Statistische Lerntheorie

Friedhelm Schwenker

Institut für Neuroinformatik

Email: friedhelm.schwenker@uni-ulm.de

- Vorlesung (2h): Di 14:15-15:45 Uhr im Raum O27/123
- Übungen (2h): Do 12:30-14:00 im Raum O27/123 (1.Übung: 3.5.)
  Schein: 50% der Punkte (ca. 8 Blätter) + aktive Übungsteilnahme; Bonusregel gilt!
- Kernfächer: Mathematische und theoretische Methoden der Informatik.
- Vertiefungsfach: Neuroinformatik.

Überblick der Vorlesung

1. Einführung in das maschinelle Lernen
2. Probabilistisches Lernen — Das PAC-Modell; Einschub W-Rechnung
3. Einführung in die Vapnik-Chervonenkis-Theorie (VC-Theorie)
4. PAC-Lernbarbeit und VC-Dimension von Funktionenmengen
5. Spezielle maschinelle Lernverfahren : Ensemble-Methoden, Support-Vektor-Lernen
6. Zusammenfassung

1. Einführung in das maschinelle Lernen

1. Allgemeine Bemerkungen: Lernen, Lerntheorie, Lernende Maschinen
2. Überwachtes und unüberwachtes Lernen; Reinforcement-Lernen
3. Klassifikationsprozess
4. Perzeptron
5. Historisches
6. Literatur zur statistischen Lerntheorie

Lernen

- Das **Lernen** bezeichnet den Vorgang der Aufnahme und der Speicherung von Erfahrungen.

- Ergebnis des Lernprozesses ist die Veränderung der Wahrscheinlichkeit, mit der Verhaltensweisen in bestimmten Situationen auftreten.

- Der Lernerfolg hängt ab: 1.) von der Zahl und der Art der Bekräftigungen und 2.) von der Zahl der Wiederholungen.

Lerntheorie

- **Mathematische Lerntheorien** sind aus Theorie der Wahrscheinlichkeitsprozesse abgeleitet worden und erlauben Voraussagen über das durchschnittliche Verhalten.

- Lerntheorien geben auch Hinweise auf die Möglichkeit der Simulation von Lernprozessen (lernende Automaten, lernende Maschinen).

Lernende Automaten/Lernende Maschinen/Lerner

- Lernende Automaten/Maschinen sind technische Systeme zur Informationsverarbeitung, denen Lernfähigkeit zugeschrieben wird.
- Die Arbeitsweise ist von den gespeicherten Arbeitsergebnissen (Erfahrungen) abhängig.
- Ziel ist es, den beabsichtigten Arbeitsprozess des Lerners zu optimieren.
- Lernende Automaten zeichnen sich durch die Möglichkeit der Belehrung aus.
- Lernen kann beispielsweise dadurch geschehen, dass die Koeffizienten einer Bewertungsfunktionen entsprechend dem Arbeitsergebnis des Lerners angepasst werden.

Zusammenfassung

1. Lernen erfolgt durch die wiederholte Präsentation von Beispielen.

2. Dabei können die Sollausgaben vorgeben sein (überwachtes Lernen).

3. Die Lernleistung wird durch eine Bewertungsfunktion gemessen.

4. Die Lernleistung kann durch Adaption (Optimierung) der Parameter einer vordefinierten Bewertungsfunktion realisiert werden.

Einordnung Lernverfahren

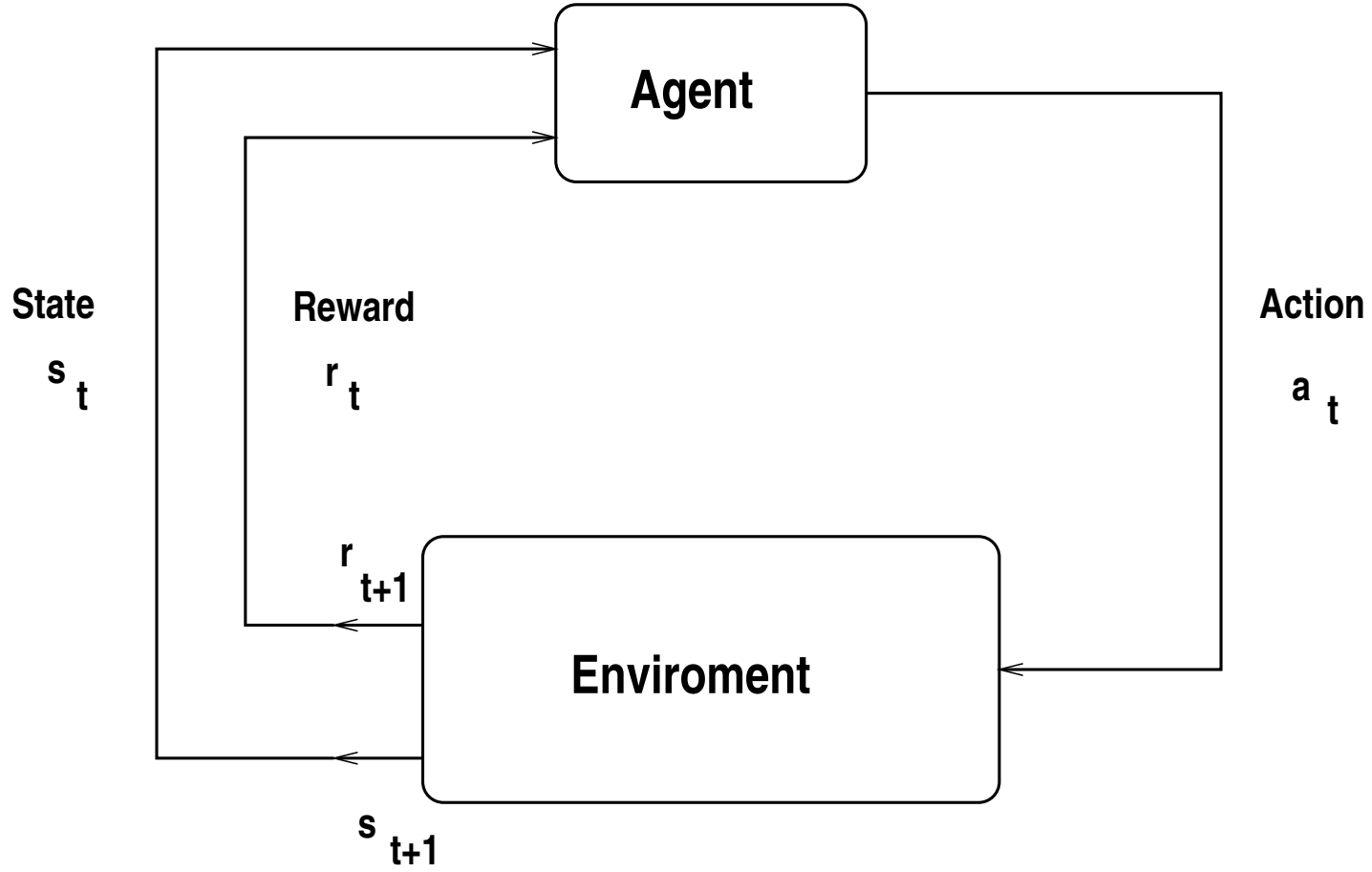
1. **Überwachtes Lernen**
   Gegeben eine (endliche) Stichprobe von Eingabe-Ausgabe-Paaren (x, y) (Trainingsmenge) mit dem Ziel eine Funktion f zu lernen, die für jede Eingabe x einen Funktionswert $f(x)$ bestimmt, mit dem Ziel, dass $f(x)$ möglichst dem zugehörigen y (dem Lehrersignal) entspricht.

   Dabei ist $x \in X$ und X ein beliebiger Eingaberaum (diskret oder kontinulierlich) und der Bildbereich von $Y := f(X)$ kann kontinuierlich sein (dann spricht man von Funktionsapproximation/Regression) oder endlich (diskret) (dann spricht man von Klassifikation).

2. **Unüberwachtes Lernen**
   Es werden keine externen Lehrersignale y für die Beispiele $x \in X$ benutzt. Anwendungsbeispiele: Vektorquantisierung, Merkmalsextraktion

3. **Reinforcement Lernen**
   Kein explizites Lehrersignal; Lerner lernt durch Reward.

Reinforcement Lernen - das Bild



RL - ein paar Begrifflichkeiten

- Agent führt eine Aktion a_t aus.
- Umwelt ändert hierdurch ihren Zustand s_t und erteilt dem Agenten einen Reward $r_t \in \mathbb{R}$, s_t und r_t werden vom Agenten wahrgenommen.
- Agent führt nächste Aktion a_{t+1} aus.
- $\mathcal{S}$ die Menge der Zustände (diskret/endlich)
- $\mathcal{A}$ die Menge der Aktionen (diskret/endlich)
- $\mathcal{A}(s_t)$ Menge der Aktion die im Zustand s_t möglich sind.
- Zeit ist diskret, d.h. $t = 1, 2, 3, \ldots$.

- Der Agent führt die Aktion gemäß einer Strategie/Taktik/Vorgehensweise (*policy*) aus, bezeichnet mit π_t.

- $\pi_t(s, a)$ ist hier die Wahrscheinlichkeit, dass die Aktion $a_t = a$ ausgeführt wird, falls der Zustand $s_t = s$ war.

- Reinforcement Lernverfahren adaptieren direkt oder indirekt die policy π_t des Agenten.

- Agent soll die in der Zukunft zu erwartenden Rewards maximieren, also den mittleren Reward

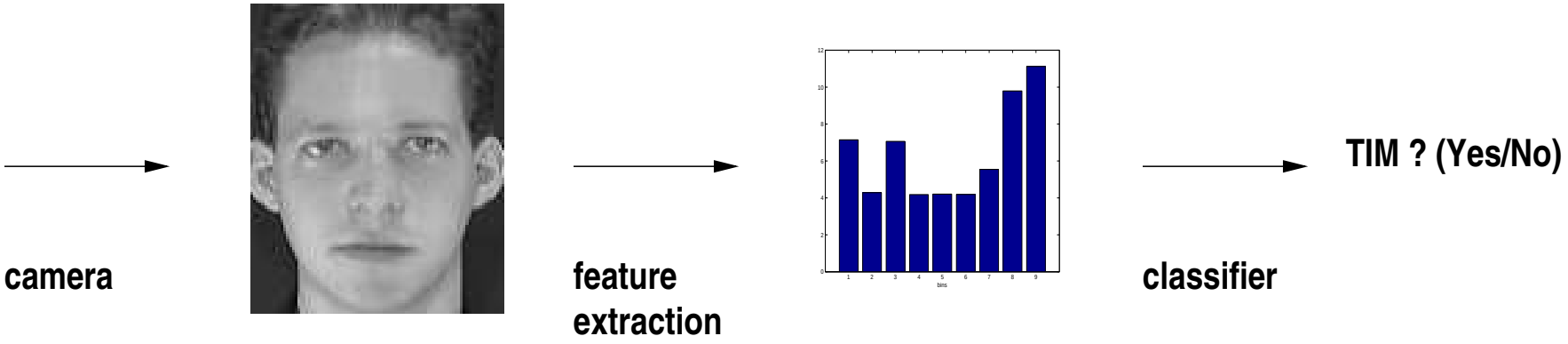
$$\frac{1}{T} \sum_{i=t+1}^{T} r_i$$

  maximieren.

- Problem: $T = \infty$ ist möglich

Klassifikation

- Klassifikation: Ausgabemenge ist diskret $Y = \{0, 1, \ldots, L\}$.
- Beschränken uns im folgenden auf binäre (oder 2-Klassen) Klassifikations-probleme, also $Y = \{0, 1\}$ oder auch $Y = \{-1, 1\}$
- Beispiel: Verifikation einer Person durch Gesichtserkennung.

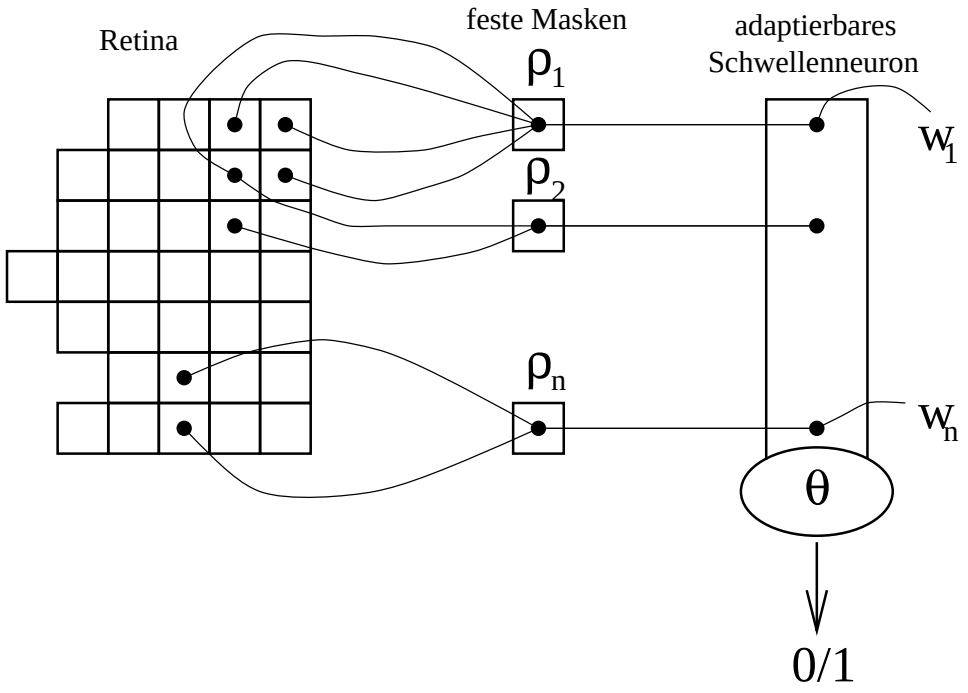


Lernende Maschinen

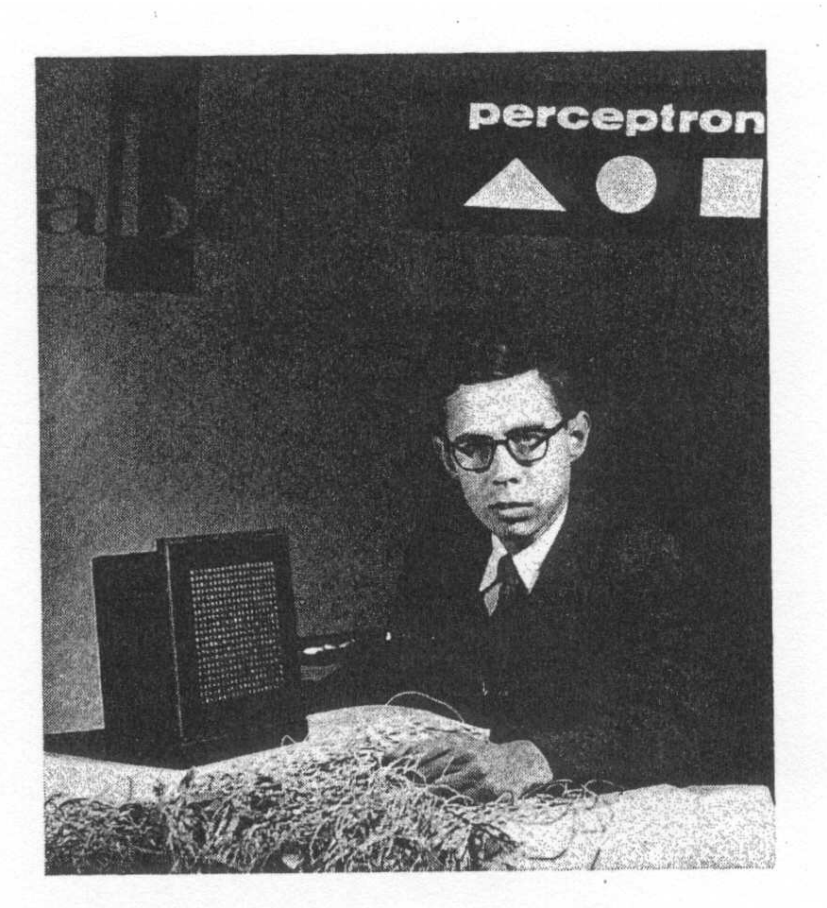
- **Lernmatrix** (neuronaler assoziativer Speicher) von Karl Steinbuch (1961).
- **Perzeptron** von Frank Rosenblatt (1958).
- **Nächste-Nachbar-Klassifikatoren**
- **Lineare Diskriminanzanalyse**
- **Entscheidungsbäume**
- **Multi-Layer-Perzeptrone, Radiale Basisfunktionennetze**
- **Support-Vektor-Maschinen**
- **Ensemble-Methoden** Bagging, Boosting

Architektur des Perzeptron

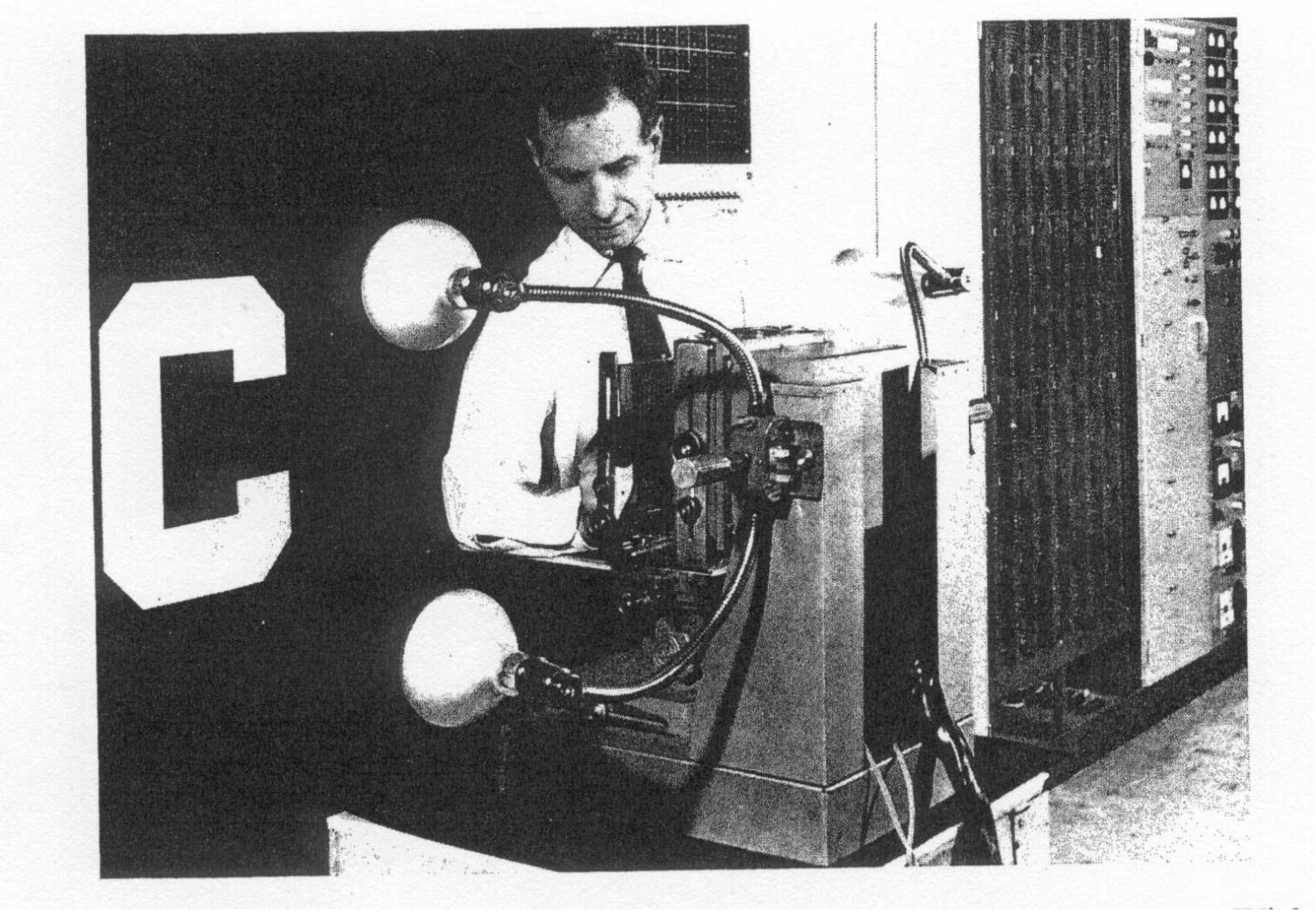
- Eingabe- oder Sensorschicht (häufig auch *Retina* genannt)
- Masken mit festen Kopplungen zur Sensorschicht
- Schwellenneuron mit adaptierbaren Gewichten und Schwellwert



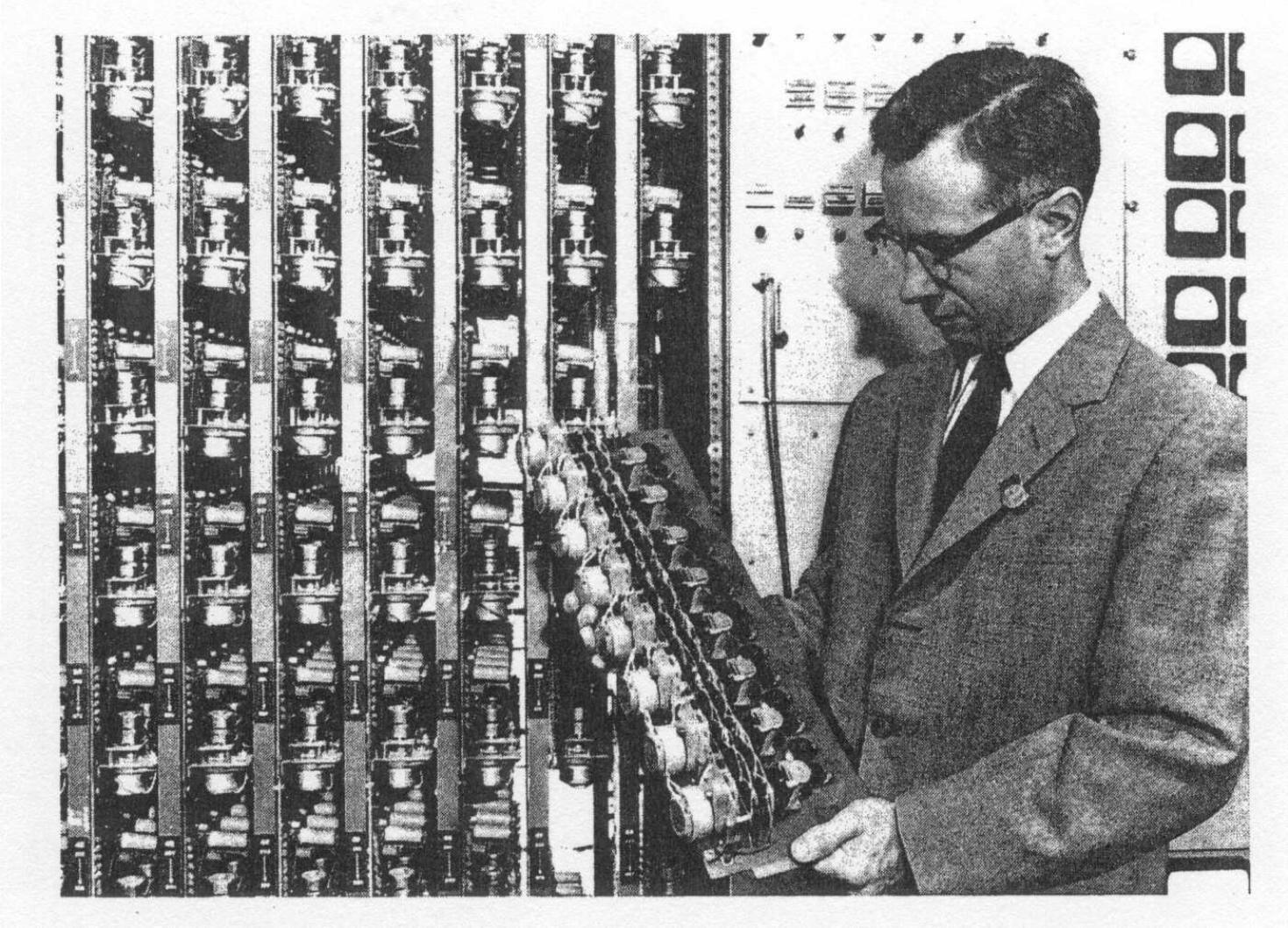
Frank Rosenblatt



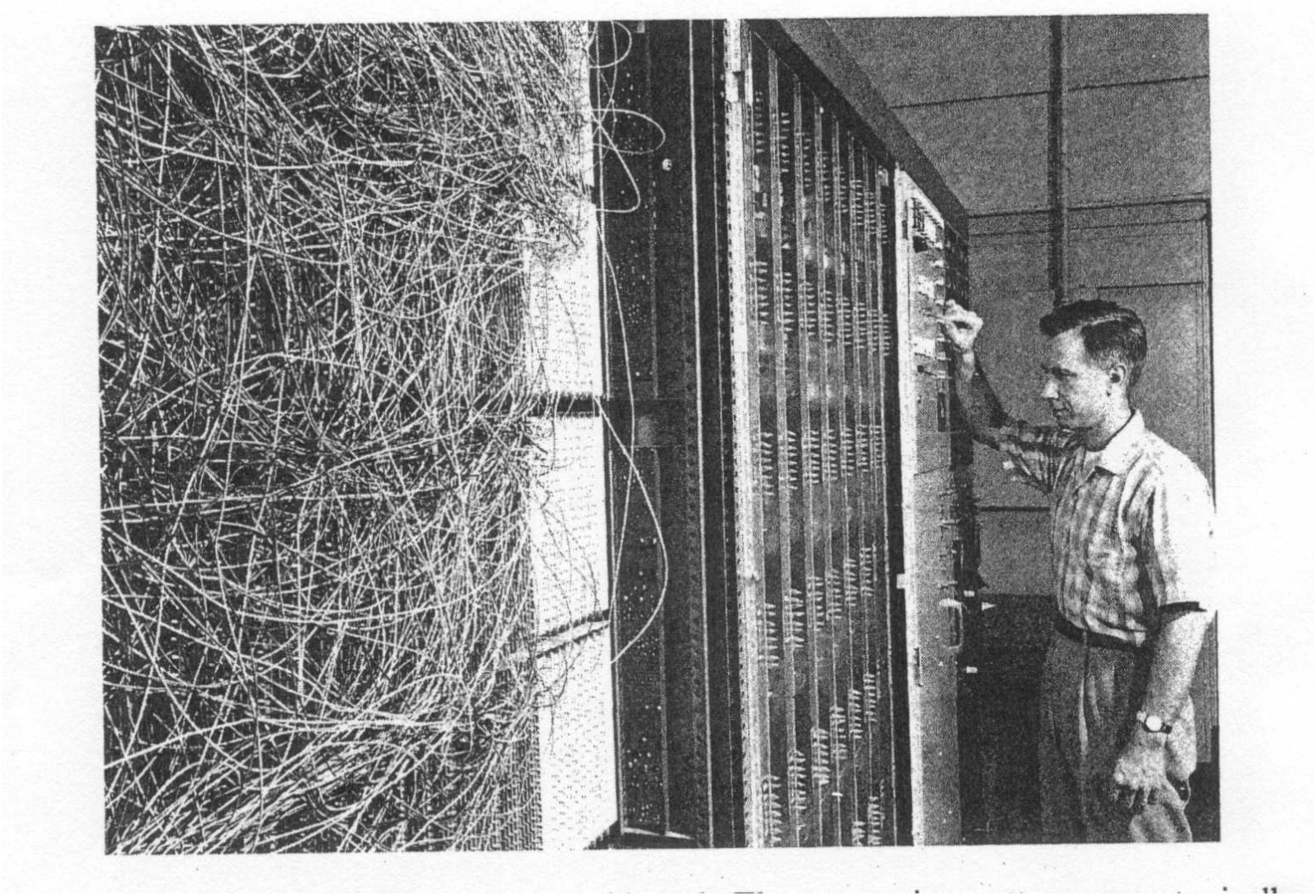
Perceptron - Mark I



Perceptron - Adaptable Weights



Perceptron - Random Connections



Literatur

[1] M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge University Press, Cambridge, 1992.

[2] L. Devroye, L. Györfi and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[3] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan College Publishing Company, 1994.

[4] R. Herbrich. *Lerning Kernel Classifiers: Theory and Applications*. The MIT Press, 2002.

[5] J A. Hertz, A Krogh, and R G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, 1991.

[6] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

[7] B. K. Natarajan. *Machine Learning*. Morgan Kaufmann Publishers, San Mateo, 1991.

[8] R. E. Schapire. *The Design and Analysis of Efficient Learning Algorithms*. The MIT Press, 1992.

[9] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2002.

[10] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[11] V. N. Vapnik. *Statistical Learning Theory*. Addison-Wesley, 1998.

2. PAC Lernen

1. Notation, Klassifikation, Konzeptmengen
2. Beispiele für Konzeptmengen
3. Lernalgorithmen
4. PAC-Lernbarkeit

Notationen

- $\Sigma \neq \emptyset$ bezeichnet das **Alphabet**
- Binäre Alphabete: $\Sigma = \{0, 1\}$ oder $\Sigma = \{-1, 1\}$.
- Abzählbare (un)endliche Alphabete: $\Sigma = \{0, 1, 2, \ldots, p\}$ bzw. $\Sigma = \mathbb{Q}$.
- Überabzählbares Alphabet: $\Sigma = \mathbb{R}$.
- Für $n \in \mathbb{N}$ ist $X = \Sigma^n$ der Produktraum

$$\Sigma^n = \{x = (x_1, \ldots, x_n) \mid x_i \in \Sigma\}.$$

- $X = \Sigma^+$ die Menge der Folgen beliebiger Länge ≥ 1 über Σ :

$$\Sigma^+ = \{x = (x_1, \ldots, x_n) \mid x_i \in \Sigma, n \in \mathbb{N}\} = \bigcup_{n=1}^{\infty} \Sigma^n$$

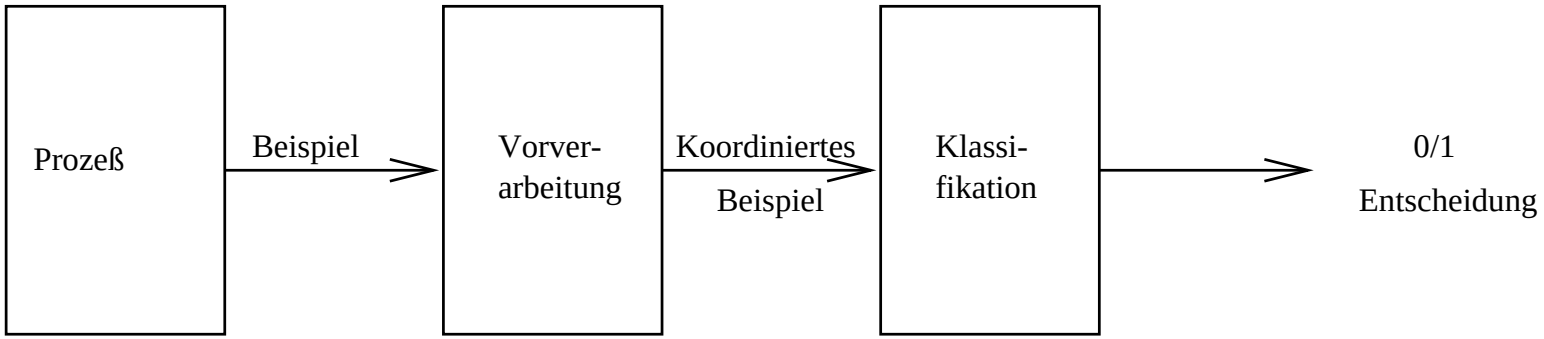
- Mit $\Sigma^0 = \{\varepsilon\}$, ε das leere Wort auch Σ^*

$$\Sigma^* = \{x = (x_1, \ldots, x_n) \mid x_i \in \Sigma, n \in \mathbb{N}_0\} = \bigcup_{n=0}^{\infty} \Sigma^n$$

Im folgenden ist entweder $\Sigma = \{1, -1\}$ oder $\Sigma = \{1, 0\}$oder $\Sigma = \mathbb{R}$.

Der Eingaberaum ist dann entweder $X = \Sigma^n$ oder eine Teilmenge aus Σ^n, also $X \subset \Sigma^n$.

Klassifikation



- Wir betrachten nur den Klassifikationsmodul.
- Klassifikationsabbildung ist von der Form $c : X \to Y$
- Beschränkung auf 2-Klassen-Probleme, d.h. Entscheidungen in oBdA. $Y = \{0, 1\}$ oder $Y = \{-1, 1\}$
- Eingaben sind reelle oder binäre Vektoren, d.h. $X = \{0, 1\}^n$ oder $X = \mathbb{R}^n$.

Konzepte

Wir betrachten also Abbildungen mit binären Bildbereich (2-Klassen-Problem), oBdA. $\{0, 1\}$

Abbildung $c : X \to \{0, 1\}$ heißt auch ein **Konzept** auf der Menge X.

Die **positiven Beispiele** eines Konzeptes c:

$$P(c) = P_c = \{x \in X \mid c(x) = 1\}.$$

Die **negativen Beispiele** (genauer **nicht-positiven Beispiele**) eines Konzeptes c:

$$N(c) = N_c = \{x \in X \mid c(x) = 0\}.$$

Für ein Menge X ist

$$\mathcal{F}_X := \{c \mid c : X \to \{0, 1\}\}$$

die Menge aller möglichen Konzepte auf X.

Einfache Konzepte

1. $\Sigma = \{0,1\}$, $X = \Sigma^2$ und $c : X \to \{0,1\}$ definiert durch

$$c(x) = c((x_1, x_2)) = x_1 \cdot x_2.$$

c ist die sogenannte AND-Funktion mit

$$P(c) = \{(1,1)\} \quad \text{und} \quad N(c) = \{(0,0),(1,0),(0,1)\}$$

2. $\Sigma = \{0,1\}$, $X = \Sigma^2$ und $c : X \to \{0,1\}$ definiert durch

$$c(x) = c((x_1, x_2)) = (x_1 + x_2) \mod 2.$$

c ist die sogenannte XOR-Funktion mit

$$P(c) = \{(1,0),(0,1)\} \quad \text{und} \quad N(c) = \{(0,0),(1,1)\}$$

3. $\Sigma = \mathbb{R}$, $X = \Sigma^2$ und $c : X \to \{0, 1\}$ definiert durch

$$c(x) = c((x_1, x_2)) = \begin{cases} 1 & : \quad {x_1}^2 + {x_2}^2 \leq 1 \\ 0 & : \quad sonst \end{cases}$$

Für die Menge der positiven Beispiele ist $P(c) = B_2$, wobei B_2 die Einheitskugel im $\mathbb{R}^2$ bzgl. der euklidischen Norm $|| \cdot ||_2$ ist.
Die Menge der negativen Beipiele ist dann durch $\mathbb{R}^2 \setminus B_2^C$.

4. $\Sigma = \mathbb{R}$, $X = \Sigma^2$, $c : X \to \{0, 1\}$ definiert durch

$$c(x) = c((x_1, x_2)) = \begin{cases} 1 & : \quad x_1 \geq 0 \\ 0 & : \quad sonst \end{cases}$$

$P(c)$ ist der positive Halbaum des $\mathbb{R}^2$ und
$N(c)$ der negative Halbraum.

Parametrisierte Konzepte

1. $\Sigma = \mathbb{R}$, $X = \Sigma^2$. Ferner $w = (w_1, w_2) \in \mathbb{R}$, $r \in \mathbb{R}$ und $c_{w,r} : X \to \{0, 1\}$ definiert durch

$$c_{w,r}(x) = c_{w,r}((x_1, x_2)) = \begin{cases} 1 & : \quad (x_1 - w_1)^2 + (x_2 - w_2)^2 \leq r^2 \\ 0 & : \quad sonst \end{cases}$$

   $P(c)$ Kugel mit Radius r im $\mathbb{R}^2$ um den Mittelpunkt $w \in \mathbb{R}^2$.

2. $\Sigma = \mathbb{R}$, $X = \Sigma^2$. Ferner $w = (w_1, w_2)$ und $c_{w,\theta} : X \to \{0, 1\}$ definiert durch

$$c_{w,\theta}(x) = c_{w,\theta}((x_1, x_2)) = \begin{cases} 1 & : \quad w_1 x_1 + w_2 x_2 \geq \theta \\ 0 & : \quad sonst \end{cases}$$

   Dann ist $P(c)$ der positive Raum des $\mathbb{R}^2$ bzgl. w und θ und $N(c)$ der negative Halbraum.

Konzeptmengen I

1. Implizite Definition eines Konzeptes c durch Festlegung der positiven oder negativen Beispielmengen.
   $\Sigma = \{0, 1\}$, $n \in \mathbb{N}$ sei P_n die Menge der Palindrome, also

$$P_n = \{w \in \Sigma^n \mid w = w^R\}.$$

2. Implizite Definition eines Konzeptes c durch einen Algorithmus.
   $\Sigma = \{0, 1\}$ und die Grammatik $G = \{\{S\}, \Sigma, P, S\}$ mit dem Startsymbol S und den Produktionen

$$P = \{S \to 0S1, \quad S \to SS, \quad S \to 01\}$$

   Für $n \in \mathbb{N}$ sei $P_n := L(G) \cap \Sigma^{2n}$ gegeben (korrekte Klammerausdrücke mit n Klammerpaaren).

Konzeptmengen II

Parametrisierte Abbildungen: Menge der Perzeptrone im $\mathbb{R}^d$

$$\mathcal{P}_d := \{c_{w,\theta} \mid w \in \mathbb{R}^d \, \theta \in \mathbb{R}\}$$

Hierbei sei $c_{w,\theta}$ eine Perzeptronabbildung definiert durch:

$$c_{w,\theta}(x) = \begin{cases} 1 & : \quad \sum_{i=1}^{d} w_i x_i \leq \theta \\ 0 & : \quad sonst \end{cases}$$

Ein Konzept $c : X \to \{0, 1\}$ heißt **Perzeptron-lernbar**, falls es ein $w \in \mathbb{R}^d$ und ein $\theta \in \mathbb{R}$ gibt mit: $c_{w,\theta}(x) = c(x)$ für alle $x \in X$.

1. Welche Konzepte sind Perzeptron-lernbar ?

2. Wie findet man w und θ für Perzeptron-lernbare Konzepte?

Lineare Separierbarkeit

$A, B \subset \mathbb{R}^d$ nichtleere Mengen. Dann heißen A und B strikt linear trennbar, falls es eine durch $a = (a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1}$ bestimmte, eine Hyperebene $H_a \subset \mathbb{R}^d$ gibt

$$H_a = \{x \in \mathbb{R}^d \mid \sum_{i=1}^{d} a_i x_i = a_0\}$$

gibt mit

$$\sum_{i=1}^{d} a_i x_i > a_0 \quad \forall x \in A \quad \text{und} \quad \sum_{i=1}^{d} a_i x_i < a_0 \quad \forall x \in B$$

Sei nun $X \subset \mathbb{R}^d$ eine endliche Menge, dann ist $c : X \to \{0, 1\}$ ein Perzeptron-lernbar, g.d.w. $N(c)$ und $P(c)$ strikt linear separierbar sind.
Dann werden durch die Perzeptronlernregel $w \in \mathbb{R}^d$ und $\theta \in \mathbb{R}$ nach endlich vielen Lernschritten gefunden, so dass gilt: $c(x) = c_{w,\theta}(x)$ für alle $x \in X$.

Konzeptmengen III

Die Menge der Hyperkugeln im $\mathbb{R}^d$

$$\mathcal{B}_d := \{c_{w,r} \mid w \in \mathbb{R}^d,\ r \in \mathbb{R}\}$$

Hierbei sei $c_{w,r}$ die Indikatorfunktion für die Hyperkugeln mit Mittelpunkt w und Radius r definiert durch:

$$c_{w,r}(x) = \begin{cases} 1 & : \quad \sum_{i=1}^{d} (x_i - w_i)^2 \leq r^2 \\ 0 & : \quad sonst \end{cases}$$

Ein Konzept $c : X \to \{0, 1\}$ heißt **RBF-lernbar**, falls es ein $w \in \mathbb{R}^d$ und ein $r \in \mathbb{R}$ gibt mit: $c_{w,r}(x) = c(x)$ für alle $x \in X$.

1. Welche Konzepte sind RBF-lernbar ?

2. Wie kann der w und r für RBF-lernbare Konzepte finden?

Radiale Separierbarkeit

$A, B \subset \mathbb{R}^d$ nichtleere Mengen. Dann heißen A und B strikt radial trennbar, falls es eine durch $a = (a_0, a_1, \ldots, a_d) \in \mathbb{R}^{d+1}$ bestimmte, eine Hypersphäre $S_a \subset \mathbb{R}^d$ gibt

$$S_a = \{x \in \mathbb{R}^d \mid \sum_{i=1}^{d} (a_i - x_i)^2 = a_0^2\}$$

gibt

$$\sum_{i=1}^{d} (a_i - x_i)^2 > a_0^2 \quad \forall x \in A \quad \text{und} \quad \sum_{i=1}^{d} (a_i - x_i)^2 < a_0^2 \quad \forall x \in B$$

oder umgekehrt, d.h. mit vertauschen Rollen von A und B.

Sei nun $X \subset \mathbb{R}^d$ eine endliche Menge, dann ist $c : X \to \{0, 1\}$ ein RBF-lernbar, g.d.w. $N(c)$ und $P(c)$ strikt radial trennbar sind.

Beispielmenge

- $X \subset \Sigma^n$ oder $X \subset \Sigma^*$;
  $c : X \to \{0, 1\}$ ist ein Konzept, *allerdings meistens nicht explizit bekannt*.

- **Stichprobe** der Länge M ist eine M-elementige Folge aus $X \times \{0, 1\}$

$$s(M, c) = ((x^1, c(x^1)), \ldots, (x^M, c(x^M)))$$

- $\mathcal{S}(M, c)$ die Menge aller Stichproben $s(M, c)$ der Länge M von c

- $\mathcal{S}(c) = \bigcap_{M=1}^{\infty} \mathcal{S}(M, c)$ die Menge aller Stichproben beliebiger Länge von c.

Schreiben auch $s(c)$ oder $s(M)$ oder auch einfach s, falls M bzw. c bzw. M und c aus dem Kontext klar sind (analog für Stichprobenmengen $\mathcal{S}(M, c)$ und $\mathcal{S}(c)$).

Lernalgorithmus

- $X = \Sigma^n$ oder $X = \Sigma^*$.
- $C \subset \mathcal{F}_X$ eine Menge von Konzepten, die gelernt werden sollen.
- $H \subset \mathcal{F}_X$ eine Menge von Konzepten, die von einer Maschine/Algorithmus berechnet werden können, genannt die Hypothesenmenge.
- Konzepte $c \in C$ sollen durch Hypothesen $h \in H$ approximiert werden.
- $c \in C$ sei ein Konzept, das nicht explizit gegeben ist.
- c ist nur auf einer endliche Stichprobe vo Beispielen bekannt, also durch $M \in \mathbb{N}$ und

$$s(M, c) = ((x^1, c(x^1)), \ldots, (x^M, c(x^M))) \in \mathcal{S}(M, c)$$

- Für $c \in C$ ist eine Hypothese $h \in H$ gesucht, so dass h **möglichst genau** mit c übereinstimmt, idealerweise $h = c$ (auf der gesamten Eingabemenge X)

- Ein **Lernalgorithmus** L ist eine Abbildung $L : \mathcal{S}(c) \to H$ die einer Stichprobe $s(c)$ eine Hypothese h zuordnet, also $L(s(c)) = h$, für die möglichst $h_{|s} = c_{|s}$ gilt (auf der Stichprobe!).

- Genauer heißt L ein (C, H)-Lernalgorithmus.

- Falls $C \neq H$ gilt i.a. kann $h \neq c$, sogar $h_{|s} \neq c_{|s}$.

- Ein (H, H)-Lernalgorithmus heißt **konsistent** falls gilt:

$$h_{|s} = c_{|s} \quad \forall s$$

- Beispiel: Das Perzeptron-Lernverfahren ist also konsistent. I.a. wird aber ein Fehler auf gesamten Eingabemenge X vorhanden sein.

Lernen von Halbgeraden - Definition

Es sei $\theta \in \mathbb{R}$. Die positive Halbgerade $r_\theta : \mathbb{R} \to \{0, 1\}$ ist definiert durch

$$r_\theta(x) = \begin{cases} 1 & : \quad x \geq \theta \\ 0 & : \quad x < \theta \end{cases}$$

Durch $\mathcal{R} = \{r_\theta \mid \theta \in \mathbb{R}\}$ ist die Menge aller positiven Halbgeraden gegeben.

Es sei nun r_θ nicht explizit gegeben, also θ nicht bekannt.

Gegeben sei aber eine M-elementige Stichprobe von r_θ:

$$s(r_\theta) = ((x^1, r_\theta(x^1)), \ldots, (x^M, r_\theta(x^M)))$$

Gesucht ist nun eine Hypothese $h \in \mathcal{R}$, die r_θ möglichst gut approximiert.

Lernen von Halbgeraden - Algorithmus

Idee: Gegeben sei die Stichprobe $s = s(r_\theta)$. Ausgabe des Algorithmus: $L(s) = r_\lambda$ mit

$$\lambda = \min P(s(r_\lambda))$$

$P(s(r_\lambda))$ die Menge der positiven Beispiele von $s(r_\lambda)$

$\lambda = \infty$
for i := 1 to M

if $r_\theta(x^i) = 1$ and $\lambda > x^i$ then $\lambda := x^i$

$L(s) := r_\lambda$

Es gilt: $\lambda \geq \theta$. Falls $\lambda \neq \theta$ gilt: $r_\theta \neq r_\lambda$.

Anforderung beim Lernen: Die Wahrscheinlichkeit, dass eine Stichprobe $s(c)$ eine große Differenz zwischen c und $h = L(s(c))$ ergibt, soll klein sein.

PAC-Lernbarkeit

Ein Lernalgorithmus L heißt **probably approximately correct – pac —** für die Hypothesenmenge H, g.d.w. für beliebiges δ, ε mit $0 < \delta < 1$ und mit $0 < \varepsilon < 1$ ein $m_0 = m_0(\delta, \varepsilon) \in \mathbb{N}$ existiert, so dass für jedes $c \in H$ mit Wahrscheinlichkeit $1 - \delta$ die Differenz zwischen $c \in H$ und $L(s(c, m)) \in H$ kleiner ist als ε falls nur $m \geq m_0$ gilt.

$1 - \delta$ heißt die **Konfidenz** und ε heißt der **Fehler**.

Probleme bei dieser Definition

1. Was bedeutet mit hoher Wahrscheinlichkeit?

2. Wie wird die Differenz zwischen Konzepten gemessen?

Hierfür brauchen wir einige Grundbegriffe der W-Theorie.

Wahrscheinlichkeitstheorie

Charakteristische Funktion einer Menge $A \subset X$.

$$\mathbf{I}_A(x) = \begin{cases} 1 & : \quad x \in A \\ 0 & : \quad \text{sonst} \end{cases}$$

σ-Algebra. Gegeben eine Menge X, $\Omega \subset \wp(X)$ eine Menge von Teilmengen aus X heißt eine *σ-Algebra* über X, gdw.

1. $X \in \Omega$

2. Für $A \in \Omega$ ist auch $A^c = X \setminus A \in \Omega$.

3. Für $A_i \in \Omega$, $i = 1, 2, 3, \ldots, \infty$ gilt: $\bigcup_{i=1}^{\infty} A_i \in \Omega$ und $\bigcap_{i=1}^{\infty} A_i \in \Omega$.

Eine σ-*Algebra* ist abgeschlossen gegenüber Komplement, abzählbarer Vereinigung und Durchschnitt.

Borel-Mengen. Sei $X = \mathbb{R}^n$, die Borel-Mengen $\mathfrak{B}_n$ sind die kleinste σ-*Algebra*, die alle offenen Intervalle

$$\{(x_1, \ldots, x_n) \ : \ x_i \in (a_i, b_i), a_i, b_i \in \mathbb{R}\}$$

in $\mathbb{R}^n$ enthält.

Wahrscheinlichkeitsraum. X gegeben. $\Omega \subset \wp(X)$ eine σ-*Algebra* über X. Ein Wahrscheinlichkeitsraum ist ein Tripel $(X, \Omega, \mathbf{P})$ wobei $\mathbf{P} : \Omega \to [0, 1]$ ein Wahrscheinlichkeitsmaß ist, dh.

1. $\mathbf{P}(X) = 1$

2. Für $A_i \in \Omega$, $i = 1, 2, 3, \ldots, \infty$ paarweise disjunkt gilt:

$$\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)$$

Das Paar (X, Ω) heißt auch Maßraum oder meßbarer Raum.

Meßbarkeit. (X, Ω) ein Maßraum. Eine reellwertige Funktion $f : X \to \mathbb{R}$ heißt Ω-meßbar (oder auch einfach nur meßbar), gdw.

$$\forall z \in \mathbb{R} \qquad \{x \in X \ : \ f(x) \leq z\} \in \Omega$$

Zufallsvariable. (X, Ω) ein Maßraum. Eine Zufallsvariable ist eine Ω-meßbar reellwertige Abbildung $f : X \to \mathbb{R}$.

Zufallsvariablen häufig in sans serif, also $\mathsf{Y} = f(\mathsf{X})$. Zufallsvariable $\mathsf{Y} = f(\mathsf{X})$

induziert ein Maß auf $\mathbb{R}$:

$$\forall Y \in \mathfrak{B}_1 \quad \mathbf{P}_{\mathsf{Y}}(Y) = \mathbf{P}_{\mathsf{X}}\{x \in X \ : \ f(x) \in Y\}$$

Verteilungsfunktion und Dichtefunktion. Für eine Zufallsvariable X heißt die Funktion $\mathbf{F}_{\mathsf{X}} : \mathbb{R}to[0,1]$ definiert durch

$$\mathbf{F}_{\mathsf{X}}(x) = \mathbf{P}_{\mathsf{X}}(\mathsf{X} \leq x)$$

die Verteilungsfunktion von X.

Die Funktion $\mathbf{f}_{\mathsf{X}} : \mathbb{R} \to \mathbb{R}$ heißt die Dichtefunktion, gdw.

$$\forall z \in \mathbb{R} \qquad \mathbf{F}_{\mathsf{X}}(z) = \int_{x \leq z} \mathbf{f}_{\mathsf{X}}(x)dx.$$

Erwartungswert. $f : X \to \mathbb{R}$ eine meßbare Abbildung. Der Erwartungswert

von f ist

$$\mathbf{E}_{\mathsf{X}}\left[f(\mathsf{X})\right] = \int_{\mathbb{R}} f(x) d\mathbf{F}_{\mathsf{X}}(x) = \int_{\mathbb{R}} f(x)\mathbf{f}_{\mathsf{X}}$$

Produktraum. Für zwei Maßräume (X_1, Ω_1) und (X_2, Ω_2) ist der Produktraum $(X_1 \times X_2, \Omega_1 \times \Omega_2)$ definiert, wobei $\Omega_1 \times \Omega_2$ die kleinste σ-*Algebra* ist, welche die Menge $\{X \times Y \ : \ X \in \Omega_1, Y \in \Omega_2\}$ enthält.

Im Folgenden werden wir auf die σ-Algebra Ω nicht weiter eingehen, für den Fall $X \subset \mathbb{R}^d$ verwenden wir eben Borel-Algebra und für eine endliche Grundmenge X treten ohnehin keine technischen Schwierigkeiten auf.

Differenz zwischen Konzepten

Sei nun $H \subset \mathcal{F}_{\mathcal{X}}$ eine Konzeptmenge, $c \in H$ ein zu beliebiges (zu lernendes) Konzept. Dann definieren wir den **Fehler** von einer Hypothese $h \in H$ bzgl. c als die Wahrscheinlichkeit des Ereignis $h(x) \neq c(x)$, also

$$err_{\mathbf{P}}(h, c) = \mathbf{P}\{x \in X \ : \ h(x) \neq c(x)\}$$

Dabei gehen wir davon aus, dass $\{x \in X \ : \ h(x) \neq c(x)\}$ eine meßbare Menge ist, also $\in \Omega$.

Schreiben auch $err_{\mathbf{P}}(h)$ oder $err(h)$ für $err_{\mathbf{P}}(h, c)$, falls c bzw. $\mathbf{P}$ aus dem Kontext eindeutig festgelegt sind.

Der Fehler lässt sich auch so schreiben:

$$err_{\mathbf{P}}(h, c) = \int_X \mathbf{I}_{\{x \in X \, : \, h(x) \neq c(x)\}} d\mathbf{P}(x) =$$

PAC-Lernbarkeit

Ein Lernalgorithmus L heißt **probably approximately correct** (pac) für die Hypothesenmenge H, g.d.w. für beliebiges δ, ε mit $0 < \delta < 1$ und mit $0 < \varepsilon < 1$ ein $m_0 = m_0(\delta, \varepsilon) \in \mathbb{N}$ existiert, so dass für jedes Zielkonzept $c \in H$ und für jedes Wahrscheinlichkeitsmaß $\mathbf{P}$

$$\mathbf{P}^m\{s \in \mathcal{S}(m, c) \ : \ err_{\mathbf{P}}(L(s)) < \varepsilon\} > 1 - \delta$$

für $m \geq m_0$.

- m_0 hängt von δ, ε
- m_0 ist unabhängig von $\mathbf{P}$
- In der Bedingung sind die Größen δ und ε über $\mathbf{P}$ verkoppelt. Deshalb ist es möglich, dass die Bedingung für alle $\mathbf{P}$ gilt.

Halbgeraden Lernalgorithmus ist PAC

- Gegeben seien $\delta, \varepsilon \in (0,1)$, ein Maß $\mathbf{P}$ auf $\mathbb{R}$ und ein zu lernendes Konzept r_θ.
- Es sei ferner $s \in \mathcal{S}(M, r_\theta)$ eine Stichprobe von r_θ der Länge M.
- Der Halbgeraden-Lernalgorithmus liefert nun $L(s) = r_\lambda$, mit $\lambda \geq \theta$. Dann ist $[\theta, \lambda)$ die Fehlermenge.
- Für ε und $\mathbf{P}$ definieren wir

$$\beta_0 = \sup\{\beta \mid \mathbf{P}[\theta, \beta) < \varepsilon\}$$

- Dann ist offenbar $\mathbf{P}[\theta, \beta_0) < \varepsilon$ und $\mathbf{P}[\theta, \beta_0] \geq \varepsilon$.

- Falls nun $\lambda \leq \beta_0$ ist, so gilt:

$$\mathrm{err}_{\mathbf{P}}(L(s)) = \mathbf{P}([\theta, \lambda)) \leq \mathbf{P}([\theta, \beta_0)) \leq \varepsilon$$

- Wie groß ist nun die Wahrscheinlichkeit für das Ereignis $\lambda \leq \beta_0$?

- Ist offenbar die Wahrscheinlichkeit dafür, dass es mindestens ein Beispiel $(x_i, r_\theta(x_i))$ in s gibt, mit $x_i \in [\theta, \beta_0]$.

- Wegen $\mathbf{P}[\theta, \beta_0] \geq \varepsilon$ ist die Wahrscheinlichkeit, dass ein einziges Beispiel nicht in $[\theta, \beta_0]$ liegt kleiner als $1 - \varepsilon$.

- Die Wahrscheinlichkeit, dass nun alle M Beispiele nicht in $[\theta, \beta_0]$ liegen ist damit kleiner als $(1 - \varepsilon)^M$. (Unabhängigkeit der Beispiele!)

- Wahrscheinlichkeit für $\lambda \geq \beta_0$ ist damit mindestens $1 - (1 - \varepsilon)^M$.

- Damit ist nun gezeigt,

$$\mathbf{P}^M\{s \in \mathcal{S}(M, r_\theta) \mid \mathrm{err}_{\mathbf{P}}(L(s)) < \varepsilon\} \geq 1 - (1-\varepsilon)^M$$

- Beobachtung: Die rechte Seite is unabhängig von r_θ und $\mathbf{P}$!

- Falls nun $M \geq M_0 = \lceil \frac{1}{\varepsilon} \ln \frac{1}{\delta} \rceil$ ist gilt:

$$(1-\varepsilon)^M \leq (1-\varepsilon)^{M_0} = \exp(M_0 \ln(1-\varepsilon)) < \exp(-M_0\varepsilon) = \exp(\ln \delta) = \delta$$

- Damit ist gezeigt, dass L pac ist.

In der **pac**-Definition wurde strikt $<$ bzw. $>$ verwendet. Dies ist kein Unterschied zu $\leq$ bzw. $\geq$

Beispiel: Fordern wir also beispielsweise $\varepsilon = 0.01$ und $\delta = 0.001$,
so folgt $M_0 = \lceil 100 \ln 1000 \rceil = 691$

Lernbarkeit von Funktionenmengen

- **pac**-Lernbarkeit ist die Eigenschaft eines Algorithmus!
- Gegeben ein Algorithmus L so können wir versuchen zu zeigen, dass er **pac** ist. Dies erfordert meist sehr spezifische Argumente. Deshalb sind wir an einer allgemeineren Methode interessiert.
- Betrachten nun (H, H)-Lernverfahren, d.h. die zu lernende Funktion c ist aus H.
- Ferner beschränken wir uns auf konsistente Lernverfahren, d.h. solche Lernverfahren, die auf der Stichprobe keine Fehler produzieren, also stets

$$L(s(c))_{|s(c)} = c_{|s(c)}$$

Definitionen

- Für eine Stichprobe $s \in \mathcal{S}(M, c)$ und eine Menge $H \subset \mathcal{F}_X$ bezeichnen wir

$$H[s] = \{h \in H \mid h(x_i) = c(x_i), i = 1, \ldots, M\}$$

  die Menge der Hypothesen $h \in H$ die mit $c \in H$ auf der Stichprobe s übereinstimmen. Die Menge der mit s konsistenten Hypothesen.

- Es sein nun wieder $\mathbf{P}$ ein W-Maß auf X. Für $\varepsilon \in (0, 1)$ definieren wir

$$B_\varepsilon = \{h \in H \mid \mathrm{err}(h, c) \geq \varepsilon\}$$

  Die Menge der Hypothesen $h \in H$, die sich von c um mindestens ε unterscheiden.

- Ein konsistenter Lernalgorithmus liefert also stets eine Approximierende $L(s(c)) \in H[s]$.

- Die **pac**-Eigenschaft erfordert nun, dass $L(s(c))$ mit hoher Wahrscheinlichkeit nicht in B_ε liegt.

- Eine Menge H heißt (potenziell) **pac-lernbar**, falls es für $\delta, \varepsilon \in (0,1)$ stets ein $M_0 \in \mathbb{N}$ gibt, so dass falls $M \geq M_0$ ist auch folgt

$$\mathbf{P}^M\{s \in \mathcal{S}(M,c) \mid H[s] \cap B_\varepsilon = \emptyset\} > 1 - \delta$$

  für jedes W-Maß $\mathbf{P}$ und jede Funktion $c \in H$.

Bemerkung: Falls also H pac-lernbar ist und L ein konsistenter Lernalgorithmus, dann ist L auch pac.

Endliche Funktionenmengen sind PAC Lernbar

- $H \subset \mathcal{F}_X$ sei endlich, also $|H| < \infty$.
- $\varepsilon, \delta, \mathbf{P}$ und $c \in H$ seien gegeben.
- Zeigen nun dass die Wahrscheinlichkeit für $H[s] \cap B_\varepsilon \neq \emptyset < \delta$ für Stichproben s mit vielen Beispielen, d.h. M groß genug.
- Für jedes $h \in B_\varepsilon$ ist

$$\mathbf{P}\{x \in X \ : \ h(x) = c(x)\} = 1 - \mathrm{err}(h, c) \leq 1 - \varepsilon$$

- Damit gilt

$$\mathbf{P}^M\{s \in \mathcal{S}(M, c) \ : \ h(x_i) = c(x_i),\ 1 \leq i \leq M\} \leq (1 - \varepsilon)^M$$

- Dies ist nun die Wahrscheinlichkeit, dass eine bestimmte Funktion $h \in B_\varepsilon$ auch in $H[s]$ ist.

- Die Wahrscheinlichkeit, dass nun irgendeine Funktion $h \in B_\varepsilon$ auch in $H[s]$ liegt, ist dann also

$$\mathbf{P}^M\{s \in \mathcal{S}(M,c) \mid H[s] \cap B_\varepsilon = \emptyset\} \leq |H|(1-\varepsilon)^M$$

- Diese Wahrscheinlichkeit ist $\leq \delta$, falls gilt:

$$M \geq M_0 = \lceil \frac{1}{\varepsilon} \ln \frac{|H|}{\delta} \rceil$$

- Dann ist

$$|H|(1-\varepsilon)^M \leq |H|(1-\varepsilon)^{M_0} < |H|\exp(-\varepsilon M_0) \leq |H|\exp(\ln(\frac{\delta}{|H|})) = \delta$$

Bemerkungen

- Theorem enthält den Bool'schen Eingaberaum $X = \{0,1\}^n$.
- Jeder konsistente Algorithmus auf $X = \{0,1\}^n$ ist pac.
- B_n der Bool'sche Funktionenraum über $X = \{0,1\}^n$:

$$|B_n| = 2^{2^n}$$

- Die Schranke für M_0 ist somit

$$M_0 = \lceil \frac{2^n}{\varepsilon} \ln \frac{2}{\delta} \rceil$$

2. Vapnik Chervonenkis Dimension

1. Wachstumsfunktion
2. Positive Halbgeraden
3. VC-Dimension
4. Perzeptrone
5. Satz von Sauer

Wachstumsfunktion

Wir haben gezeigt: Falls $|H| < \infty$, dann ist H auch pac-lernbar.

Beweis benötigt die Endlichkeit von H.

Offenbar gibt es viele nichtendliche Hypothesenmengen, etwa auf $\mathbb{R}$ definierte Funktionenklassen.

Viele nicht endliche Hypothesenmengen haben eine sehr spezielle Struktur, z.B. lineare Funktionen.

Idee: Statt der Kardinalität soll die Ausdrucksmächtigkeit von H betrachtet werden.

Diese muss aber zuerst formalisiert werden!

Es sei $X = \mathbb{R}^n$ oder $X = \{0,1\}^n$ und eine Funktionenmenge aus $\mathcal{F}_X$

$$H := \{h \mid h : X \to \{0,1\}\}$$

Sei nun $x \in X^M$, $x = (x^1, \ldots, x^M)$ dann sei

$$\pi_H(x) = |\{(h(x^1), \ldots, h(x^M) \mid h \in H\}|$$

die Zahl der möglichen Klassifikationen von x durch H.

Offenbar gilt: $\pi_H(x) \leq 2^M$

Wir definieren nun die Wachstumsfunktion $\Pi_H : \mathbb{N} \to \mathbb{N}$ durch

$$M \mapsto \max\{\pi_H(x) \mid x \in X^M\}$$

Positive Halbgeraden

Sei $\theta \in \mathbb{R}$ und die positive Halbgerade $r_\theta : \mathbb{R} \to \{0, 1\}$ mit

$$r_\theta(x) = \begin{cases} 1 & : \quad x \geq \theta \\ 0 & : \quad x < \theta \end{cases}$$

$H = \{r_\theta \mid \theta \in \mathbb{R}\}$ die Menge aller positiven Halbgeraden.

$x = (x^1, \ldots, x^M) \in \mathbb{R}^M$ eine Stichprobe mit M Elementen und oBdA gelte

$$x^1 < x^2 < \cdots < x^M$$

Die Menge aller möglichen Klassifikationen ist nun

$$(1, 1, \ldots, 1, 1), (0, 1, \ldots, 1, 1), \ldots, (0, 0, \ldots, 0, 1), (0, 0, \ldots, 0, 0) \in \{0, 1\}^M$$

Somit folgt: $\pi_H(x) = M + 1$ und offenbar auch $\Pi_H(M) = M + 1$.

VC Dimension

In den 1970er Jahren wurde das Konzept der Vapnik-Chervonenkis-Dimension (kurz VC-Dimension) als ein Maß für die Ausdrucksstärke von Mengen binär-wertiger Funktionen entwickelt.

Die VC-Dimension wurde erstmals von Blumer im Jahre 1986 mit der PAC-Lernbarkeit in Verbindung gebracht, es gilt:

$$H \text{pac-lernbar} \iff \text{VCdim}(H) < \infty$$

Es gilt:

$$\text{VCdim}(H) = \max\{M \mid \Pi_H(M) = 2^M\}$$

Falls das Maximum nicht existiert, so setzen wir $\text{VCdim}(H) = \infty$.

Definition: Eine Stichprobe $x = (x^1, \ldots, x^M)$ wird durch H *zertrümmert*, gdw. es für jede Teilmenge $A \subset \{x^1, \ldots, x^M\}$ eine Hypothese $h \in H$ gibt mit $h(x^i) = 1$ gdw. $x^i \in A$.

VCdim$(H) = d \in \mathbb{N}$ dann gibt es eine Stichprobe von Länge d die H zertrümmert wird, aber es gibt keine Stichprobe der Länge $d+1$ die H zertrümmert wird.

VCdim$(H) = \infty$, dann gibt es für jede Zahl $d \in \mathbb{N}$ eine Stichprobe der Länge d die H zertrümmert wird.

Beispiel: H sei die Menge der positiven Halbgeraden, dann ist offenbar VCdim$(H) = 1$.

Perzeptrone

Menge der Perzeptrone im $\mathbb{R}^n$

$$P_n := \{c_{w,\theta} \mid w \in \mathbb{R}^n, \theta \in \mathbb{R}\}$$

Hierbei sei $c_{w,\theta}$ eine Perzeptronabbildung definiert durch:

$$c_{w,\theta}(x) = \begin{cases} 1 & : \quad \sum_{i=1}^{n} w_i x_i \leq \theta \\ 0 & : \quad sonst \end{cases}$$

Satz: VCdim$(P_n) = n + 1$

Satz von Cover (1964):

$$\Pi_{P_n}(M) = 2 \sum_{i=0}^{n} \binom{M-1}{i}$$

Satz von Sauer

Es sei H eine Hypothesenmenge mit endlicher VC-Dimension, also $0 \leq$ VCdim$(H) = d < \infty$. Ferner $M \in \mathbb{N}$.
Dann gilt:

$$\Pi_H(M) \leq 1 + \binom{M}{1} + \binom{M}{2} + \cdots + \binom{M}{d}$$

Mit

$$\phi(d, M) := 1 + \binom{M}{1} + \binom{M}{2} + \cdots + \binom{M}{d}$$

gilt für $M \geq d \geq 1$

$$\phi(d, M) < (\frac{eM}{d})^d \quad e = 2,71...$$

d.h. also gilt

$$\Pi_H(M) < (\frac{eM}{d})^d$$

Beweis: Satz von Sauer

Sei $d = \text{VCdim}(H) = 0$, dann ist für jedes $x \in X$ offenbar $h(x) = c$ mit $c \in \{0, 1\}$ für alle $h \in H$.

Also ist $\pi_H(x) = 1$ für alle $x = (x^1, \ldots, x^M)$.

Somit ist dann $\Pi_H(M) = 1 = \phi(0, M)$.

Damit ist der Satz für $d = 0$ gezeigt.

Für $d \geq 1$ und $M = 1$ gilt dann

$$\Pi_H(1) \leq 2^1 = \phi(d, 1) = \binom{1}{0} + \binom{1}{1}$$

Damit ist der Satz auch für $d \geq 1$ und $M = 1$ gezeigt.

Beweis nun durch Induktion über $d + M$.

Der Fall $d + M = 2$ ist bereits gezeigt: Dies betrifft $d = 0$ und $M = 2$ oder $d = M = 1$.

Der Satz gelte nun für $d + M \leq k$ mit $k \geq 2$.

Sei nun H eine Hypothesenmenge mit VCdim$(H) = d$ und sei $x = (x^1, \ldots, x^M)$ mit $d + M = k + 1$.

Die Fälle $(d, M) = (0, k + 1)$ und $(d, M) = (k, 1)$ sind bereits gezeigt, deshalb können wir $d \geq 1$ und $M \geq 2$ annehmen.

Also sei $x = (x^1, \ldots, x^M)$ eine Stichprobe mit M paarweise verschiedenen Beispielen (sonst kann man die Stichprobe verkürzen und die Induktionshypothese anwenden).

Setze $E := \{x^1, \ldots, x^M\}$, dann ist $|E| = M$.

Wir betrachten die Einschränkung von H auf E, also $H_E := H|E$.

Dann ist offenbar $\pi_H(x) = |H_E|$ und wir zeigen deshalb:

$$|H_E| \leq \phi(d, M)$$

Es sei $F = E \setminus \{x^M\}$ und entsprechend $H_F := H|F$

Zwei verschiedene Funktionen $g, h \in H_E$ sind identisch auf F, gdw. $h = g$ auf F und $h(x^M) \neq g(x^M)$.

H_* sei nun die Menge der Hypothesen aus H_F, die auf diese beiden Arten aus 2 verschiedenen Hypothesen von H_E entstehen.

Sei also $h_* \in H_*$, so gibt es 2 mögliche Hypothesen in H_E. Damit gilt dann:

$$|H_E| = |H_F| + |H_*|$$

Wir untersuchen nun $|H_F|$ und $|H_*|$.

Sei $\tilde{x} = (x^1, \ldots, x^{M-1})$ dann gilt

$$|H_F| = \pi_H(\tilde{x}) \leq \Pi_H(M-1)$$

Nach Induktionsvoraussetzung gilt dann

$$|H_F| \leq \Pi_H(M-1) \leq \phi(d, M-1)$$

wegen $d + (M-1) \leq k$.

Wir zeigen nun VCdim$(H_*) \leq d-1$.
Zum Beweis nehmen wir an, das H_* die Stichprobe $z = (z^1, \ldots, z^d)$ zertrümmert.

Für jede Hypothese $h_* \in H_*$ gibt es $h_1, h_2 \in H_E$ mit $h_1 = h_2$ auf F und mit $h_1(x^M) \neq h_2(x^M)$.

Dann aber zertrümmert H_E und damit auch H die Stichprobe $(z^1, \ldots, z^d, x^M)$ von Länge $d+1$.

Dies steht aber im Widerspruch zu VCdim$(H) = d$.
Also gilt VCdim$(H_*) \leq d - 1$.

Damit folgt

$$|H_*| = \pi_{H_*}(\tilde{x}) \leq \Pi_{H_*}(M-1) \leq \phi(d-1, M-1)$$

denn $(d-1) + (M-1) \leq k$.

Insgesamt folgt so:

$$\pi_H(x) = |H_E| = |H_F| + |H_*| \leq \phi(d, M-1) + \phi(d-1, M-1) = \phi(d, M)$$

Hierbei nutzt man die bekannte Rekursionsformel für Binomialkoefizienten:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

.

Es bleibt noch zu Zeigen, dass für $M \geq d \geq 1$ gilt:

$$\phi(d, M) < \left(\frac{eM}{d}\right)^d \qquad e = 2,71...$$

Offenbar ist für alle $0 \leq i \leq d$ wegen $M \geq d$:

$$(M/d)^d (d/M)^i \geq 1$$

Somit folgt:

$$\sum_{i=1}^{d} \binom{M}{i} \leq (M/d)^d \sum_{i=1}^{d} \binom{M}{i} (d/M)^i \leq (M/d)^d (1 + (d/M))^M < (Me/d)^d$$

denn es gilt $(1 + x/n)^n < e^x$ für alle $x \in R$ und $n \in \mathbb{N}$.

Ferner $(1 + x/n)^n \to e^x$ bei $n \to \infty$.

Folgerung : Für eine Hypothesenmenge H mit VCdim$(H) = d$ und $M \geq d$ gitl:

$$d \leq \log_2 \Pi_H(M) \leq d \log_2(eM/d)$$

Satz : Es sei F ein endlichdimensionaler Vektorraum rellwertiger Funktionen, auf X, $g : X \to \{0, 1\}$ ein Konzept auf X und

$$H := \{\text{sgn}(f + g) \ : \ f \in F\}$$

Dann gilt:

$$\text{VCdim}(H) = \dim(F)$$

4. VC Dimension und Lernbarkeit

1. VCdim nicht beschränkt

2. VCdim endlich

5. VC-Dimension und Architekturen aus der Praxis

1. Mehrschicht Netze

2. Perzeptrone und Support-Vektor-Maschinen

3. Ensemble Methoden

Mehrschicht Netze

Perzeptron Lernen

Lernregel:

$$\Delta w = l\,(T - y) \cdot x \text{ mit Lehrersignal } T \qquad (1)$$

andere Schreibweise der Lernregel:

$$\Delta w = -l\,\text{sign}(x \cdot w) \cdot x = l\,T \cdot x \quad \text{falls } T \neq y \text{ (Änderungsschritt)} \qquad (2)$$

Zu bestimmen: $S =$ Anzahl der Änderungsschritte

Problem lösbar, falls $\exists\, w$ mit $\text{sign}(x^\mu \cdot w) = T^\mu\ \forall \mu$,
d.h. $T^\mu\,(x^\mu \cdot w) > 0\ \forall \mu$, d.h. $D(w) := \min_{\mu=1}^{M} T^\mu\,(x^\mu \cdot w) > 0$.

$D(w)$ nimmt auf der Einheitskugel $K = \{w\ :\ w \cdot w = 1\}$ das Maximum d an.
Also gibt es w^* mit $w^* \cdot w^* = 1$ und $D(w^*) = d$.
Problem lösbar, falls $d > 0$. Sei nun $c := \max_{\mu=1}^{M}(x^\mu \cdot x^\mu)$.

Betrachte das Gewicht w_S nach S Änderungsschritten: $w_S = \sum_{i=1}^{S}(\Delta w)_i$.

Dann gilt:

$$(\Delta w) \cdot w^* \overset{(2)}{=} l\, T^\mu\, (x^\mu \cdot w^*) \geq l\, D(w^*) = l\, d \tag{3}$$

$$\begin{aligned} (w + \Delta w) \cdot (w + \Delta w) - w \cdot w &= 2\,((\Delta w) \cdot w) + (\Delta w) \cdot (\Delta w) \\ &\overset{(2)}{=} -2\, l\, \text{sign}(x^\mu \cdot w)\ (x^\mu \cdot w) + l^2\, (x^\mu \cdot x^\mu) \\ &\leq l^2\, (x^\mu \cdot x^\mu) \leq l^2\, c \end{aligned} \tag{4}$$

Also gilt: $w_S \cdot w_S \overset{(4)}{\leq} S l^2 c$ und $w_S \cdot w^* \overset{(3)}{\geq} S l d$. Daraus folgt:

$$S\, l\, d \overset{(3)}{\leq} w_S \cdot w^* \leq \sqrt{(w_S \cdot w_S)(w^* \cdot w^*)} = \sqrt{w_S \cdot w_S} \leq \sqrt{S\, l^2\, c} \quad \Longrightarrow S \leq c/d^2$$

Support Vektor Lernen

Ist zunächst einmal eine spezielle Form des Perzeptron-Lernverfahrens.

Lernverfahren entsteht durch eine Kombination von 2 Zielen, diese legen im Fall linear separierbarer Mengen eine eindeutige Trennhyperebene fest.

Wieder gegeben Trainingsdaten

$$\mathcal{M} = \{(x^\mu, T^\mu) \ : \ \mu = 1, \ldots, M\} \subset \mathbb{R}^d \times \{-1, 1\}$$

Wir nehmen zunächst einmal an, die Mengen

$$P = \{x^\mu \mid T^\mu = 1\} \quad \text{und} \quad N = \{x^\mu \mid T^\mu = -1\}$$

seien linear separierbar.

Das Perzeptron-Lerntheorem sichert die Konverganz in endlich vielen Schritten gegen eine Lösung w (erweiterter Gewichtsvektor).

Wir suchen nun nach einer Lösung w^*, welche

1. Die Separationsbedingungen erfüllt:

$$T^\mu(\langle w, x^\mu \rangle + w_0) > 0 \quad \text{für alle } \mu = 1, \ldots, M$$

2. möglichst weit vonn den Mengen N und P entfernt ist (*maximal margin*)

Es sein

$$\min_\mu \; T^\mu(\langle w, x^\mu \rangle + w_0) = \delta > 0$$

Nun reskalieren wir und erhalten $w = \frac{1}{\delta} w$ und $w_0 = \frac{1}{\delta} w_0$ und erhalten

$$T^\mu(\langle w, x^\mu \rangle + w_0) \geq 1 \quad \text{für alle } \mu = 1, \ldots, M$$

Offenbar gibt es mindestens einen Punkt $x^\nu \in P$ und $x^\mu \in N$ mit

$$\langle w, x^\nu \rangle + w_0 = 1$$

und mit

$$\langle w, x^\mu \rangle + w_0 = -1$$

Daraus folgt $\langle w, x^\nu - x^\mu \rangle = 2$ und damit ist $D(w)$ die Breite des Randes der separierenden Hyperebene gegeben durch

$$D(w) = \langle \frac{w}{\|w\|_2}, (x^\nu - x^\mu) \rangle = \frac{2}{\|w\|_2}$$

Also Maximierung des Randes bedeutet Minimierung von

$$\varphi(w) = \frac{\|w\|_2^2}{2} \to \min$$

unter den Nebenbedingungen

$$T^\mu(\langle w, x^\mu \rangle + w_0) \geq 1 \quad \text{für alle } \mu = 1, \ldots, M$$

Dies ist ein quadratisches Optimierungsproblem unter Nebenbedingungen.

VC-Dimension optimaler Trennebenen

Gegeben Datendatz (x^μ, T^μ), $x^\mu \in \mathbb{R}^d$ mit

$$\|x^\mu - a\| \leq C$$

Betrachten nun die Menge der Trennebenen mit der Normalisierungsbedingung

$$\min_{x^\mu} |\langle w, x^\mu \rangle + b| = 1$$

Gilt zusätzlich $\|w\| \leq D$, dann hat die Menge der der Funktionen

$$f(x) = \text{sign}(\langle w, x \rangle + b)$$

die VC-Dimension h mit

$$h \leq \min\{d, \lceil C^2 D^2 \rceil\} + 1$$

Konstruktion der optimalen Trennebene

Mit der Einführung von sogenannten Lagrange Multiplikatoren $\alpha_\mu \geq 0$ für $\mu = 1, \ldots, M$ wird es in ein Optimierungsproblem ohne Nebenbedingungen überführt, dieses muss bzgl w und w_0 minimiert und bzgl. α maximiert werden.

$$L(w, w_0, \alpha) = \frac{\|w\|_2^2}{2} - \sum_{\mu=1}^{M} \alpha_\mu \left(T^\mu (\langle w, x^\mu \rangle + w_0) - 1 \right)$$

Setzt man nun wie gehabt die partiellen Ableitungen $\frac{\partial L}{\partial w} = 0$ und $\frac{\partial L}{\partial w_0} = 0$ so erhält man die Bedingungen

$$\sum_{\mu=1}^{M} \alpha_\mu T^\mu = 0 \quad \text{und} \quad w = \sum_{\mu=1}^{M} \alpha_\mu T^\mu x^\mu$$

Außerdem folgt aus den Optimierungsbedingungen

$$\alpha_\mu \left[T^\mu \left(\langle w, x^\mu \rangle + w_0 \right) - 1 \right] = 0 \quad \text{für alle } \mu = 1, \ldots, M$$

Falls nun $\alpha_\mu \neq 0$ so folgt: $T^\mu \left(\langle w, x^\mu \rangle + w_0 \right) = 1$, d.h. für solche Trainingsbeispiele liegt x^μ genau auf dem Rand.

Diese Vektoren heißen **Support Vektoren**. Offensichtlich ist w eine Linearkombination der Support Vektoren (geometrisch ist dies (jedenfalls im $\mathbb{R}^2$) klar).

$$w = \sum_{x^\mu \in SV} \alpha_\mu T^\mu x^\mu$$

Dann setzen wir diese Resultate in L ein und erhalten das quadratische Funktion

$$W(\alpha) = \sum_{\mu=1}^{M} \alpha_\mu - \frac{1}{2} \sum_{\nu=1}^{M} \sum_{\mu=1}^{M} \alpha_\nu \alpha_\mu T^\nu T^\mu \langle x^\nu, x^\mu \rangle$$

das mit $\alpha_\mu \geq 0$ für alle $\mu = 1, \ldots, M$ zu maximieren ist.

Die Lösung α^* liefert nun fest:

$$w^* = \sum_{\mu=1}^{M} \alpha_\mu^* T^\mu x^\mu$$

Die Schwelle $w_0^* \in \mathbb{R}$ läßt sich mit Hilfe eines Support Vektors x^{μ_0} bestimmen, denn dann gilt $\alpha_{\mu_0} = 0$ und damit

$$T^{\mu_0} \left(\langle w, x^{\mu_0} \rangle + w_0 \right) = 1$$

Also folgt

$$w_0^* = \frac{1}{T^{\mu_0}} - \langle w, x^{\mu_0} \rangle$$

damit liegt die Entscheigungsfunktion fest:

$$F(x) = \text{sgn } \left(\langle w^*, x \rangle + w_0^* \right) = \text{sgn } \left(\sum_{x^\mu \in SV} \alpha_\mu^* T^\mu \langle x^\mu, x \rangle + w_0^* \right).$$

Nicht separierbares Problem

$P = \{x^\mu \mid T^\mu = 1\}$ und $N = \{x^\mu \mid T^\mu = -1\}$ linear nicht separierbar

Soft Separationsbedingungen durch Schlupfvariable $\delta_\mu \geq 0$ (*slack variables*)

$$T^\mu \left(\langle w, x^\mu \rangle + w_0 \right) \geq 1 - \delta_\mu \quad \text{für alle } \mu = 1, \ldots, M$$

Nun minimieren wir mit $C > 0$

$$\varphi(w, \delta) = \frac{1}{2} \|w\|_2^2 + \frac{C}{M} \sum_{\mu=1}^{M} \delta_\mu$$

Dies führt wiederum auf die quadratische Funktion

$$W(\alpha) = \sum_{\mu=1}^{M} \alpha_\mu - \frac{1}{2} \sum_{\nu=1}^{M} \sum_{\mu=1}^{M} \alpha_\nu \alpha_\mu T^\nu T^\mu \langle x^\nu, x^\mu \rangle$$

die mit $0 \leq \alpha_\mu \leq C/M$ für alle $\mu = 1, \ldots, M$ zu maximieren ist.

Nichtlineares Support Vektor Lernen

$P = \{x^\mu \mid T^\mu = 1\}$ und $N = \{x^\mu \mid T^\mu = -1\}$ linear nicht separierbar

Nun transformieren wir x^μ gemäß einer Transformation $\phi : \mathbb{R}^d \to \mathcal{H}$, in einen Vektorraum mit Skalarprodukt (genauer ein Hilbertraum), z.B. $\mathcal{H}$ kann endlichdimensional sein, also $\mathcal{H} = \mathbb{R}^N$, aber auch ein unendlichdimensonaler Raum, etwa der Folgenraum $l^2(\mathbb{R})$.

Idee: Zunächst Transformation $z^\mu := \phi(x^\mu)$ nach $\mathcal{H}$ durchführen und dann das Support-Vektor-Lernproblem in $\mathcal{H}$ lösen (ist nichts Neues).

Die Entscheidgungsfunktion hat die Gestalt

$$F(x) = \text{sgn}\left(\sum_{\phi(x^\mu)\in SV} \alpha_\mu^* T^\mu \langle \phi(x^\mu), \phi(x)\rangle + w_0^*\right).$$

Abbildungen der Form

$$(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \to (\phi(x), \phi(y)) \in \mathcal{H} \times \mathcal{H} \to \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \in \mathbb{R}$$

lassen sich u.U. durch sogenannte **Mercer Kernfunktionen** $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ direkt darstellen.

Satz von Mercer: Sei $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ symmetrisch und gelte

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x) k(x, y) f(y) dx dy > 0$$

für alle $f \in L^2$ (quadratische integrierbare Funktionen). Dann gibt es einen Hilbertraum $\mathcal{H}$ und eine Abbildung $\phi : \mathbb{R}^d \to \mathcal{H}$ mit

$$k(x, y) = \langle \phi(x), \phi(y) \rangle \quad \text{für alle } x, y \in \mathbb{R}^d$$

Damit läßt sich die Entscheidungsfunktion darstellen durch

$$F(x) = \text{sgn} \left(\sum_{\mu=1}^{M} \alpha_\mu^* T^\mu k(x^\mu, x) + w_0^* \right) .$$

Die Koeffizienten ergeben sich durch Maximierung von

$$W(\alpha) = \sum_{\mu=1}^{M} \alpha_\mu - \frac{1}{2} \sum_{\nu=1}^{M} \sum_{\mu=1}^{M} \alpha_\nu \alpha_\mu T^\nu T^\mu k(x^\nu, x^\mu)$$

die mit $0 \leq \alpha_\mu \leq C/M$ für alle $\mu = 1, \ldots, M$ erreichen

Beispiele für Mercer Funktionen

1.

$$k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right) \quad \sigma^2 > 2$$

2.

$$k(x,y) = \tanh\left(\alpha\langle x,y\rangle + \theta\right) \quad \alpha,\theta \in \mathbb{R}$$

3.

$$k(x,y) = \left(\langle x,y\rangle + 1\right)^d$$