

Aufgabe 1**9 Punkte**

- a) Geben Sie die Definition von Güte (Accuracy) und Präzision (Precision) eines binären Klassifikators jeweils in Abhängigkeit von True Positives (TP), True Negatives (TN), False Positives (FP) und False Negatives (FN) an. (___/2P)

$$\text{Accuracy: } \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision: } p = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- b) Betrachtet wird ein binäres Klassifikationsproblem mit möglichen Ausgabewerten $y \in \{0,1\}$ und Eingabewerten $x \in \mathbb{R}$. Der Hypothesenraum sei gegeben durch die Menge $H = H^+ \cup H^-$, wobei $H^\pm = \{h_\theta^\pm | \theta \in \mathbb{R}\}$ mit $h^+_\theta(x) := \begin{cases} 1 & \text{falls } x > \theta \\ 0 & \text{sonst} \end{cases}$ und $h^-_\theta(x) := 1 - h^+_\theta(x)$. Außerdem seien 2 Trainingsbeispiele gegeben:
- Eingabewert $x^{(1)} = 1$ mit Ausgabewert $y^{(1)} = 0$,
 - Eingabewert $x^{(2)} = 3$ mit Ausgabewert $y^{(2)} = 1$.

Geben Sie den Versionsraum $VS_{H,D}$ bezüglich des Hypothesenraums H und den Trainingsbeispielen $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)})\}$ an, indem Sie den entsprechenden Parameterbereich für θ spezifizieren.

Tipp: Vielleicht hilft Ihnen eine Skizze zur Veranschaulichung des Problems.

$$x^{(1)} = 1, \quad \text{if } y^{(1)} = 0 = h^+_\theta(x^{(1)}) \quad x \leq \theta \quad \theta \geq 1$$

$$x^{(2)} = 3, \quad y^{(2)} = 1 = h^+_\theta(x^{(2)}) \quad x \geq \theta \quad 3 \geq \theta$$

$$(\leq \theta \leftarrow)$$

c)

Begründen Sie, warum die **VC-Dimension** eines Hypothesenraums bestehend aus linearen Klassifikatoren in \mathbb{R} (wie zum Beispiel H) gleich 2 ist.

(__/2P)

Tipps:

- Wie ist die VC – Dimension definiert?
- Überlegen Sie ob zwei, drei oder mehrere Eingabewerte so gewählt werden können, dass es für jede Kombination von Ausgabewerten einen θ -Wert gibt, sodass die Eingabewerte korrekt klassifiziert werden.

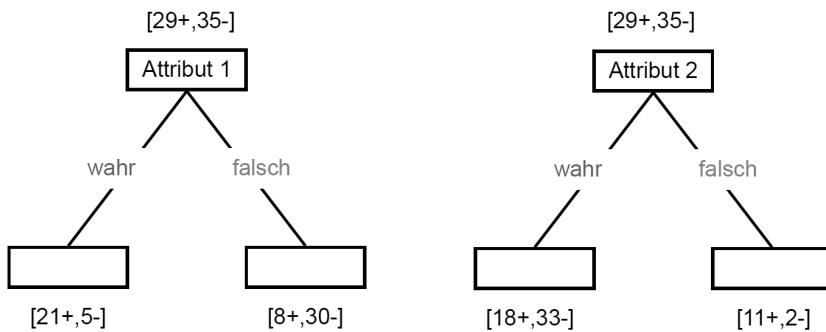
The VC dimension of H is equal to the maximum number of data points which can be arbitrarily separated from H .

- d) Betrachten Sie die nachfolgende Darstellung zum 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 (Notation wie in der Vorlesung)

(___ /3P)

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

besser als Entscheidungskriterium? Begründen Sie Ihre Entscheidung mit zwei Rechnungen (ohne diese auszurechnen) und einer Fallunterscheidung.



Schreibweise: [Anzahl Positive Bsp. (+), Anzahl Negative Bsp. (-)]

$$A1: \quad IG(A_1) = -\frac{29}{64} \log_2 \frac{29}{64} - \frac{35}{64} \log_2 \frac{35}{64} = 0.99$$

$$IG(S_1) = -\frac{21}{26} \log_2 \frac{21}{26} - \frac{5}{26} \log_2 \frac{5}{26} = 0.71$$

$$IG(S_2) = -\frac{8}{38} \log_2 \frac{8}{38} - \frac{30}{38} \log_2 \frac{30}{38} = 0.14$$

$$IG(S, A_1) = 0.99 - \frac{26}{64} \cdot 0.71 - \frac{30}{64} \cdot 0.14 = 0.26$$

$$A2: \quad IG(A_2) = 0.99$$

$$IG(S_3) = -\frac{18}{51} \log_2 \frac{18}{51} - \frac{33}{51} \log_2 \frac{33}{51} = 0.94$$

$$IG(S_4) = -\frac{11}{15} \log_2 \frac{11}{15} - \frac{2}{15} \log_2 \frac{2}{15} = 0.62$$

$$IG(S, A_2) = 0.99 - \frac{51}{64} \cdot 0.94 - \frac{15}{64} \cdot 0.62 = 0.12$$

$A_{Attribut 1}$ ist besser

Aufgabe 2**11 Punkte**

- a) Es soll ein Naiver Bayes-Klassifikator zur Detektion von Spam in E-Mails entwickelt werden. Als Attribute/Features stehen folgende Beobachtungen zur Verfügung, die typischerweise aus empfangenen E-Mails extrahiert werden können:

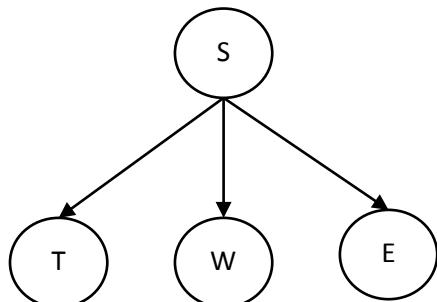
(___/2P)

- die Tageszeit, zu der die E-Mail empfangen wurde ($T \in \{\text{Morgen, Mittag, Abend}\}$),
 - ob sie das Wort "kostenlos" enthält ($W \in \{\text{ja, nein}\}$),
 - und ob die E-Mail-Adresse des Absenders in Ihrem Adressbuch bekannt ist, vorher in Ihrem Posteingang gesehen oder vorher nicht gesehen wurde.
- ($E \in \{B: \text{bekannt}, S: \text{gesehen}, U: \text{nicht gesehen}\}$).

1. Gegeben ist das Bayes'sche Netz. Füllen Sie die folgenden Tabellen mit den entsprechenden (bedingten) Wahrscheinlichkeiten aus.

T	W	E	S
Morgen	ja	nicht gesehen	Spam (ja)
Mittag	nein	gesehen	Ham(Nein) ✓
Abend	ja	gesehen	Spam (ja)
Abend	ja	bekannt	Ham(Nein) ✓
Mittag	nein	nicht gesehen	Ham(Nein) ✓
Abend	nein	gesehen	Ham(Nein)
Mittag	nein	nicht gesehen	Spam (ja)
Morgen	nein	gesehen	Ham(Nein) ✓
Morgen	Ja	nicht gesehen	Spam (ja)

S	P(S)
Spam (ja)	$\frac{4}{9}$
Ham (nein)	$\frac{5}{9}$



S	E	P(E S)
Spam	Bekannt	0
Spam	Gesehen	$\frac{1}{4}$
Spam	nicht gesehen	$\frac{3}{4}$
Ham	Bekannt	$\frac{1}{5}$
Ham	Gesehen	$\frac{3}{5}$
Ham	nicht gesehen	$\frac{1}{5}$

w=ja ↳ bekannt

2. Eine neue E-Mail wurde empfangen. Diese E-Mail enthält das Wort „kostenlos“, aber die E-Mail-Adresse des Absenders existiert in Ihrem Adressbuch. Aus technischen Gründen fehlt die Angabe der Tageszeit. (____/2P)

Was sollte die Vorhersage Ihres Naive Bayes Spam-Detektors, basierend auf diesen Attributen sein?

Begründen Sie Ihre Entscheidung formal.

$$p(\text{Spam}=\text{ja}) = \frac{4}{9} \quad p(\text{Ham}=\text{nein}) = \frac{5}{9}$$

$$p(\text{Spam}=\text{ja}) \cdot p(w=\text{ja} | \text{Spam}=\text{ja}) \cdot p(\bar{b}=\text{bekannt} | \text{Spam}=\text{ja})$$

$$= 0$$

$$p(\text{Ham}=\text{nein}) \cdot p(w=\text{ja} | \text{Ham}=\text{nein}) \cdot p(\bar{b}=\text{b} | \text{Ham}=\text{nein}) = \frac{5}{9} \times \frac{1}{5} \times \frac{1}{5} = \frac{1}{45}$$

Rechtf.: Ham (Nein)

- b) Was kann bei einem Bayes'schen Netz gelernt werden? (____/1P)

Mit welcher Methode erfolgt dies, wenn die Struktur bekannt ist und Variablen nur teilweise beobachtbar sind?

p. P42 P46.

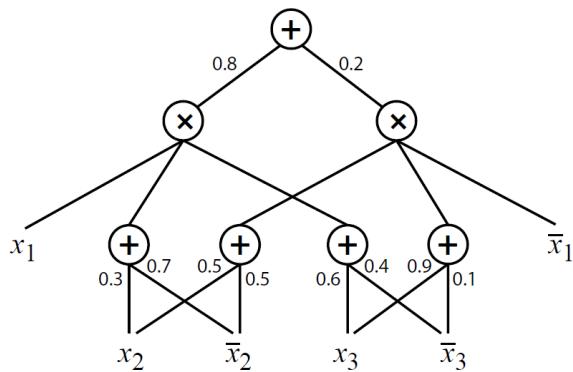
describes conditional dependencies with respect to subsets of variables,

Gradient ascent, expectation-maximization algorithm



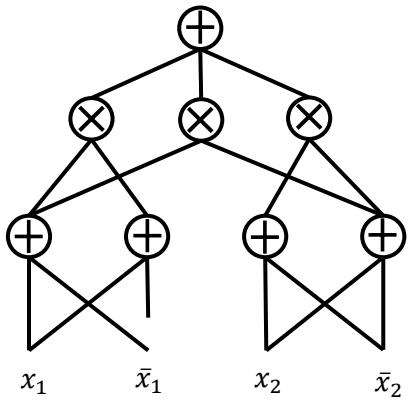
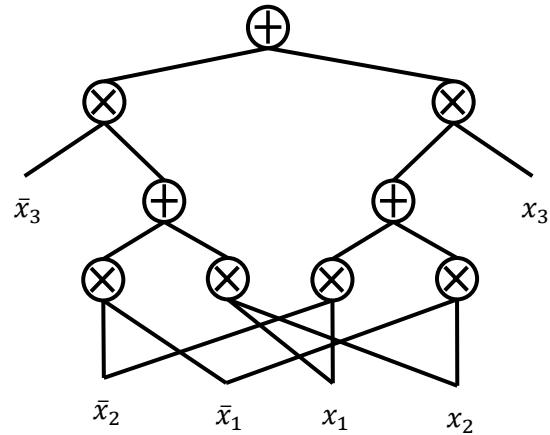
Gegeben ist folgendes Sum-Product Netz (SPN), das eine Wahrscheinlichkeitsverteilung über die Zufallsvariablen X_1, X_2 und X_3 mit Hilfe der Indikatorvariablen $x_1, \bar{x}_1, x_2, \bar{x}_2, x_3, \bar{x}_3$ kodiert. Berechnen Sie die Wahrscheinlichkeiten der Belegung der folgenden Zufallsvariablen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

(____/3P)



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

d) Gegeben sind die folgenden beiden Sum-Product Netze. Geben Sie jeweils an, ob das SPN valide ist oder nicht valide ist. Begründen Sie Ihre Wahl, kurz. (___/3P)

A**B**

Aufgabe 3**10 Punkte**

- a) Geben Sie eine quadratische Fehlerfunktion E eines neuronalen Netzes, sowie /2P die Formel der iterativen Gewichtsoptimierung $\Delta \vec{w}$ in Abhängigkeit von E an.
Benennen Sie die verwendeten Variablen.
(Gefragt ist **nicht** die explizite Rückpropagierung der Gradienten)

- b) Geben Sie die Funktionsterme von zwei häufig verwendeten nichtlinearen Aktivierungsfunktionen an. Für welches Verfahren im Kontext Neuronaler Netze wird die Ableitung der Aktivierungsfunktion benötigt? /2P

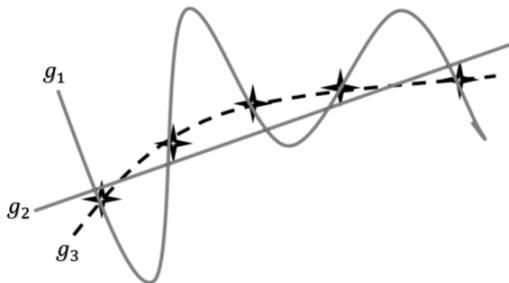
ReLU: $f(x) = \max(0, x)$ Leaky ReLU $f(x) = \begin{cases} x, & x > 0 \\ ax, & \text{else} \end{cases}$
Backpropagation

- c) Welcher Bagging-Ansatz für (tiefe) neuronale Netze wurde in der Vorlesung vorgestellt? Wie äußert sich die Verwendung dieser Methode jeweils in Training und Inferenz des neuronalen Netzes? /2P

δ, P_{dj}

Dropout
during training randomly deactivate neurons with probability p
during inference, multiply the output of each neuron with its dropout probability p

- d) 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). (___ /2P)
1. Wie nennt man das Phänomen, das während des Lernens auftritt und zu g_1 führt? Nennen Sie zwei Ansätze, die dieses Phänomen bei neuronalen Netzen verhindern können.



\exists : over fitting
 early stopping increase number and types
 of instances

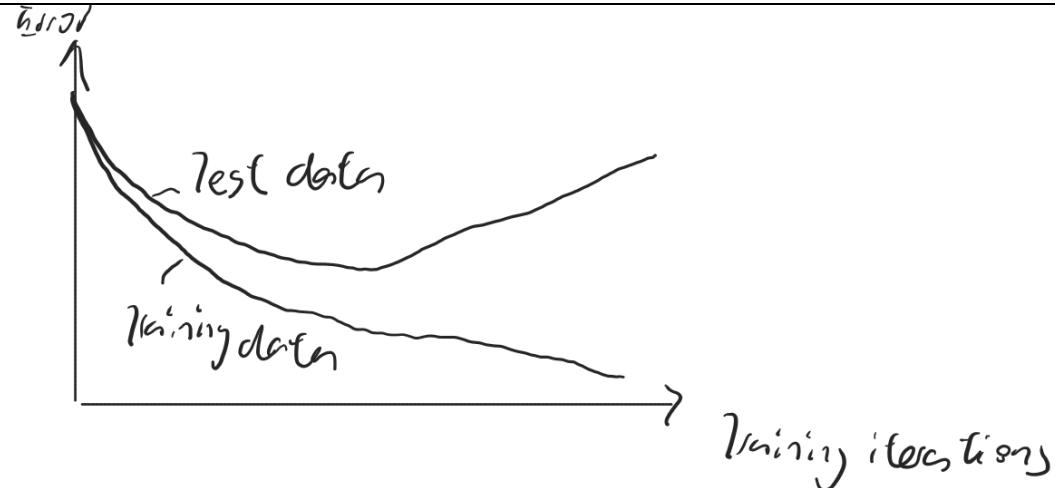
2. Welche Vermutung kann über die VC-Dimension des Netzes, das die Kurve g_2 als Approximation für die Trainingsdaten liefert, getroffen werden? (___ /1P)

O L. P63

The larger $VC(h)$, the better a system can learn to solve a problem

3. Betrachten Sie die Trainingsphase des Netzwerks im Fall von Kurve g_1 :
Skizzieren Sie qualitativ den Verlauf der Fehlerfunktion, jeweils für Trainings- und Testdaten in Abhängigkeit der Anzahl der Iterationen eines Gradientenabstiegverfahrens (mit sinnvoll gewählter Lernrate). (____/1P)

Tipp: Es geht hierbei um den groben Verlauf: Eventuelle kleine Fluktuationen des Verlaufs sollen nicht berücksichtigt werden.



Aufgabe 4**10 Punkte**

- a) Gegeben sind die Ein- und Ausgabe eines Conv-Layers in Form von Feature Maps. Es wurde kein Padding durchgeführt und der Stride beträgt 2. (/2P)
Eingabedimension: 11x11x3
Ausgabedimension: 5x5x5

Wie viele Kernel welcher Größe beinhaltet der Conv Layer?

$$\frac{11-n}{2} + 1 = 5$$

$$n = 3$$

$$3, 3 \times 3$$

- b) Nennen Sie drei in der Vorlesung genannte Möglichkeiten um Gewichte eines CNN zu initialisieren. (/1,5P)

06. 141

Xavier Kaiming He
 transfer learning

- c) Was versteht man unter Padding bei CNN? Nennen Sie drei in der Vorlesung genannten Möglichkeiten um die Eingabe eines CNN zu erweitern (Padding). (/1,5P)

06 P15

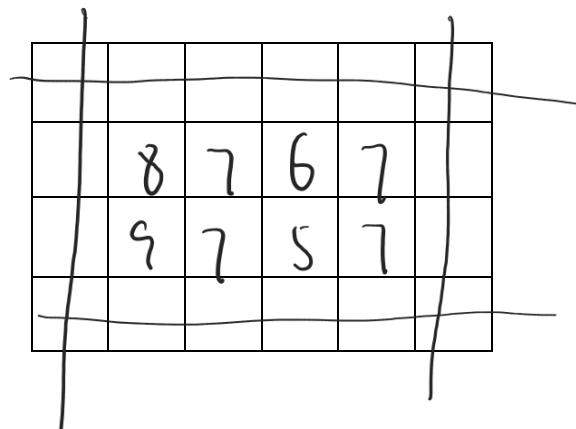
Increase resolution of input feature map

Zero Padding Reflect Padding Circular Padding

- d) Führen Sie ein Max Pooling durch und streichen Sie überflüssige Zeilen und Spalten in der Ergebnisvorlage.
 Filtergröße p=3, Stride s=1. Padding wird in diesem Netz nicht genutzt.

(___/3P)

0	4	6	6	3	0
8	7	5	4	3	2
7	0	3	5	0	7
9	5	3	4	2	2



- e) Bewerten Sie die folgenden Aussagen mit wahr oder falsch?

(___/2P)

In einem CNN werden Gewichte lokal wiederverwendet.	✓
Die letzten Schichten eines CNN sind <u>nicht immer</u> Fully Connected Layer.	✓
Die zu lernenden Gewichte befinden sich sowohl in den Kernen der Convolution- als auch <u>Pooling Layer</u> .	✗
In einem CNN wird für die Erstellung der Features Expertenwissen benötigt.	✗

Aufgabe 5**12 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 Action Space
 Reward function Transition function

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

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

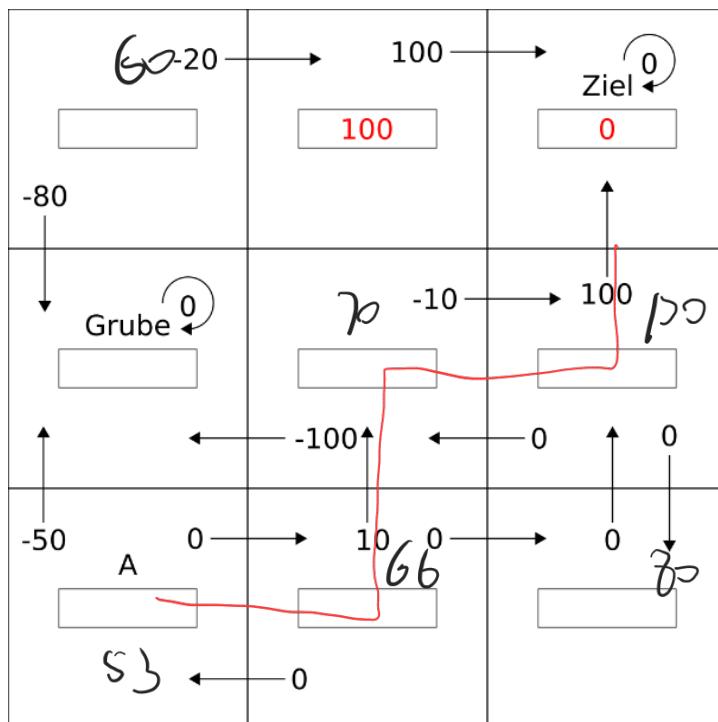
- c) Gegeben ist $V^\pi(s)$: Bestimmen Sie die Optimale Strategie $\pi^*(s)$ formal. (___/2P)
 Außerdem, definieren Sie die rekursive Form der Bellmann Optimalitätsgleichung für den nicht-deterministischen Fall in Abhängigkeit der optimalen Wertfunktion $V^*(s_t)$.

$$\pi^*(s) = \arg\max_a V^\pi(s)$$

$$V^*(s_t) = \max_{a \in A} (r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V^*(s'))$$

- d) Beschreiben Sie die Begriffe des Reinforcement Learning und der Planung. Wo liegen die Unterschiede? (___/2P)

- e) Betrachten Sie die unten stehende 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. Zeichnen Sie den Pfad der optimalen Strategie von Zelle A zum Ziel ein. (___/4P)



Aufgabe 6**8 Punkte**

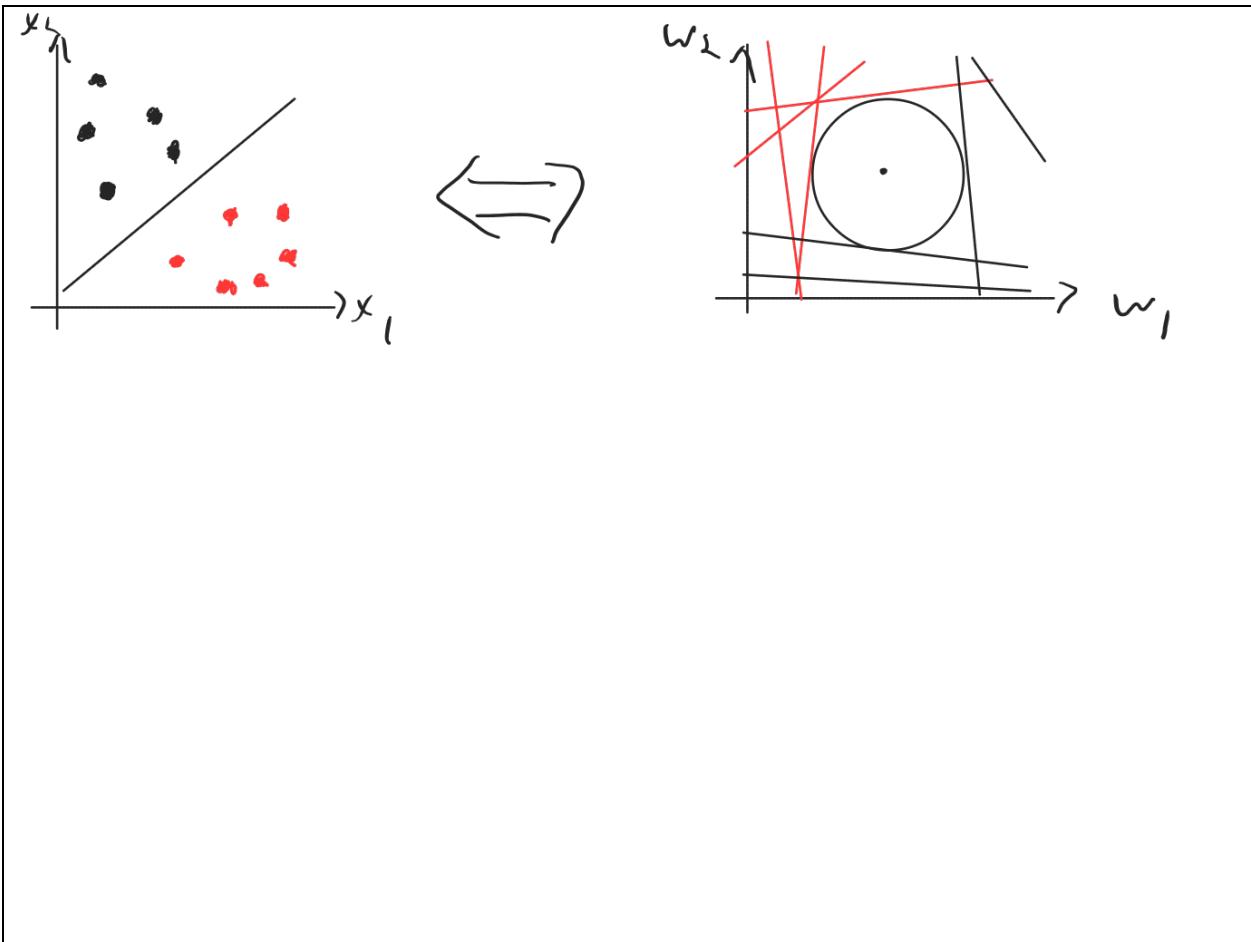
- a) Welches Ziel verfolgt eine Support Vector Machine bei der Klassifikation? (___/1P)

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

- b) Geben Sie das Optimierungskriterium der optimalen Hyperebene sowie die Randbedingung für die korrekte Klassifikation an (gegeben Trainingsbeispiele der Form (\vec{x}_i, y_i)). (___/2P)

$$\begin{aligned} \text{Maximize } & \bar{w}^2 \rightarrow \text{Minimize } (\bar{w}')^2 \\ \text{Subject to } & y_i (\bar{w}' \vec{x}_i + b) \geq 1, \quad i=1 \dots n \end{aligned}$$

- c) Skizzieren Sie die Dualität von Merkmals- und Hypothesenraum der SVM (Version Space Duality). Wie ist die optimale Hyperebene mit größtem Rand im Hypothesenraum repräsentiert? 07. P31 3P



- d) Nennen Sie eine Möglichkeit um die SVM auch für die Lösung nichtlinearer Probleme zu verwenden? Warum ist diese Möglichkeit anwendbar? 07. P16 2P

Non Linear kernel methods.

Transform the data into another (higher-dimensional)
space where the data can be separated
linearly.

