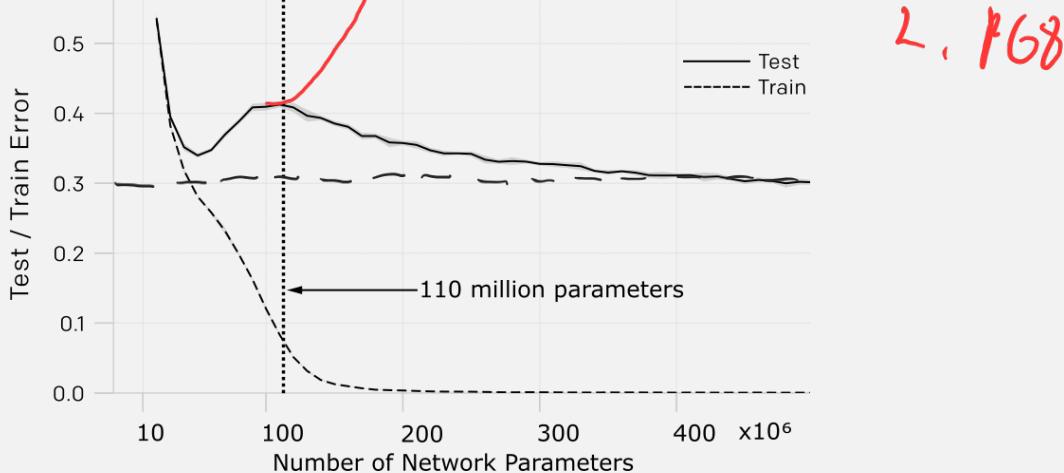
02. P19 PLL

A hypothesis overfits the training examples, if some other hypothesis, that fits the training examples less well, actually performs better over the entire distribution of instances.

Learning system memorizes training data rather than early stopping learning the underlying structure

- c) Neuronale Netze zeigen ein bisher unerklärbares Phänomen welches „Deep Double Descent“ genannt wird und scheinbar den klassischen Regeln des maschinellen Lernens widerspricht. (___/4P)

Das folgende Schaubild zeigt dieses Phänomen. Auf der X-Achse befindet sich die Anzahl der trainierbaren Netzwerkparameter in Millionen und auf der Y-Achse der dazugehörige Trainings- und Test-Loss.



Ab 110 Millionen Parametern (rechts der senkrecht gestrichelten Linie) verhält sich der Test-Loss ungewöhnlich.

- Wieso ist ab dieser Position der Verlauf des Test-Loss ungewöhnlich? Ihre Erklärung sollte die Anzahl der Parameter in Kontext mit dem Test-Loss setzen.
Sie dürfen ihre etwaige Lösung aus b) referenzieren.
- Zeichnen Sie in das Schaubild wie die klassische ML-Lehre den Test-Loss vermuten würde.

i) With increasing complexity of hypothesis space the model starts to capture the underlying structure that leads to test error reduce.

- d) Definieren Sie den Begriff des induktiven Bias (___/1P)

~~Set of assumptions or prior knowledge that a learning system incorporates to generalize from data~~
Certain hypotheses are preferred over other hypotheses in the hypothesis space.

e) Welche zwei Bias-Arten wurden in der Vorlesung vorgestellt?

(__/1P)

Inductive Bias

Inherent Bias

f) Welche zwei entscheidenden Designentscheidungen müssen für den k-

Means festgelegt werden?

Welche Probleme können dadurch entstehen?

(__/2P)

Number of Clusters (k) Initialization of Cluster Centers

Suboptimal cluster formation

overfitting or underfitting

wrong number of initial clusters

Aufgabe 2 – Neuronale Netze / 10P

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

$$04 \quad 1472$$

$$f(x) = \begin{cases} x, & x > 0 \\ ax, & x \leq 0 \end{cases} \quad \frac{df(x)}{dx} = \begin{cases} 1, & x > 0 \\ a, & \text{else} \end{cases}$$

- b) Warum wird bei Dropout die Ausgabe der Neuronen im Training mit der Dropoutwahrscheinlichkeit dividiert, bzw. in der Inferenz multipliziert? (/1,5P)

① During training backpropagation averages gradients of all different subnets used in minibatch and updates the weights
 ② During testing, we want to make one final prediction
 $\hat{y} = \prod_i h_{\Theta}(x^{(i)}, \mu)$ where μ is the dropout rate.

- c) Nennen Sie eine Eigenschaft, die jede Aktivierungsfunktion besitzen muss. (/0,5P)

hot-linear

d) Gegeben ist ein Neuron mit Inputvektor \vec{x} , Gewichten \vec{w} und dem Bias (____/4P)

b. Das Neuron verwendet eine LeakyReLU Aktivierungsfunktion mit $\alpha = \frac{1}{3}$

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 $\frac{\partial L}{\partial a}, \frac{\partial a}{\partial z}$ und $\frac{\partial z}{\partial w_0}, \frac{\partial z}{\partial w_1}, \frac{\partial z}{\partial b}$ an.

Eingabevektor: $\vec{x} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$,

Gewicht und Bias: $\vec{w} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, b = 2$

Lossfunktion, Label, Ausgabe: $L = (\hat{y} - a)^2, \hat{y} = 0, a = -2$,

Als Hilfestellung geben wir Ihnen die Neuronenformel an:

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

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

$$\frac{\partial L}{\partial a} = -2(\hat{y} - a) = -2(0 - (-2)) = -4$$

$$\frac{\partial a}{\partial z} = \begin{cases} 1, & z \geq 0 \\ \alpha = \frac{1}{3}, & \text{else} \end{cases} \quad z = -2 \cdot 3 + 1 \cdot (-2) + 2 = -6 \quad \leftarrow$$

$$\frac{\partial z}{\partial w_0} = x_0 = 3$$

$$\frac{\partial z}{\partial w_1} = x_1 = -2$$

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

$$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_0} = -4 \times \frac{1}{3} \times 3 = -4$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_1} = -4 \times \frac{1}{3} \times (-2) = \frac{8}{3}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial b} = -4 \times \frac{1}{3} \times 1 = -\frac{4}{3}$$

- e) Das Neuron aus der vorherigen Aufgabe erhält nun eine andere Eingabe und Label. Dabei errechnete es im Backpropagation-Schritt folgende Gradienten: (___/1,5P)

$$\begin{aligned}\frac{\partial L}{\partial w_0} &= 5 \\ \frac{\partial L}{\partial w_1} &= 0 \\ \frac{\partial L}{\partial b} &= -5\end{aligned}$$

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 \bar{w}^{n+1} und \bar{b}^{n+1} an

$$\bar{w}_0^{n+1} = \bar{w}_0^n + (-y \cdot \frac{\partial L}{\partial w_0}) = -2 + (-0.1) \times 5 = -2.5$$

$$\bar{w}_1^{n+1} = \bar{w}_1^n + (-y \cdot \frac{\partial L}{\partial w_1}) = 1 + 0 = 1$$

$$\bar{b}^{n+1} = \bar{b}^n + (-y \cdot \frac{\partial L}{\partial b}) = 2 + 0.5 = 2.5$$

- f) Sie trainieren ein klassisches neuronales Netz mit Gewichtsregularisierung. Ab einem bestimmten Zeitpunkt ist die Ausgabe des Netzes konsistent mit dem Label und Sie erhalten einen Crossentropyloss mit Wert 0. Wenn sie nun das Netz weiter trainieren lassen: Verändern sich die trainierbaren Gewichte des Netzes weiter? Erläutern Sie Ihre Antwort. (___/1,5P)

The weights will be changed.
By applying parameter regularization the loss function is
 $\tilde{L} = L + \frac{\lambda}{2} \|\tilde{w}\|_2^2$ with $\lambda > 0$
Crossentropy = $\frac{\partial \tilde{L}}{\partial \tilde{w}} = \lambda \tilde{w}$
the gradient still exists,

Aufgabe 3 – Convolutional Neural Networks**___ / 10P**

- a) Bei Computer Vision Aufgaben werden heutzutage keine herkömmlichen neuronalen Netze verwendet sondern CNNs. Nennen Sie dafür zwei Gründe

Gründe

1. CNNs can use convolutional operations to detect local patterns
2. CNNs employs parameter sharing and weight sharing to reduce the number of parameters.

- b) Sie haben ein 70×70 großes RGB Eingabebild. Sie wollen Convolutional Kernel mit Größe 9×9 und Stride 1 verwenden. Die resultierende Feature Map soll die gleiche Dimension wie das Eingabebild haben
- Wie viele Convolutional Filter werden benötigt?
 - Auf welche Größe muss das ursprüngliche Bild gepaddet werden?
 - Wie viele trainierbare Parameter hat das resultierende Convolutional Layer?

06. P29

i)

$$\text{ii) } D = \frac{m-9}{1} + 1 \quad m = 78 \quad 78 \times 78 \times 3$$

iii)

$$(9 \times 9 \times 3) \times 3$$

- c) Welche wesentliche Technik wurde in der ResNet-Architektur eingeführt? Erklären Sie diese kurz

06. P55

Skip Connection: Identity parallel to each residual block
 add identity to output of block
 learn the change of existing feature map

- d) Sie haben folgende Feature Map mit nur einem Channel: (___/3P)

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

- i) Was ist die Ausgabe, wenn sie einen 2×2 Maxpool mit Stride 2 verwenden?

6 p34

3	0
2	0

- ii) Sie haben einen Convolutional Filter, welcher einen 3×3 Kernel besitzt, kein Padding verwendet, Stride = 1 benutzt und bei dem jeder Parameter (Gewichte und Bias) den Wert 1 besitzt. Verwenden Sie diesen Filter auf die ursprüngliche Feature Map und schreiben Sie die Ausgabe in den folgenden Tensor

1	1	1
1	1	1
1	1	1

6	5
4	4

1f2-1f3-2f2

- e) Sie trainieren ein CNN, welches Hunde, Katzen und Menschen in dieser Reihenfolge klassifizieren soll. Sie haben ein Bild von einer Katze von diesem CNN klassifizieren lassen und folgenden Output erhalten:

$$[1 \ 2 \ 3]$$

- i) Berechnen Sie zuerst den Softmax der Prädiktion.
ii) Berechnen Sie anschließend den Crossentropy Loss.

06. 137

Runden Sie auf 2 Nachkommastellen und verwenden sie für die Crossentropy den natürlichen Logarithmus.

$$\text{Hunde: } \frac{e^1}{e^1 + e^2 + e^3} \approx 0.09$$

$$\text{Katzen: } \frac{e^2}{e^1 + e^2 + e^3} \approx 0.24$$

$$\text{Menschen: } \frac{e^3}{e^1 + e^2 + e^3} \approx 0.67$$

$$\begin{aligned}\therefore \text{Crossentropy Loss} &= -\sum_j p_j \ln p_j \quad [0, 1, 0] \\ &= -1 \cdot \ln 0.24 \\ &\approx 1.43\end{aligned}$$

Aufgabe 4 – Support Vector Machine / 5P

- a) Beschreiben Sie in einem Satz auf welche Art die SVM Daten für die Klassifikation trennt und welches Kriterium dieser Trennung zu Grunde liegt. / 1P)

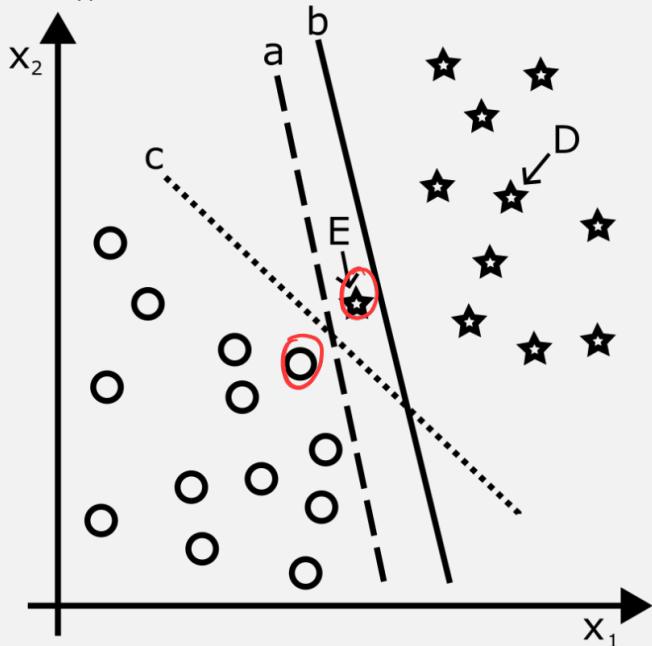
07. P12

Hyperplane with maximum margin
size of the margin determines the generalization capability

- b) Nennen Sie eine Möglichkeit um die SVM auch auf nicht-linear trennbaren Daten zu verwenden. / 1P)

Kernel-Trick

- c) Gegeben ist ein Datensatz mit den zwei Klassen **Stern** und **Kreis**. Dieser soll mit (___/3P) einer klassischen SVM korrekt klassifiziert werden. In der unteren Abbildung finden Sie eine grafische Darstellung der Datenpunkte und verschiedenen Hypothesen im Musterraum/Merkmalsraum.



- Geben Sie an welche der Hypothesen (a, b, c) das optimale Ergebnis des SVM Algorithmus auf diesem Datensatz ist.
- Markieren Sie die entsprechenden Stützvektoren (Support Vectors) in der Abbildung durch Einkreisen. (Info: Achten Sie auf eindeutige Darstellung, falsch markierte Stützvektoren führen zu Punkteverlust.)
- Wie verändert sich das Ergebnis aus ii), wenn der Datenpunkt D aus dem Datensatz entfernt wird?
- Wie verändert sich das Ergebnis aus ii), wenn der Datenpunkt E aus dem Datensatz entfernt wird?

i) C.

ii), no change

iii). the hyperplane will change and
mag close to b

Aufgabe 5 --Reinforcement Learning / 11P

- a) Durch welches Modell lässt sich die Problemstellung beim Reinforcement Learning formal darstellen? Welche vier Bestandteile werden für die Modellierung benötigt? (___/2P)

markov decision process

State Space (S) Action Space (A)

Transition Function (P), Reward Function (R)

- 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
problems

- c) Was ist der wesentliche Unterschied zwischen Reinforcement Learning und überwachtem Lernen mit Bezug auf die Fehlerberechnung? (___/1P)

In supervised learning, the error is measured between predicted and actual outputs, while in reinforcement learning, the error represents the difference between expected and estimated future rewards

- d) Ein Agent wird mit Hilfe von Q-Learning trainiert. Der Agent führt eine Aktion a in s aus und erhält dabei eine Belohnung r in Höhe von 1 und wird in einen neuen Zustand s' überführt. Die approximierte Aktionswertfunktion hat in s' ein Maximalwert von 10, Die approximierte Aktionswertfunktion hat für s, a einen Wert von 8, der Diskontierungsfaktor beträgt 0,9. Wie lautet die Formel zur Berechnung des TD-Fehlers? Berechnen Sie den TD-Fehler. 1L, p3 (___ /2P)

$$\begin{aligned}
 \delta_t &= r + \gamma Q_{\text{old}}(s'_t, a'_t) - Q_{\text{old}}(s_t, a_t) \\
 &= 1 + 0.9 \times 10 - 8 \\
 &= 2
 \end{aligned}$$

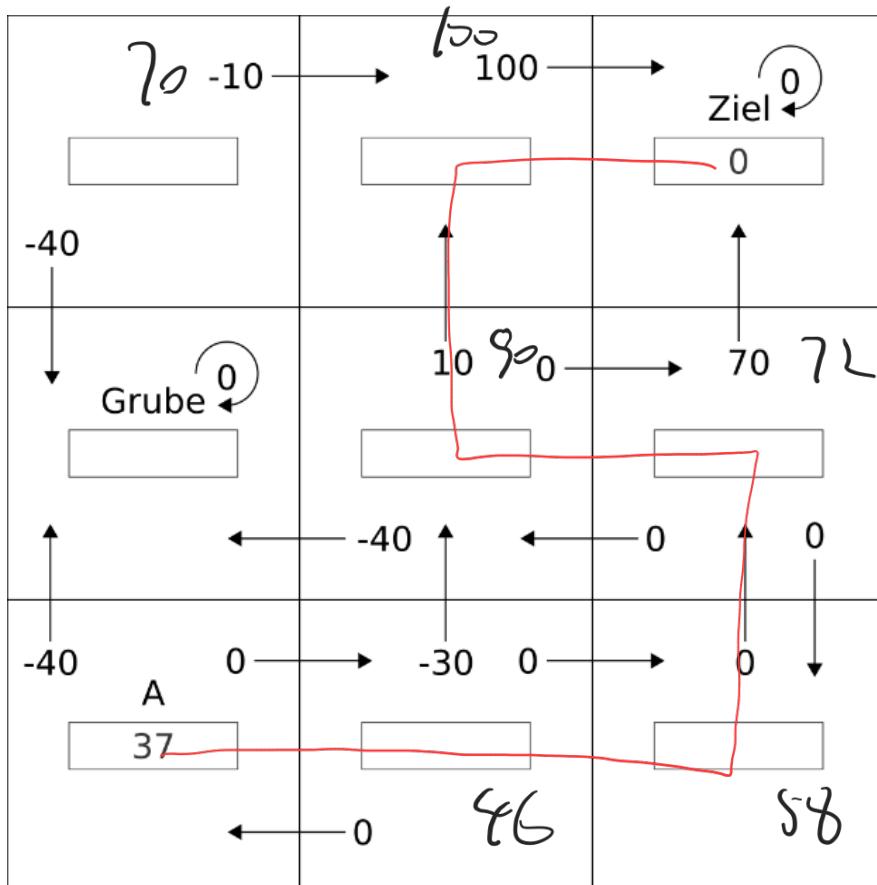
- e) Nennen Sie einen Vorteil und einen Nachteil des strategiebasierten Verfahrens (___ /1P)

flexibility in addressing complex problems
high expertise requirement
not enough exploitation

- f) 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 ($V^*(s)$) dieser Strategie in die entsprechenden Kästen ein (Diskontierungsfaktor $\gamma = 0,8$). Runden Sie ihre Ergebnisse auf ganze Zahlen.

Nehmen Sie an, dass die optimale Strategie gelernt wurde. Tragen Sie die Zustandswerte ($V^*(s)$) dieser Strategie in die entsprechenden Kästen ein (Diskontierungsfaktor $\gamma = 0,8$). Runden Sie ihre Ergebnisse auf ganze Zahlen.

Zeichnen Sie den Pfad der optimalen Strategie von Zelle A zum Ziel ein.



Aufgabe 6 -- HMM, Lernen nach Bayes, SPN / 12P

- a) Nennen Sie die drei, bei der Anwendung von HMMs zu lösenden, grundlegenden Probleme sowie jeweils einen effizienten Lösungsansatz für das jeweilige Problem. (/1,5P)

- b) Gegeben sind zwei unabhängige Ereignisse A und B. (/1P)
 Weiterhin sind $P(B)$ und $P(A|B)$ bekannt.
 Geben Sie $P(B|A)$ sowie $P(A, B)$ im Bezug zu den bekannten Größen an.
 Vereinfachen Sie so weit wie möglich.

$$P(B|\alpha) = \frac{P(\alpha, B)}{P(\alpha)} = \frac{P(\alpha|\beta) \cdot P(\beta)}{P(\alpha)} = P(\beta)$$

$$P(\beta|\alpha) = P(\beta)$$

$$P(\alpha, \beta) = P(\beta) \cdot P(\alpha|\beta)$$

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

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

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

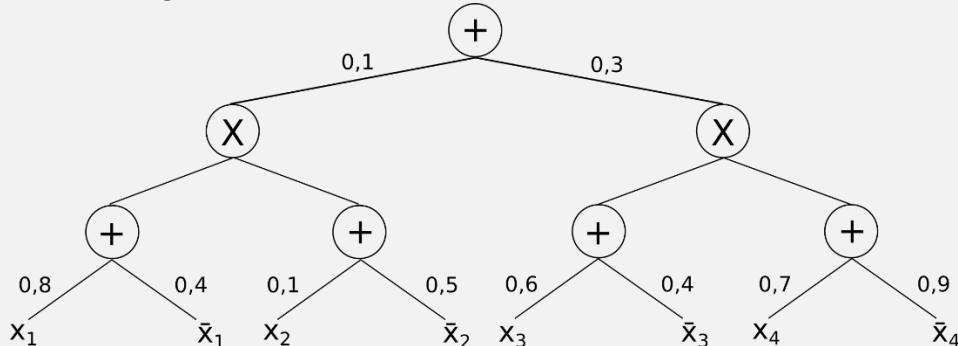
$$p(V=j|) = \frac{1}{2} \quad p(V=N) = \frac{1}{2}$$

$$\begin{aligned} & p(V=j) \cdot p(W=S|V=j) \cdot p(T=N|V=j) \\ &= \frac{1}{2} \times \frac{4}{5} \times \frac{3}{5} \\ &= \frac{6}{20} \end{aligned}$$

$$\begin{aligned} & p(V=N) \cdot p(W=S|V=N) \cdot p(T=N|V=N) \\ &= \frac{1}{2} \times \frac{1}{5} \times \frac{3}{5} \\ &= \frac{3}{20} \end{aligned}$$

Classification: Verspätung = ja

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



- e) Berechnen Sie die Wahrscheinlichkeit $P(\bar{X}_1)$ für das gegebene SPN, welches die Verteilung $P(X)$ repräsentiert. Geben Sie alle Zwischenergebnisse an. Sie sollen hierzu in die Abbildung schreiben. (____/2,5)

