

Aufgabe 1 – Induktives Lernen, Lerntheorie, Entscheidungsbäume und Unüberwachtes Lernen

____ /12P

- a) Welche Ziele verfolgen Induktion und Deduktion? Wie erreichen sie diese Ziele? 01, 16, 17 (____/2P)

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

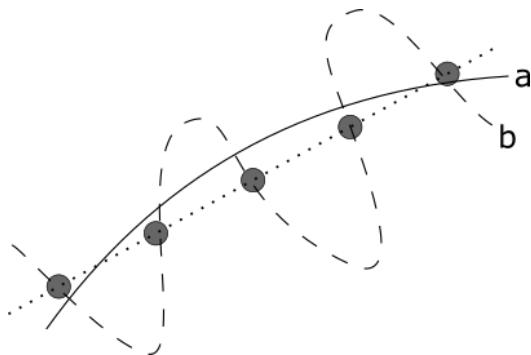
Deductive reasoning is a method of reasoning that starts with a general principle and examines the possibility to reach a specific, logical conclusion.

- b) Definieren Sie Konsistenz und Vollständigkeit im Hypothesenraum. 01, 16, 17 (____/2P)

Consistent: No negative examples are classified positive

Complete: All positive examples are classified positive

- c) Welche Lernmaschine weißt die höhere Kapazität auf? Die Lerndaten werden durch die Kreise dargestellt (Zielfunktion gepunktet). Welches Verhalten bringt eine höhere Kapazität bei Lernmaschinen mit sich? 02, 17 (____/1P)

bOverfitting

- d) Wie verhalten sich Lern- und Testfehler beim „Overfitting“? (___/1P)

Test Error higher than training Error

- e) Geben Sie die Formeln für Genauigkeit (Precision) und True-Positive-Rate (Recall) an. (___/2P)

$$P = \frac{TP}{TP + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

- f) Welche Klassenverteilung wird bei Entscheidungsbäumen angestrebt? Wie wirkt sich dies auf die Entropie aus? Definieren Sie den Informationsgewinn. (___/2P)

Partition the space so that all instances in a region where Entropy will reduce close to zero. the same class

Information gain: expected entropy reduction of s by splitting on attribute A

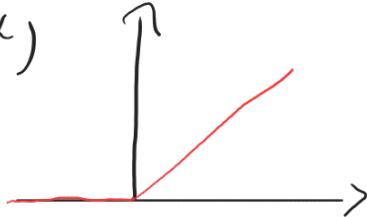
- g) Welche Schritte sind für das „Reduced Error Pruning“ bei Entscheidungsbäumen nötig? (___/2P)

Remove a node (subtree) and replace it with a leaf node of the most common class.

Aufgabe 2 – Neuronale Netze /8P

- a) Nennen Sie eine in der Vorlesung behandelte nichtlineare Aktivierungsfunktion und zeichnen Sie das zugehörige Schaubild der Aktivierungsfunktion. (____/1P)

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



- 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)

$$E(w) = \frac{1}{N} \sum_{i=1}^N (y_i - w \cdot x_i)^2 \quad \Delta w = -\gamma \cdot \frac{\partial E(w)}{\partial w}$$

y_i : *truth* of *input*

f_i : *output*

f_i : *predicted output*

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

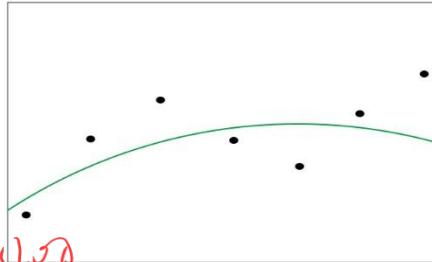
1. Local minima: SGD, Momentum

2. Flat regions or plateaus: Learning Rate Adjustment

Slow convergence because of zig-zags \Rightarrow Adam

- d) Was lässt sich über die VC-Dimension eines neuronalen Netzes sagen, das aus den untenstehenden Lerndaten (Punkte) die eingezeichnete Kurve approximiert? Wie muss die Topologie des Netzes angepasst werden um die Approximation zu verbessern. (___/1P)

02,



VC Dimension

The system with low capacity that leads to underfitting.
Choose an optimal hypothesis

- e) „Je höher die VC-Dimension, umso besser kann das Netz aus einem bestehenden Datensatz lernen, d.h. generalisieren.“ Ist diese Aussage wahr oder falsch, begründen Sie Ihre Entscheidung. (___/1P)

02. P 63

False. Because models with a high VC-Dimension will increase the risk of Overfitting

- f) Was versteht man unter „residual learning“ und wodurch wird damit das Training verbessert? (___/1P)

08 160

~~skip connection~~ ReLU halves the variance in each layer and manually increase the variance by that factor

Aufgabe 3 – Convolutional Neural Network /8P

- a) Für welche Arten von Daten werden CNNs typischerweise verwendet?

(/1P)

Image Data

- b) Warum besitzen CNNs weniger Parameter/Gewichte als vollvernetzte Neuronale Netze?

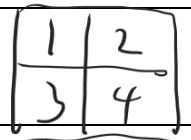
Copy(/2P)*weight sharing*

*Parameter Sharing through Convolution; Translation-Invariance
Pooling layers; Hierarchical structure; Global Receptive field*

- c) Wenden Sie die "max-pooling" Operation mit einer Filtergröße von 2x2, einem Padding von 0 und einem Stride von 1 an. Die Eingabe sieht wie folgt aus:

(/2P)

1	1	2
0	1	0
3	1	4



- d) Wie sehen die Filter von auf realen Bilddaten trainierten CNNs typischerweise aus? Gibt es Unterschiede je nach Position (Tiefe) des Filters im CNN? Beschreiben Sie diese.

(/2P)*6.180**odd number symmetry C. input channel*

*Features from low receptive fields are mostly edge detectors
Features from high receptive fields contain semantic meaning*

- e) Was ist ein Vorteil von Fully Convolutional Networks gegenüber CNNs mit Fully-Connected Schichten am Ende und wofür werden diese typischerweise verwendet?

Fully convolutional networks can be used

for different resolutions
image segmentation

Aufgabe 4 – Support Vector Machines /8P

- a) Bezogen auf einen Support Vektor Klassifikator, beschreiben Sie das Problem welches gelöst wird, die Lösung die gefunden wird und Intuition für die Lösung. 07. 18 (____/1,5P)

Problem: Classification

Solution: Find the best separating line / hyperplane with maximum margin to the classes

Intuition: Size of the margin determines the generalization capability

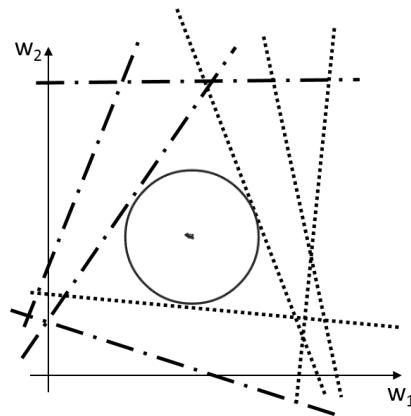
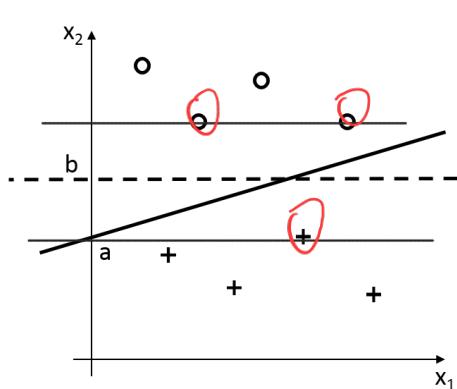
- b) Formulieren Sie mathematisch das grundlegende Optimierungsproblem für einen linearen Support Vector Klassifikator. Geben Sie außerdem die Nebenbedingungen für die Optimierung an. 07. 18 (____/1,5P)

Maximize $\frac{1}{2} \|\mathbf{w}\|^2$ –> minimize $\|\mathbf{w}\|^2$

Under the conditions: $y_i(\mathbf{w}' \mathbf{x}_i + b) \geq 1, i=1, \dots, n$

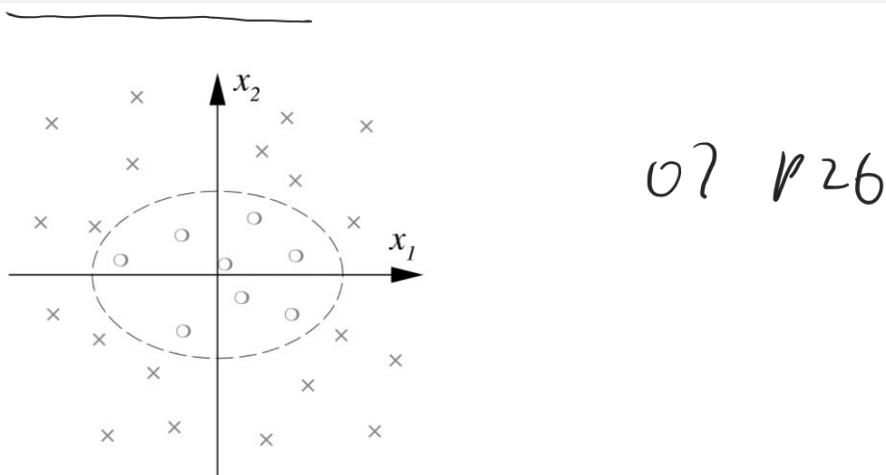
- c) Wie viele Stützvektoren werden für die eindeutige Lösung eines binären Klassifikationsproblems mindestens benötigt, wenn der Merkmalsraum $n > 0$ Dimensionen hat? htl (____/1P)

- d) Welche der beiden Hypothesen im linken Graph ist die optimale Hypothese? (___/3P)
 Markieren Sie die entsprechenden Stützvektoren (Support Vectors) im linken Graphen (Merkmalsraum) und zeichnen Sie die Hypothese (Parameter w) in den rechten Graphen (Hypothesenraum) ein.



b

- e) Geben Sie eine gültige Transformationsregel an um die Daten in einem anderen Raum linear trennen zu können. (___/1P)



Möglichkeit - Transformation

$$\varphi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

$$(x_1, x_2) \rightarrow (z_1, z_2, z_3) = (x_1^1, \sqrt{x_1}x_2, x_1^2)$$

Aufgabe 5 – Reinforcement Learning /12P

- a) 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 reward.

- b) Beschreiben Sie die Begriffe Zustand und Beobachtung und geben sie je ein Beispiel für einen Zustand und eine Beobachtung. (___/2P)

A state is a complete description of the state of the world. Schach

An observation is a partial description of the state of the world. Poker

- c) Beschreiben Sie die Begriffe „Bootstrapping“ und „Sampling“ für die Wertaktualisierungen beim Reinforcement Learning. (___/3P)
Wertaktualisierungen beim Reinforcement Learning.
Ordnen sie die Verfahren: Policy Iteration, SARSA und Monte Carlo Methoden den beiden Begriffen zu.

12. P 37

Bootstrapping: Update involves an estimate.
 dynamic programming; temporal-difference learning

Sampling: Update samples an expectation
 Monte-Carlo Methods,
 temporal-difference learning

- d) Ein Agent wird mit Hilfe von Q-Learning trainiert. Wie lautet die Formel zur Berechnung des TD Fehlers? (___/2P)

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. Bestimmen sie den TD-Fehler.

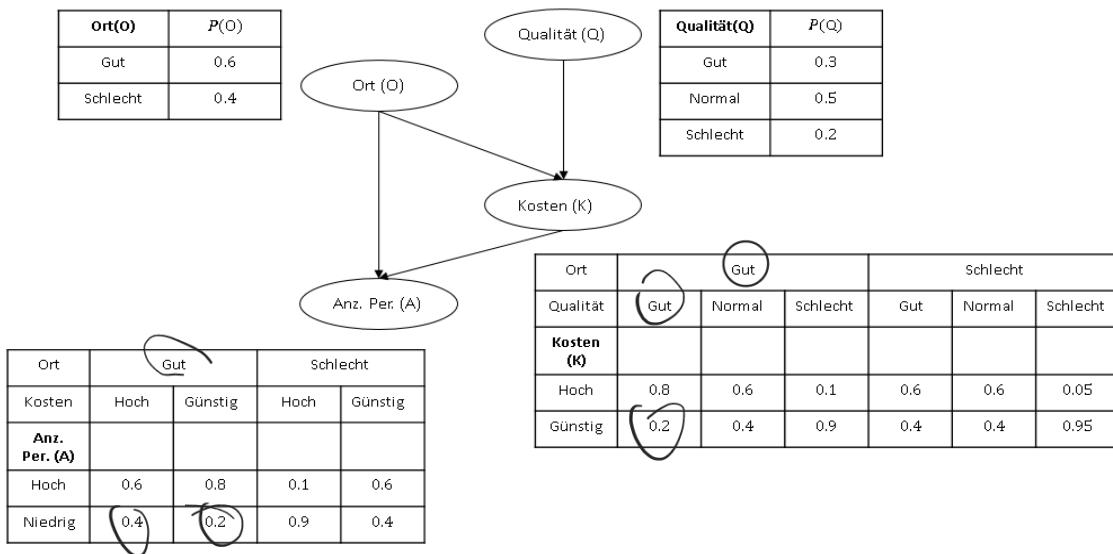
$$\begin{aligned} \delta_t &= r_t + \gamma Q_{\pi^t}(s_t', a_t') - Q_{\pi^t}(s_t, a_t) \\ &= 1 + 0.9 \times 10 - 8 \\ &= 2 \end{aligned}$$

- e) Um die Varianz bei dem REINFORCE Algorithmus zu reduzieren kann eine „Baseline“ verwendet werden. Definieren sie eine geeignete „Baseline“ formal. Ist die Verwendung einer konstanten „Baseline“ zulässig? Begründen Sie Ihre Antwort. (___/2P)

- f) Warum bezeichnet man strategiebasierte Verfahren wie den REINFORCE Algorithmus als „on-policy“ Algorithmen? Warum kann die Strategieaktualisierung beim REINFORCE Algorithmus ohne „Baseline“ zu einer Verschlechterung führen? (___/2P)

Aufgabe 6 – Bayes HMM und SPN**____ /12P**

Um ein gutes Restaurant auszuwählen, können vier Faktoren berücksichtigt werden: die Qualität (Q) und die Kosten (K) des Essens, der Ort des Restaurants (O) und die Anzahl der Personen (A), die das Restaurant besuchen. Das folgende Bayes-Netzwerk beschreibt die Beziehung zwischen den Faktoren



- a) Geben Sie die zugehörige Faktorisierung der Verbundwahrscheinlichkeitsdichte an:

$$P(O, Q, K, A)$$

(____/1P)

$$P(O, Q, K, A) = p(O) \cdot p(Q) \cdot p(K|O, Q) \cdot p(A|O, K)$$

- b) Wie hoch ist die Wahrscheinlichkeit, ein Restaurant zu besuchen, dessen Ort und Qualität des Essens gut sind, dessen Kosten niedrig sind und dessen Besucherzahl gering ist?

(Hinweis: Es ist hinreichend bei der Berechnung die korrekten Multiplikanden aufzuschreiben)

$$\begin{aligned}
 P(O = \text{Gut}, Q = \text{Gut}, K = \text{Günstig}, A = \text{niedrig}) &= \\
 p(O = \text{Gut}) \cdot p(Q = \text{Gut}) \cdot p(K = \text{Günstig} | O = \text{Gut}, Q = \text{Gut}) & \\
 \cdot p(A = \text{niedrig} | O = \text{Gut}, K = \text{Günstig}) & \\
 = 0.6 \times 0.3 \times 0.1 \times 0.4 &= 0.0072
 \end{aligned}$$

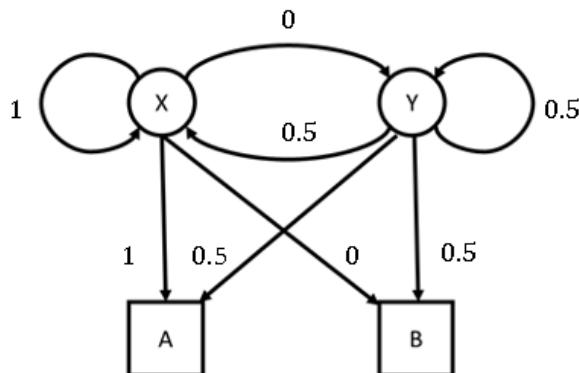
- c) Angenommen, der Preis des Essens ist für die Entscheidung nicht wesentlich, wie hoch ist dann die Wahrscheinlichkeit, ein Restaurant zu besuchen, dessen Ort gut, die Qualität des Essens gut und die Anzahl der Personen gering sind? (Hinweis: Es ist hinreichend bei der Berechnung die korrekten Multiplikanden aufzuschreiben)

(___/3P)

$$P(O = \text{Gut}, Q = \text{Gut}, A = \text{niedrig}) =$$

$$\begin{aligned} & P(O = \text{Gut}) \cdot P(Q = \text{Gut}) \cdot P(A = \text{niedrig}) \\ &= 0.6 \times 0.3 \times \frac{0.4 \times 0.2}{2} \end{aligned}$$

Gegeben sei das folgende Hidden Markov Modell mit der Zustandsmenge $S=\{X, Y\}$ und der Menge der möglichen Beobachtungen $O=\{A, B\}$. Die Startverteilung bei $t=0$ sei $S_0 = (0,1; 0,9)$.



- d) Sagen Sie den (System-)Zustand des Hidden-Markov-Modells zum Zeitpunkt $t = 2$ voraus (Prädiktion). (___/2P)

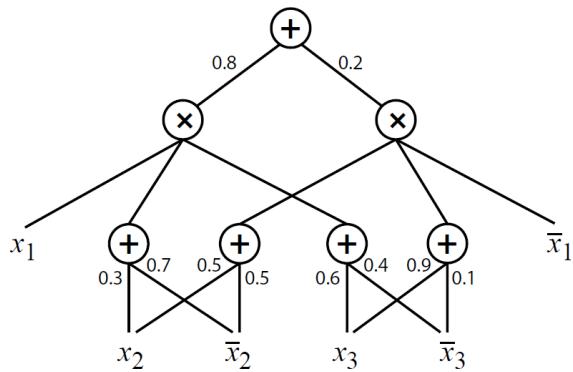
(___/2P)

- e) Sagen Sie den (System-)Zustand des Hidden-Markov-Modells zum Zeitpunkt $t = \infty$ voraus (Prädiktion). (___/1P)

(___/1P)

f) Gegeben ist folgendes Sum-Product Netz (SPN), das eine Wahrscheinlichkeitsverteilung über die Zufallsvariablen X_1, X_2 und X_3 mit Hilfe der Indikatorvariablen $x_1, \bar{x}_1, x_2, \bar{x}_2, x_3, \bar{x}_3$ kodiert. Berechnen Sie die Wahrscheinlichkeiten der Belegung der folgenden Zufallsvariablen und tragen Sie diese in die Tabelle ein.

(___ /3P)



X_1	X_2	X_3	$\Phi(X)$
1	1	1	
0	1	1	
0	0	0	

