Kapitel 9

Klassifikation und statistisches Lernen

9.1 Einführung

In diesem Kapitel soll die Fragestellung behandelt werden, wie Ereignisse einer Stichprobe optimal in Klassen eingeteilt werden können. Beispiele für Klassifizierungsprobleme sind die Unterscheidung von Signal und Untergrund in einem Teilchenphysikexperiment (Trigger, Datenselektion), die Zuordnung von Treffern in einem Detektor zu verschiedenen Spuren, die Zuordnung von Pixeln eines Bildes zu einem Buchstaben oder einem Gesicht, die Zuordnung zu 'arm' oder 'reich' ('gesund' oder 'krank') in einer Bevölkerungsstichprobe oder die Zuordnung SPAM oder Nicht-SPAM bei E-Mails.

Formal betrachten wir Ereignisse, die gewisse Eigenschaften oder Merkmale (englisch 'features') haben, nach denen sie klassifiziert werden sollen und die wir in einem Merkmalvektor $\vec{x} = (x_1, x_2, \ldots, x_m)$ zusammenfassen. Die Klasseneinteilung wird im Allgemeinen schwieriger mit wachsender Dimension m des Merkmalraums (deshalb versucht man häufig als ersten Schritt, wenig aussagekräftige oder redundante Variable zu eliminieren). Weitere Erschwernisse ergeben sich, wenn die Ereignisklassen im Merkmalraum überlappen oder sich auf unzusammenhängende Gebiete verteilen. Häufig ist in dem Merkmalraum nicht von vornherein ein 'Abstand' zwischen verschiedenen Merkmalen definiert, so dass man zunächst eine sinnvolle Abstandsmetrik zu definieren hat, um die Merkmale vergleichbar zu machen. In der Regel werden die Merkmale zunächst aufgearbeitet, um die Klassifikation zu erleichtern. Mögliche Maßnahmen sind:

- Normierung der einzelnen Merkmale x_j auf eine Varianz 1 oder ein festes Intervall, zum Beispiel $[0, 1]$;
- Diagonalisieren der Kovarianzmatrix der Merkmale, so dass die transformierten Merkmale (Linearkombinationen der ursprünglichen) unkorreliert sind (Hauptkomponenten-Analyse, 'principle component analysis (PCA)');
- als Verallgemeinerung von PCA die Suche nach Merkmalskombinationen, die besonders signifikante Aussagen machen (Faktorenanalyse);

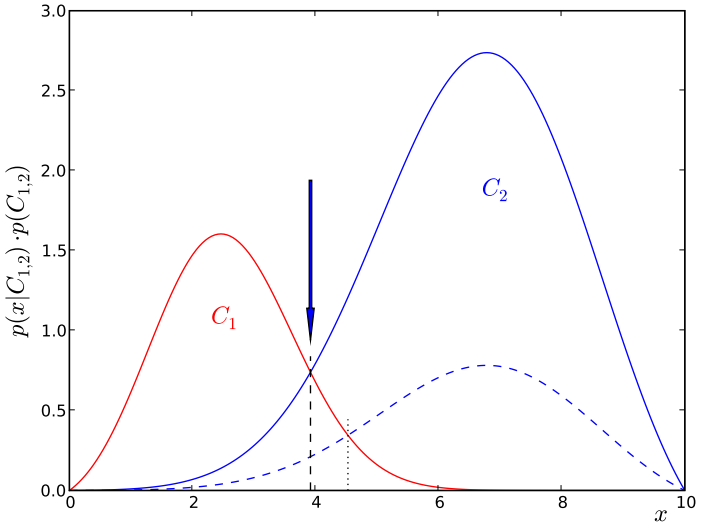


Abbildung 9.1: Wahrscheinlichkeitsdichte für das Merkmal x für zwei Klassen mit unterschiedlichen A-Priori-Wahrscheinlichkeiten. Im Fall $p(C_1) < p(C_2)$ (durchgezogenen Linien) ist die optimale Trennung bei niedrigerem x-Wert als im Fall $p(C_1) > p(C_2)$ (gestrichelte Linie für C_2).)

- Reduktion der Dimensionalität des Merkmalraumes durch Beseitigung redundanter oder unsignifikanter Information (zum Bespiel die Merkmalskombinationen mit den kleinsten Eigenwerten bei PCA).

Bayes-Diskriminante: Ein naheliegendes Klassifizierungsschema ist die Zuordnung eines Ereignisses e_i zu einer Klasse k, wenn die Wahrscheinlichkeit für die Klasse C_k (entsprechend einer 'Hypothese' im vorigen Kapitel) bei gegebenen Merkmalen $\vec{x}_i$ größer ist als für alle anderen Klassen:

$$e_i \to C_k \iff p(C_k|\vec{x}_i) > p(C_j|\vec{x}_i) \quad \forall\, j \neq k. \tag{9.1}$$

Die Wahrscheinlichkeit für eine Klasse ergibt sich wieder aus dem Bayes-Theorem (1.18):

$$p(C_k|\vec{x}_i) = \frac{p(\vec{x}_i|C_k) \cdot p(C_k)}{\sum_{j=1}^{n} p(\vec{x}_i|C_j) \cdot p(C_j)} \tag{9.2}$$

Das Klassifizierungsschema ist anschaulich in Abb. 9.1 anhand nur eines Merkmals x dargestellt: das Merkmal tritt in den zwei betrachteten Klassen normalverteilt mit unterschiedlichen Mittelwerten und Breiten auf. Die Normierungen entsprechen den A-priori-Wahrscheinlichkeiten für die Klassen ($p(C_1),\ p(C_2)$), die in der Abbildung mit zwei unterschiedlichen Verhältnissen angenommen sind. Die Trennung der beiden Klassen nach (9.1) ergibt sich, wo sich die beiden Kurven schneiden.

Das ist natürlich ein besonders einfaches Beispiel, insbesondere wollen wir im Folgenden multi-dimensionale Merkmalsräume betrachten ('multivariate analysis'). In multi-dimensionalen Räumen werden die Klassen durch Hyperflächen getrennt, die durch (9.1) festgelegt werden. Im einfachsten Fall ist die Fläche eine lineare

Funktion, im allgemeinen eine komplizierte Funktion der Merkmale, eventuelle auch nicht zusammenhängend.

Training: Im Allgemeinen werden die Wahrscheinlichkeitsdichten (9.2), auf deren Basis die Klassentrennung erfolgt, nicht bekannt sein. Mit wachsender Dimensionalität wird es auch immer schwieriger, diese Wahrscheinlichkeitsdichten aus Simulationen zu konstruieren, weil zunehmend weniger Ereignisse in ein diskretes Bin fallen. Es sind deshalb Algorithmen entwickelt worden, die Klassentrennung mit Hilfe von Trainigsdatensätzen "lernen" können. Trainiert wird mit simulierten oder auch realen Daten auf eine Ausgabegröße des Algorithmus, die ein Maß für die Zugehörigkeit zu einer Klasse ist. Zum Beispiel kann bei zwei disjunkten Klassen die Ausgabegröße 0 oder 1 sein je nachdem, ob der Merkmalvektor in die Klasse 1 oder 2 gehört. Bei sich überlappenden Verteilungen kann die Ausgabe eine kontinuierliche Zahl sein, die ein Maß für die Wahrscheinlichkeit für eine Klassenzugehörigkeit ist. Das Trainingsergebnis wird mit einem unabhängigen Datensatz getestet, um damit Effizienz und Reinheit der Klassenzuordnung zu bestimmen.

9.2 Schätzung von Wahrscheinlichkeitsdichten

Das Trennungskriterium (9.1) kann man direkt anwenden, wenn man die Wahrscheinlichkeiten $p(C_k|\vec{x})$ in (9.2) als Funktion der Merkmale $\vec{x}$ numerisch zur Verfügung hat. Häufig muss man sich die Wahrscheinlichkeiten aus Simulationen beschaffen. Dazu simuliert man Ereignisse entsprechend der Wahrscheinlichkeitsdichte $p(\vec{x}_i|C_k)$ für jede Klasse C_k und wendet dann das Bayes-Theorem mit den relativen Häufigkeiten der C_k an, um $p(C_k|\vec{x})$ zu bestimmen.

Es gibt verschiedene Methoden aus simulierten, diskreten Ereignissen die Wahrscheinlichdichte zu schätzen:

- Falls eine parametrisierte Modellfunktion bekannt ist, können mit den MC-Ereignissen die Parameter, zum Beispiel durch ML-Anpassung, bestimmt werden.

- Als Modellfunktion kann man auch eine Linearkombination von orthogonalen Funktionen benutzen, zum Beispiel Wavelets.

- Die Dichte wird an jedem Punkt durch Mittelung der Ereignisse über ein Nachbarschaftsvolumen mit vorgebbarer Größe bestimmt.

- Bei der Mittelung kann man die nahen Ereignisse mehr wichten als die weiter entfernten, zum Beispiel durch eine Gauss-Funktion. Die Wichtungsfunktion nennt man 'Kernfunktion' ('kernel funktion') und die Methode 'Kernel Probability Density Estimation (kernel PDE)'.

Im Folgenden wird beispielhaft nur die letzte Methode besprochen.

'Kernel Probability Density Estimation': Gegeben sei eine Stichprobe $\vec{x}_i$ ($i = 1, ..., N$). Die Wahrscheinlichkeitsdichte an einem Punkt $\vec{x}$ wird abgeschätzt durch:

$$\hat{p}(\vec{x}) = \frac{1}{Nh^m} \sum_{i=1}^{N} K\left(\frac{\vec{x} - \vec{x}_i}{h}\right). \tag{9.3}$$

Dabei ist K die Kern-Funktion, h ein Parameter, der die Reichweite der Mittelung bestimmt, und m ist die Dimension von $\vec{x}$. Der Reichweiteparameter h muss so gewählt werden, dass genügend Ereignisse in der Nachbarschaft liegen. Als mögliche Wahl findet sich zum Beispiel in der Literatur $h = N^{-1/(m+4)}$ (man beachte, dass $V \cdot N^{-1/m}$ der mittlere Abstand zwischen zwei Ereignissen in dem m-dimensionalen Volumen V ist).

Gauss-Kern: Wenn die Kern-Funktion eine Gauss-Funktion ist, kann man auch mögliche Korrelationen der Merkmale mit deren Kovarianzmatrix V einbeziehen, wobei V aus der Simulation geschätzt wird, entweder global für den ganzen Datensatz oder lokal um $\vec{x}$ für die Ereignisse, die wesentlich zu $\hat{p}(\vec{x})$ beitragen. Die Formel für die geschätzte Wahrscheinlichkeitsdichte lautet für den Gauss-Kern:

$$\hat{p}(\vec{x}) = \frac{1}{N\sqrt{2\pi \det V}\, h^m} \sum_{i=1}^{N} \exp\left(\frac{(\vec{x} - \vec{x}_i)^T V^{-1} (\vec{x} - \vec{x}_i)}{2h^2}\right). \tag{9.4}$$

9.3 Lineare Diskriminanten

9.3.1 Klassentrennung durch Hyperebenen

Ein Trennungskriterium wie (9.1) definiert Hyperflächen im Merkmalsraum, die in die verschiedenen Klassen aufteilen. Im einfachsten Fall sind diese Flächen Hyperebenen, die zwei Klassen trennen. Die Hessesche Normalform einer Ebene ist:

$$\vec{n}(\vec{x} - \vec{x}_0) = 0, \tag{9.5}$$

wobei $\vec{n}$ der Normalenvektor auf der Ebene, $\vec{x}$ einen beliebigen Punkt und $\vec{x}_0$ einen bestimmten Punkt auf der Ebene beschreibt (der Differenzvektor $\vec{x} - \vec{x}_0$ liegt in der Ebene, siehe Abb. 9.17).

Wenn der Punkt mit dem Ortsvektor $\vec{x}$ nicht auf der Ebene liegt, ist die Gleichung (9.5) nicht erfüllt und es ergibt sich:

$$\vec{n}(\vec{x} - \vec{x}_0) = d \quad \text{mit} \quad d > 0 \text{ oder } d < 0, \tag{9.6}$$

wobei d der Abstand des durch $\vec{x}$ gegebenen Punktes von der Ebene ist und das Vorzeichen die beiden Hemisphären kennzeichnet. Insbesondere ergibt sich für $\vec{x} = 0$ aus $\vec{n}\vec{x}_0 = -d_0$ der Abstand der Ebene vom Ursprung (mit dem durch die Ebenenorientierung festgelegten Vorzeichen).

Im Folgenden wird ein Festlegung der Ebene eingeführt, die eine optimale Trennung zwischen zwei Klassen ergeben, wenn sich deren Verteilungen im Merkmalsraum annähernd durch Normalverteilungen beschreiben lassen.

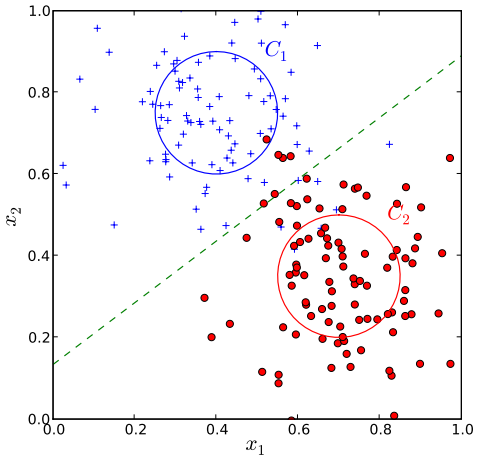


Abbildung 9.2: Stichprobe von Ereignissen mit Merkmalen (x_1, x_2), die aus zwei Klassen gezogen wurden (Kreuze und Kreise). Die Klassenzuordnung kennt man nur für die Trainings- und Testdatensätze. Die Linie zwischen den beiden Anhäufungen ist die Fisher-Diskriminante, die beide Klassen optimal trennt.

9.3.2 Fisher-Diskriminante

Gegeben sei eine Stichprobe von Ereignissen, die zwei Klassen C_1 und C_2 entnommen sind und jeweils durch einen Merkmalvektor $\vec{x}$ gekennzeichnet sind (Abb. 9.2). Die Wahrscheinlichkeitsdichten der Merkmale seinen $f(\vec{x}|C_1)$ und $f(\vec{x}|C_2)$. Wir bilden nun aus einer Linearkombination der Komponenten von $\vec{x}$ eine Testfunktion:

$$t(\vec{x}) = \sum_{j=1}^{m} a_j x_j = \vec{a}^T \vec{x} \tag{9.7}$$

Diese Testfunktion hat unterschiedliche Wahrscheinlichkeitsdichten für die beiden Klassen, die sich durch die Projektion der Ereignisse auf eine Achse senkrecht zur Ebene ergeben (das ist die t-Achse):

$$g(t|C_k), \;\; k = 1, 2. \tag{9.8}$$

Der Koeffizientenvektors $\vec{a}$ soll nun so bestimmt werden, dass die beiden Wahrscheinlichkeitsdichten (9.8) möglichst optimal getrennt sind. Man kann die Testfunktion so interpretieren, dass der Vektor $\vec{a}$ die Orientierung einer Ebene definiert und für den Ortsvektor $\vec{x}_0$ eines Punktes in der Ebene gibt $t(\vec{x}_0)$ den Abstand vom Ursprung an (siehe (9.6)). Durch Anpassung des Koeffizientenvektors $\vec{a}$ und durch Festlegung eines kritischen Wertes t_c der Testfunktion soll nun eine optimale Trennung zwischen zwei Klassen C_1 und C_2 erreicht werden.

Dazu berechnen wir die Erwartungswerte der $\vec{x}$ und die Kovarianzmatrizen für beide Klassen getrennt:

$$\vec{\mu}^{(k)} = \int \vec{x} f(\vec{x}|C_k)\, dx_1 \ldots dx_m, \quad k = 1, 2; \tag{9.9}$$

$$V_{ij}^{(k)} = \int (x_i - \mu_i^{(k)})(x_j - \mu_j^{(k)}) f(\vec{x}|C_k)\, dx_1 \ldots dx_m, \quad k = 1, 2; \;\; i, j = 1, \ldots, m \tag{9.10}$$

In der Regel werden diese Erwartungswerte mit Hilfe von simulierten Datensätzen für beide Klassen geschätzt ('gelernt').

Wegen der linearen Abhängigkeit von t von den Merkmalen, sind die Erwartungswerte von t und deren Varianzen für die beiden Klassen einfach zu berechnen:

$$t_k = \int t\, g(t|C_k)\, dt = \vec{a}^T \vec{\mu}^{(k)} \tag{9.11}$$

$$\sigma_k = \int (t - t_k)^2\, g(t|C_k)\, dt = \vec{a}^T V^{(k)} \vec{a} \tag{9.12}$$

Die Trennungsebene soll jetzt durch Wahl von $\vec{a}$ so gelegt werden, dass der Abstand $|t_1 - t_2|$ möglichst groß wird und die t-Werte möglichst dicht um die Erwartungswerte konzentriert sind, was durch die Varianzen der t_k gegeben ist. Durch Maximierung des χ^2-artigen Ausdrucks

$$J(\vec{a}) = \frac{(t_1 - t_2)^2}{\sigma_1^2 + \sigma_2^2} = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}} \tag{9.13}$$

in Bezug auf $\vec{a}$ ergibt sich die optimale Trennung. Die Matrix B auf der rechten Seite von (9.13) ist die Kovarianzmatrix von $\vec{\mu}^{(1)} - \vec{\mu}^{(2)}$,

$$(t_1 - t_2)^2 = \sum_{i,j=1}^{m} a_i a_j (\mu^{(1)} - \mu^{(2)})_i (\mu^{(1)} - \mu^{(2)})_j = \sum_{i,j=1}^{m} a_i a_j B_{ij} = \vec{a}^T B \vec{a}, \tag{9.14}$$

und die Matrix $W = V^{(1)} + V^{(2)}$, die Summe der Kovarianzmatrizen der beiden Klassen, ergibt sich aus

$$\sigma_1^2 + \sigma_2^2 = \sum_{i,j=1}^{m} a_i a_j (V^{(1)} + V^{(2)})_{ij} = \vec{a}^T W \vec{a}. \tag{9.15}$$

Die Maximierung von $J(\vec{a})$ legt $\vec{a}$ bis auf einen Skalenfaktor fest:

$$\vec{a} \sim W^{-1}(\vec{\mu}^{(1)} - \vec{\mu}^{(2)}) \tag{9.16}$$

Die rechte Seite der Gleichung kann aus Simulationen bestimmt werden. Für die Trennung der beiden Klassen muß noch ein kritischer Wert t_c der Testfunktion festgelegt werden, so dass die Klassenzugehörigkeit nach $t < t_c$ oder $t > t_c$ entschieden wird. Das Kriterium für die Wahl von t_c sind Effizienz und Reinheit der klassifizierten Ereignisse.

9.4 Neuronale Netze zur Datenklassifikation

9.4.1 Einleitung: Neuronale Modelle

Die Entwicklung der Neuroinformatik hat seit Beginn der 80er Jahre einen großen Aufschwung erfahren. Der wesentliche Grund dafür ist sicherlich die große Leistungssteigerung bei den Computern. Damit wurden Computersimulationen von komplexeren Gehirnmodellen und künstlichen neuronalen Netzen (KNN) erst möglich. Dagegen gehen die ersten aussagekräftigen Theorien über die Informationsverarbeitung im Gehirn und den Nervenzellen bis in die 40er Jahre zurück.

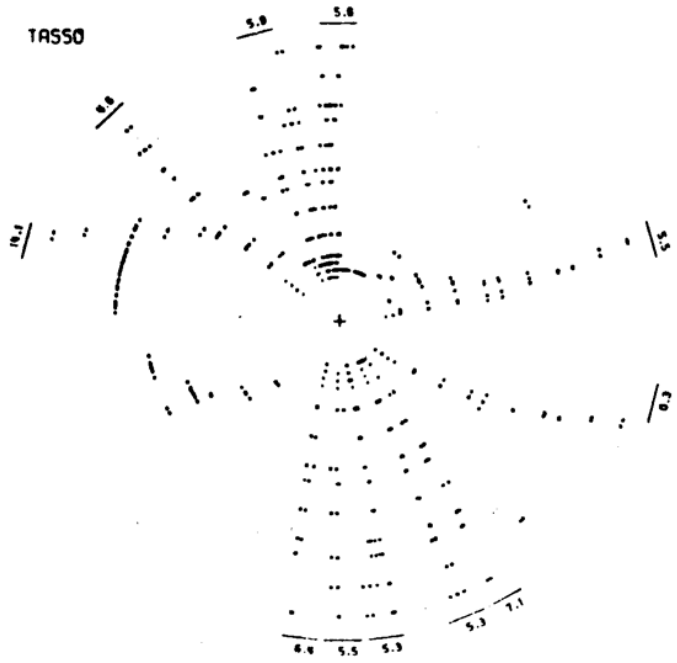


Abbildung 9.3: Hit-Muster, die von Teilchenspuren in einer Driftkammer (TASSO-Experiment) hinterlassen wurden.

Es ist offensichtlich, dass von-Neumann-Computer bei kognitiven Aufgaben (Hören, Sehen, Mustererkennen, etc.) und bei unvollständiger, inkonsistenter oder verrauschter Information im Vergleich zum Gehirn versagen. Das Hit-Muster, das zum Beispiel Teilchenspuren in einer Driftkammer hinterlassen (Abb. 9.3), hat unser Auge 'momentan', innerhalb O(0.1s), als stetig aufeinanderfolgende Punkte erkannt und miteinander verbunden. Der Zeitbedarf eines Computers ist nur dank seiner sehr viel größeren Geschwindigkeit pro einzelnem Rechenschritt vergleichbar. Mit künstlichen neuronalen Netzen könnte dieselbe Leistung innerhalb von O(μs) erzielt werden.

Gehirn-Architektur: Die charakteristischen Merkmale der Datenverarbeitung im Gehirn machen den Unterschied zu dem heutigen Standard für Computerarchitekturen klar:

- sehr viele parallele Prozessoren, O(10^{11}), insgesamt kompakt, geringer Energieverbrauch;
- langsame Einzelschritte, O(ms);
- massiv parallele Verarbeitung (O(10^{13}) Synapsen);
- keine Hardware-Software-, Algorithmen-Daten-Trennung;
- lernfähig:
  - evolutionäres, dynamisches Lernen gibt hohe Flexibilität für die Informationsverarbeitung,

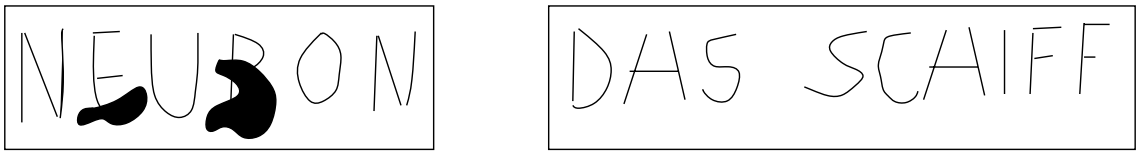


Abbildung 9.4: Beispiele für Fehlertoleranz und Ausgleich von Ungenauigkeiten im Gehirn: auf der linken Seite ist die Information verstümmelt; rechts wird exakt das gleiche Symbol einmal als 'A' und dann als 'H' im Zusammenhang richtig erkannt.

  - evolutionäre Selbstorganisation gibt dem Netz eine gewisse Plastizität zur Anpassung an Neues;

- fehlertolerant (Abb. 9.4), Information kann bis zu einem gewissen Grade
  - unvollständig,
  - inkonsistent,
  - verrauscht sein;
- Stärke: schnelle Erfassung komplexer Zusammenhänge, kognitive Aufgaben, Mustererkennung, assoziative Verknüpfungen.

Literatur zu Neuronalen Netzen: Einführende Literatur zu neuronalen Netzen findet man unter [5, 6, 7, 8, 9, 10, 11, 12]. Siehe auch Spektrum der Wissenschaft, Nov. 79 und Nov. 92, beide Hefte sind dem Gehirn gewidmet [13, 14].

9.4.2 Natürliche neuronale Netze

Die intellektuellen Leistungen werden in der Hirnrinde (Neokortex) erzielt (Fläche etwa 0.2 m^2, Dicke 2-3 mm). Die Hirnrinde ist in Felder für verschiedene Teilaufgaben organisiert (zum Beispiel visuelle, motorische, somatosensorische, Assoziations-Felder).

Ein Schnitt durch die Hirnrinde zeigt ein vertikal ausgerichtetes Netz von Neuronen (Nervenzellen) mit ihren Verzweigungen (Abb. 9.5). In einer vertikalen Säule von $1\,mm^2$ befinden sich etwa 10^5 Neuronen, insgesamt gibt es etwa 10^{11} Neuronen im Gehirn.

Aufbau und Funktion der Neuronen:

Es gibt viele unterschiedliche Neuron-Typen. Um die uns interessierenden wesentlichen Eigenschaften von Neuronen zu beleuchten, konzentrieren wir uns auf die schematische Darstellung eines typischen Neurons in Abb. 9.6. Solch ein Neuron besteht aus

- dem Zellkörper, Durchmesser 5-80 μm,
- den Dendriten, die sich zu Dendritenbäumen mit einer Reichweite von 0.01-3 mm verzweigen,

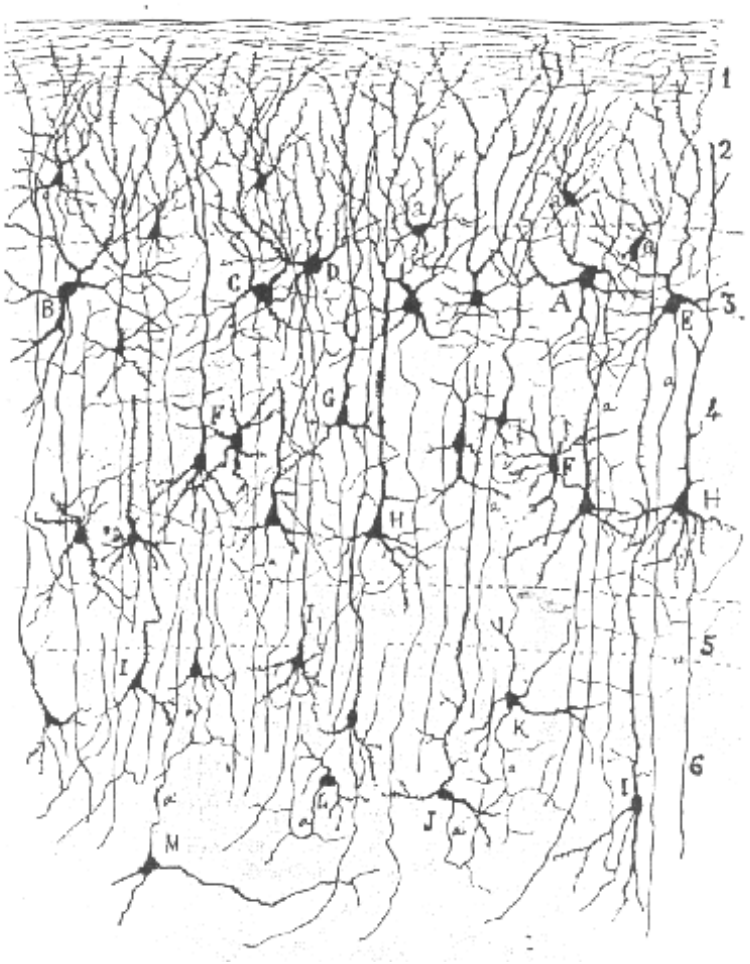


Abbildung 9.5: Vertikaler Schnitt durch die Hirnrinde. Die Dichte der Neuronen ist um einen Faktor 100 untersetzt

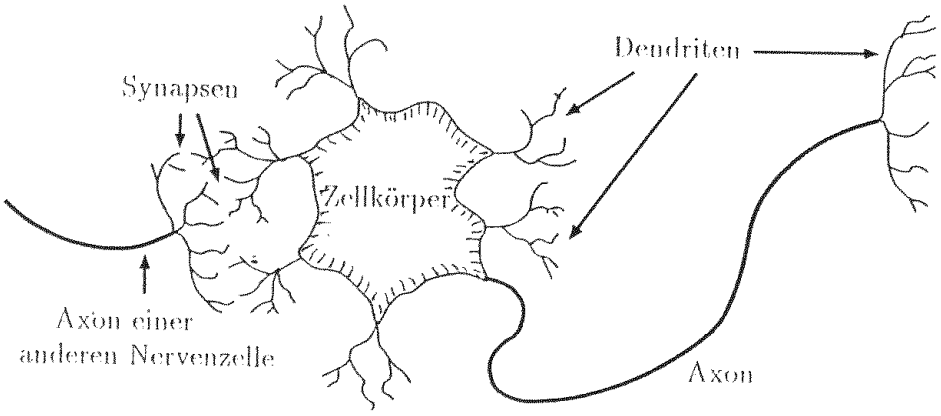


Abbildung 9.6: Schematische Darstellung eines Neurons.

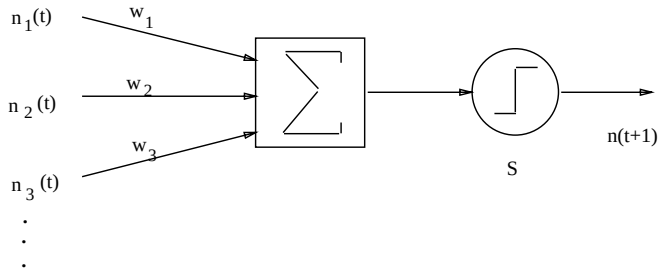


Abbildung 9.7: Neuron als logisches Schaltelement

- den Axons, die bis zu 1 m lang sein können.

Funktionsweise eines Neurons:

- Die Dendriten sammeln in einer Umgebung bis zu etwa 400 μm Signale von benachbarten Neuronen oder von den Axonen weiter entfernter Neuronen.
- Die Signalübertragung auf die Dendriten oder direkt auf den Zellkörper erfolgt über chemische Kontakte (Neurotransmitter) an den Synapsen innerhalb von O(1 ms). In der Hirnrinde hat jedes Neuron O(10^3) Synapsen (allgemein im Gehirn O(1) bis O(10^5)). Die Zeitskala für die Übertragung ist 1 ms, d.h. dass zum Beispiel die visuelle Erkennung eines Bildes mit nicht mehr als O(10) seriellen Schritten erfolgen muß.
- Das Summensignal aller Dendriten verändert das elektrische Potential des Zellkörpers.
- Bei Überschreiten einer Schwelle erzeugt diese Potentialänderung einen Nadelpuls (Spike) auf dem Axon (Signalgeschwindigkeit etwa 10 m/s).

Einfaches Modell: das McCulloch-Pitts-Neuron: Abbildung 9.7 zeigt das McCulloch-Pitts-Neuron, das einem logischen Schaltelement entspricht. Die binären Eingangssignale n_i erzeugen ein binäres Ausgangssignal n (n_i, $n = 0$ oder 1) nach der Vorschrift:

$$n(t+1) \;=\; \Theta\left(\sum_j w_j n_j(t) - s\right) \tag{9.17}$$

Dabei ist t eine diskrete Zeitvariable. Die Heaviside-Funktion ist definiert als:

$$\Theta(x) = \begin{array}{ll} 1 & x \geq 0 \\ 0 & sonst \end{array}$$

Die Gewichte w_i entsprechen den Synapsenstärken, s ist der Schwellenwert. Das Neuron 'feuert' also, wenn die gewichtete Summe der Eingangssignale die Schwelle s überschreitet. Die Gewichte können > 0 (erregend) oder < 0 (hemmend) sein, wie es auch tatsächlich für Synapsen beobachtet wird.

Neuronale Vernetzung: Wesentlich für die Funktion des Gehirns ist das kollektive Verhalten eines Systems von nichtlinear gekoppelten Neuronen. Im Beispiel Abb. 9.8 werden die Eingangsreize x_i (zum Beispiel visuelle Signale) in Ausgangssignale y_i (zum Beispiel zur Bewegung eines Muskels) transformiert.

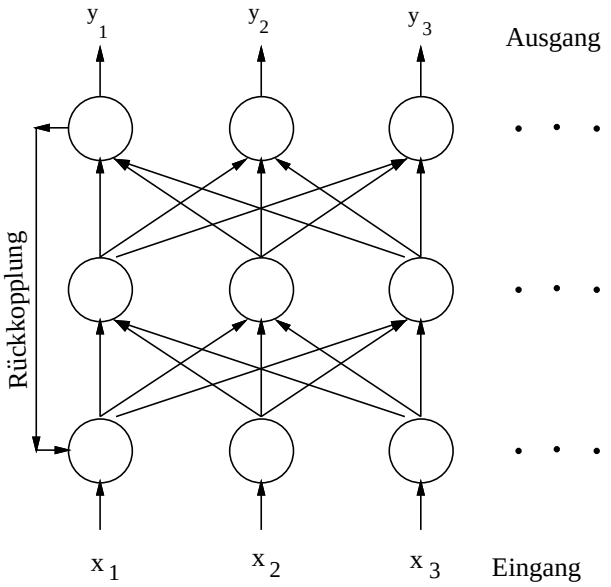


Abbildung 9.8: Beispiel für ein neuronales Netz.

Lernen und Selbstorganisation:

Aus eigener Erfahrung wissen wir, dass das Gedächtnis auf unterschiedlichen Zeitskalen arbeitet. Manches ist bereits nach Sekunden verpflogen, wie die dauernd einwirkenden sensorischen Reize, anderes behalten wir für Minuten oder Tage oder Jahre. Das Behalten im Gedächtnis ist also ähnlich einem evolutionärem Prozess. Generell scheint zu gelten, dass die Stärke und Häufigkeit eines Reizes das Lernen wesentlich beeinflußt. Man beachte, dass wir zum Lernen offensichtlich in der Regel nicht zu wissen brauchen, ob das Gelernte richtig ist ('Lernen ohne Lehrer').

Auf diese Beobachtungen ist die Lernregel von Hebb begründet: Die Synapsenstärke ändert sich proportional zu der Korrelation zwischen prä- und postsynaptischem Signal:

$$\Delta w_i = \eta \cdot y(x_i) \cdot x_i, \text{ mit } 0 < \eta < 1 \tag{9.18}$$

Der Lernparameter η legt die Lerngeschwingigkeit fest. Es ist ein besonders empfindlicher Parameter: einerseits möchte man schnell lernen, andererseits birgt zu schnelles Lernen die Gefahr, dass zuviel Unsinn abgespeichert wird.

Strukturbildung: Mit den etwa 10^{13} Synapsen ergeben sich etwa $10^{10^{14}}$ mögliche Konfigurationen des Gehirns. Das kann nicht alles genetisch festgelegt sein! Genetisch kodiert sind wahrscheinlich nur Organisationsschemata und ein Strukturbildungsmechanismus. Die Verbindungen zwischen den Neuronen werden zum Teil evolutionär aufgrund sensorischer Reize gebildet und können meistens auch später noch verändert werden.

Topographische Abbildungen: Der Lernvorgang führt offensichtlich zu Strukturen im Gehirn, die vorgegebene topographische Zusammenhänge bei den einlaufenden Sinnesreizen intakt lassen. Beispielsweise wird im somatosensorischen Kortex der Tastsinn der Hautoberfläche so abgebildet, dass benachbarte Körperbereiche

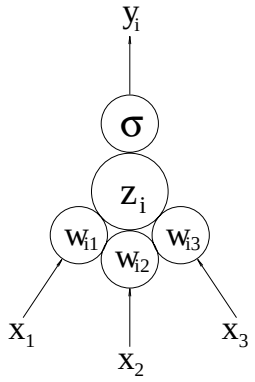


Abbildung 9.9: Struktur eines künstlichen Neurons

benachbart bleiben. Eine wesentliche Eigenschaft der Abbildung ist die Anpassung der Größe der Bildbereiche entsprechend der Wichtigkeit und das jeweils benötigte Auflösungsvermögen.

9.4.3 Künstliche neuronale Netze (KNN)

Künstliche neuronale Netze und neuronale Algorithmen sind in den letzten Jahren intensiv theoretisch untersucht, auf Computern simuliert und – seltener – als Hardware realisiert worden. Bei der Entwicklung von NN-Modellen wird man sich natürlich von den biologischen Befunden inspirieren lassen. Für die Anwendung ist es aber nicht wichtig, ob ein Modell tatsächlich in der Natur realisiert wird. Hier ist der praktische Erfolg ausschlaggebend.

Ausgehend von den im vorigen Abschnitt entwickelten Vorstellungen über natürliche neuronale Netze definieren wir im folgenden, welches die gemeinsamen Elemente der KNN-Modelle sein sollen. Diese Aufstellung ist nicht strikt, sondern soll eher eine Orientierung sein.

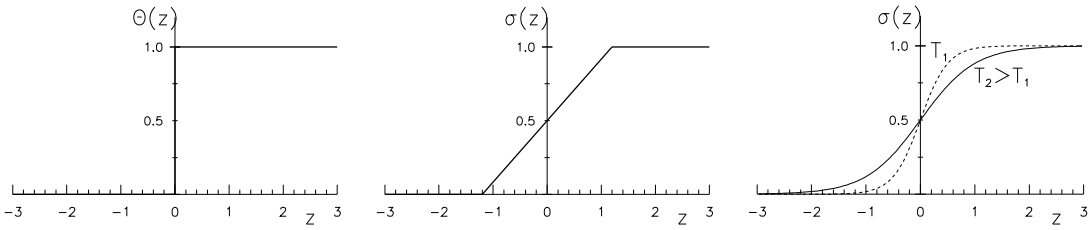
- Prozessorelement: (formales) Neuron, Netzwerk-Knoten (Abb. 9.9).
- Eingabeaktivitäten x_j (Signale auf den Dendriten) sind reelle Zahlen (oder Spannungen, Ströme), eventuell binär (-1,1) oder (0,1).
- Gewichte (entspricht den Synapsen) w_{ij}, > 0 (erregend), < 0 (hemmend)
- Aktivitätsfunktion, zum Beispiel:

$$z_i = \sum_j w_{ij} x_j - s_i$$

- Ausgabefunktion (oder Transferfunktion) g:

$$y_i = g(z_i)$$

  I.a. liegt y_i im Intervall [-1,1] oder [0,1] und hat häufig ein Schwellwertverhalten mit Sättigung an den Intervallgrenzen. Neben der Θ-Funktion werden häufig



Stufe linear sigmoid

Abbildung 9.10: Beispiele von Schwellenfunktionen

folgende 'sigmoide' Funktionen gewählt (Abb. 9.10):

$$\sigma(z) = \frac{1}{1+e^{-z/T}} \tag{9.19}$$

$$\sigma(z) = \tanh(z/T) \tag{9.20}$$

$$\sigma(z) = 1/2(1+\tanh(z/T)) \tag{9.21}$$

Die Funktionen (9.19) und (9.21) haben Werte im Intervall [0,1] und die Funktion (9.20) im Intervall [-1,1]. Sigmoide Funktionen haben den Vorteil im Bereich der Schwelle differenzierbar zu sein. Die 'Temperatur' T bestimmt den Bereich mit variabler Verstärkung:

Für T$\rightarrow$ 0 geht σ in die Θ-Funktion über (binäres Neuron).

T groß: weiche Entscheidung.

- Netzwerk-Architektur: Netzwerk mit Knoten und Verbindungen
  - 'jeder mit jedem'
  - Nachbarschaftsverknüpfung
  - uni- oder bi-direktional
  - Schicht-Struktur mit hierarchischer Anordnung (zum Beispiel feed-forward)
  - mit oder ohne Rückkopplung
  - ...
- Lernen:
  - Anpassung der Gewichte
  - Anpassung der Architektur: Erzeugen und Löschen von Neuronen und Verbindungen
- Lernregel:
  - selbständig (ohne Lehrer, unsupervised), zum Beispiel Hebb-Regel
  - angeleitet (mit Lehrer, supervised) Vergleich des Netzwerk-Outputs mit der (vom Lehrer vorgegebenen) Erwartung, Anpassung durch Fehlerminimierung (zum Beispiel Backpropagation- Algorithmus).

- Update-Regel: Neubestimmung eines Netzzustandes kann synchron, sequentiell oder iterativ (wegen nichtlinearer Kopplungen) gemacht werden.
- Netzwerk-Phasen:
  - Trainingsphase (Verwendung eines Trainings-Datensatzes)
  - Generalisierungsphase (Anwendung auf unbekannte Daten)

Feed-Forward-Netzwerke

In dieser Vorlesung wollen wir uns auf sogenannte Feed-Forward-Netzwerke beschränken, in denen die Neuronen geschichtet angeordnet sind und die Verbindungen streng nur in eine Richtung, jeweils zur nächsthöheren Schicht, von der Eingabeschicht bis zur Ausgabeschicht laufen (Abb. 9.8, ohne Rückkopplung). Feed-Forward-Netze (FFN) werden häufig zur

- Lösung von Klassifikationsaufgaben,
- Mustererkennung und
- Funktionsapproximation

benutzt. Für praktische Anwendungen sind sie wahrscheinlich der wichtigste Netzwerktyp. Ihre Bedeutung haben FFN wohl durch die von herkömmlichen Computern gut ausführbaren, im Prinzip sequentiellen, Algorithmen und insbesondere die Backpropagation-Lernvorschrift erhalten.

Das einfachste Beispiel ist das (einfache) Perzeptron mit nur einer Eingangsschicht und einer Ausgangsschicht. Mit Computersimulationen konnte gezeigt werden, dass ein Perzeptron 'intelligenter' Leistungen fähig ist: Es kann angebotene Muster unterscheiden und kann diese Musterklassifizierung mit Hilfe eines Lehrers lernen (supervised learning).

9.4.4 Das einfache Perzeptron

Definition und Eigenschaften des Perzeptrons:

Abbildung 9.11 zeigt das einfache Perzeptron mit einer Eingangsschicht (oder -lage) und einer Ausgangsschicht (wir ordnen den Eingängen eine Schicht zu, ist manchmal auch anders definiert). Jeder der k Eingänge ist mit jedem der l Ausgänge verbunden, den Verbindungen werden die Gewichte w_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, l)$ zugeordnet. Die Eingänge $x_1,\ x_2, \ldots, x_k$ lassen sich in einem 'Mustervektor' $\vec{x}$ zusammenfassen, der einen Punkt im 'Musterraum' (pattern space) darstellt. Die einzelnen Komponenten sind 'Merkmale' (features). Über die folgende Vorschrift wird einem Mustervektor $\vec{x}$ ein Ausgabevektor $\vec{y}$ zugeordnet:

$$y_i = g\left(\sum_j w_{ij} x_j\right) = g(\vec{w}_i \vec{x}) \tag{9.22}$$

Im letzten Teil wurden die Gewichte zu einem Ausgangsknoten i zu einem Vektor zusammengefaßt. Die Transferfunktion g ist gewöhnlich eine sigmoide Funktion (ursprünglich beim Perzeptron einfach die Θ-Funktion, wir wollen uns hier nicht darauf

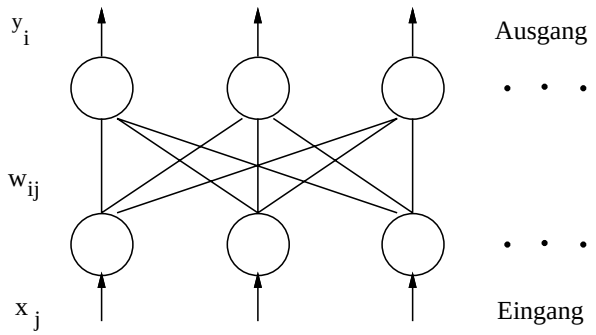


Abbildung 9.11: Perzeptron-Netzwerk

beschränken). In Gl. (9.22) kommen keine expliziten Schwellen s_i vor wie in der Formel (9.17) für das McCulloch-Pitts-Neuron. Schwellen können durch eine zusätzliche konstante Eingabe $x_0 = 1$ und die Gewichte $w_{i0} = -s_i$ berücksichtigt werden.

Beispiel: Darstellung der Boolschen Funktionen AND und OR: Wir wollen hier binäre Ein-und Ausgabegrößen betrachten mit Werten 0 und 1. Dann muß die Transferfunktion die Θ-Funktion sein, $g = \Theta$. Im folgenden wird gezeigt, dass sich die Funktionen AND und OR entsprechend der Wahrheitstafel in Abb. 9.12 durch ein Netz mit den 2 Eingängen x_1 und x_2 und einem Ausgang y realisieren lassen ('Ja-Nein-Maschine').

Wir wollen an dieser Stelle zunächst nicht der Frage nachgehen, wie das Netz die richtigen Antworten lernt; das wird dann allgemeiner für mehrschichtige FFN gezeigt (siehe Abschnitt 9.4.6). Man kann sich aber leicht davon überzeugen, dass die Gewichte

$$\text{AND}: \ (\mathrm{w}_0, \mathrm{w}_1, \mathrm{w}_2) \ = \ (-1.5, 1.0, 1.0)$$

$$\text{OR} \ : \ (\mathrm{w}_0, \mathrm{w}_1, \mathrm{w}_2) \ = \ (-0.5, 1.0, 1.0)$$

das Problem lösen (Abb. 9.12). Die Bedeutung dieses Resultates ist sehr anschaulich: Nach Gl. (9.22) wird der Raum der Muster (x_1, x_2) in 2 Klassen geteilt, die der Bedingung

$$\vec{w}\vec{x} \ < \ 0 \text{ bzw. } \vec{w}\vec{x} \ < \ 0$$

genügen. Die Trennung zwischen beiden Klassen

$$\vec{w}\vec{x} \ = \ 0$$

definiert eine Hyperebene im Musterraum, auf der der Vektor $\vec{w}$ senkrecht steht. In unserem Fall sind die Hyperebenen Geraden, die folgenden Gleichungen genügen:

$$\text{AND}: x_1 + x_2 = 1.5$$

$$\text{OR} \ : x_1 + x_2 = 0.5$$

Abbildung 9.12 zeigt die Lage der Geraden in dem Musterraum.

Allgemein gilt, dass durch Gl. (9.22) für jeden Ausgabeknoten eines Perzeptrons eine Hyperebene definiert wird, die jeweils den Musterraum in zwei Klassen einteilt.

x_1	x_2	$y(AND)$	$y(OR)$	$w_1x_1 + w_2x_2$
0	0	0	0	0
1	0	0	1	1
0	1	0	1	1
1	1	1	1	2

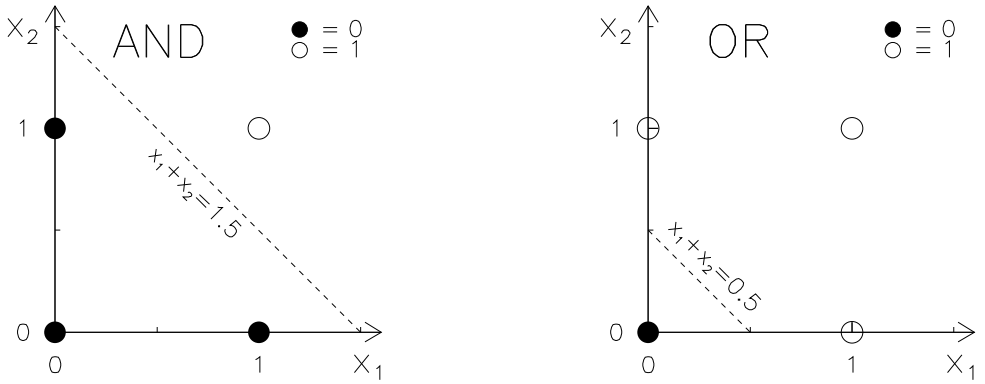


Abbildung 9.12: Oben: Wahrheitstafel für die Boolschen Funktionen AND und OR zusammen mit der Summe der gewichteten Eingänge wie vom Perzeptron berechnet. Unten: Klasseneinteilung im Musterraum für das AND- und OR-Problem. Die gestrichelten Geraden geben die von dem Perzeptron jeweils gefundene Klassentrennung an.

x_1	x_2	$y(XOR)$	$\vec{w}\vec{x}$
0	0	0	$w_0 < 0$
1	0	1	$w_0 + w_1 > 0$
0	1	1	$w_0 + w_2 > 0$
1	1	0	$w_0 + w_1 + w_2 < 0$

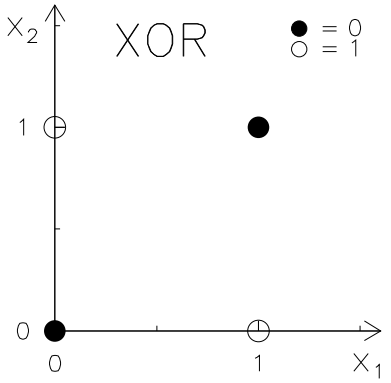


Abbildung 9.13: Links: Wahrheitstafel für die Boolschen Funktionen XOR zusammen mit den Bedingungen an die Gewichte. Rechts: Klasseneinteilung im Musterraum für das XOR-Problem.

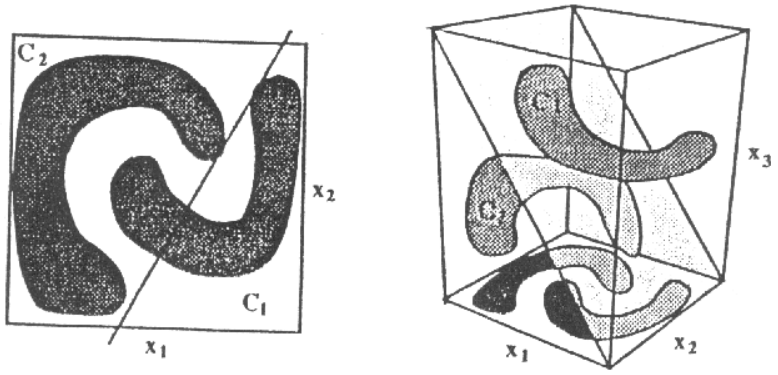


Abbildung 9.14: Lineare Separierbarkeit: a) in 2 Dimensionen nicht separierbar, b) in 3 Dimensionen separierbar.

Die Trennung ist scharf für $g = \Theta$, was für eine Klasse $y = 0$ und für die andere $y = 1$ liefert. Bei einer sigmoiden Funktion ist die Ausgangsaktivität y ein (im Allgemeinen nichtlineares) Maß für den Abstand von der Hyperebene, solange man sich noch so nahe an der Hyperebene befindet, dass g noch nicht in Sättigung ist.

Limitierung des einfachen Perzeptrons:

Aus der vorangehenden Diskussion ergibt sich sofort, dass ein Perzeptron nur dann Muster in Klassen einteilen kann, wenn diese durch eine Hyperebene zu trennen sind. Man sagt in diesem Fall: die Klassen sind 'linear separierbar'; die Hyperebenen werden 'lineare Diskriminanten' genannt (siehe Abschnitt 9.3). Ein bekanntes, einfaches Beispiel, bei dem das einfache Perzeptron keine Lösung findet, ist die XOR-Funktion (Exclusive-OR) definiert in der Tabelle in Abb. 9.13. Man erkennt sofort, dass die Bedingungen an die Gewichte nicht gleichzeitig erfüllt werden können. Das entspricht der Tatsache, dass in Abb. 9.13 keine Gerade gefunden werden kann, die die $(y = 0)$- von der $(y = 1)$-Klasse trennt.

Ein anderes Beispiel von nicht linear separierbaren Punktemengen ist in Abb. 9.14a gezeigt. In solchen Fällen kann man eventuell doch noch eine Perzeptron-Lösung finden, wenn man ein weiteres Merkmal findet, dass die Klassen diskriminiert. Die trennende Hyperebene läge dann in einem um eine Dimension erweiterten Raum (Abb. 9.14b). Das Problem ließe sich auch mit Hilfe komplizierterer Transferfunktionen lösen, was aber dem grundlegenden Konzept für neuronale Netze (möglichst einfache Einzelschritte) widerspräche.

Eine allgemein anwendbare Lösung findet man durch Erweiterung des Perzeptron-Modells auf mehrschichtige Netze.

9.4.5 Das Mehrlagen-Perzeptron

Lösung des XOR-Problems:

Wir haben gesehen, dass ein einfaches Perzeptron durch

$$\vec{w}\vec{x} \;=\; 0 \tag{9.23}$$

Hyperebenen im Musterraum definiert, die den Raum in die beiden Klassen

$$\begin{array}{lll} \vec{w}\vec{x} & < & 0 \qquad \text{Klasse 1} \\ \vec{w}\vec{x} & > & 0 \qquad \text{Klasse 2} \end{array} \tag{9.24}$$

unterteilt. Mit der Kombination von Hyperebenen lassen sich offensichtlich Volumina im Musterraum definieren. Eine solche Kombination gelingt tatsächlich durch die Erweiterung des einfachen Perzeptrons um eine (oder mehrere) Lagen. Dieses Mehrlagen-Perzeptron hat dann neben den Eingangs- und Ausgangslagen auch versteckte Lagen (hidden layers).

Bei dem XOR-Problem (Abb. 9.13) sehen wir, dass die 1-Klasse zwischen den beiden für das AND und das OR gefundenen Hyperebenen (Abb. 9.12) liegt. Das liegt natürlich daran, dass sich das XOR aus einer entsprechenden AND-OR-Kombination ergibt:

$$y(XOR) \;=\; \overline{y(AND)} \;\wedge\; y(OR).$$

Wir definieren also ein dreilagiges Netz mit 2 Knoten in der Eingangslage, 2 Knoten in der versteckten Lage, 1 Knoten in der Ausgangslage (Netz-Konfiguration: 2 - 2 - 1). Die Aktivitäten der Knoten und die Gewichte sind:

$\vec{x}$: Eingangsaktivitäten,

$\vec{x}'$: Aktivitäten der versteckten Knoten,

y : Ausgangsaktivität (im Allgemeinen auch ein Vektor),

$\vec{w}_i$: Gewichte für die Eingänge ($i = 1, 2$ ist der Index der versteckten Knoten),

$\vec{w}'$: Gewichte für die Ausgänge $\vec{x}'$ der versteckten Knoten.

In Abb. 9.15 sind an die Netz-Verbindungen die Gewichte w_{i1}, w_{i2} bzw. w'_1, w'_2 und an die Knoten die Schwellen $-w_{i0}$ bzw. $-w'_0$ geschrieben. Mit der Tabelle sieht man, dass in diesem Netz die beiden versteckte Knoten jeweils das AND und OR realisieren und die Ausgangslage die logische Verknüpfung von beiden. Die 1-Klasse des Netzes liegt also zwischen den beiden Geraden in Abb. 9.15b, die 0-Klasse außerhalb.

Für das Anlernen von Netzen ist es wichtig zu sehen, dass die Lösungen für die Klassenseparation nicht eindeutig sind. In unserem Beispiel gibt es eine unendliche Schar von Hyperebenen, die kontinuierlich durch Translationen und Rotationen auseinanderhervorgehen und die, solange sie nicht einen der Musterpunkte überspringen, dasselbe leisten. Problematischer für die Kontrolle des Lernens ist allerdings, dass es auch Lösungen geben kann, die nicht kontinuierlich zusammenhängen. Für das XOR-Problem finden wir zum Beispiel die in Abb. 9.16 angegebene Lösung, bei der die zwei Hyperebenen diesmal die 0-Klasse einschließen, während die 1-Klasse außerhalb liegt.

Die Hessesche Normalform für die Hyperebenen:

Die Gleichung einer Hyperebene, $\vec{w}\vec{x} = 0$, ist offensichtlich invariant gegenüber einer Transformation

$$\vec{w} \rightarrow -\vec{w} \tag{9.25}$$

x_1	x_2	x'_1	x'_2	$w'_1x'_1 + w'_2x'_2$	y
0	0	0	0	0	0
1	0	0	1	1	1
0	1	0	1	1	1
1	1	1	1	0	0

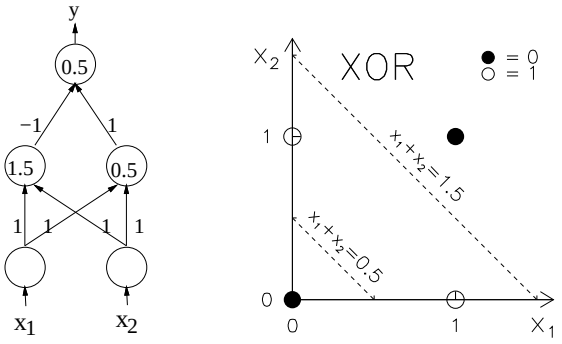


Abbildung 9.15: Links: Wahrheitstafel für das XOR-Netz auf der rechten Seite. Mitte: Netzwerk mit Gewichten und Schwellen zur Lösung des XOR-Problems. Rechts: Musterraum des XOR-Problems mit den durch das Netz bestimmten Hyperebenen.

x_1	x_2	x'_1	x'_2	$w'_1x'_1 + w'_2x'_2$	y
0	0	0	1	-1	0
1	0	0	0	0	1
0	1	1	1	0	1
1	1	0	1	-1	0

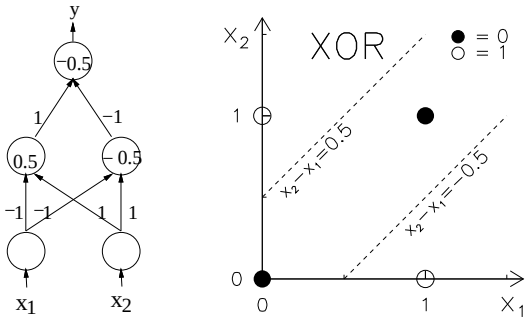


Abbildung 9.16: Links: Wahrheitstafel für das XOR-Netz auf der rechten Seite. Mitte: Netzwerk mit Gewichten und Schwellen zur Lösung des XOR-Problems (alternativ zu Abb. 9.15). Rechts: Musterraum des XOR-Problems mit den durch das Netz bestimmten Hyperebenen.

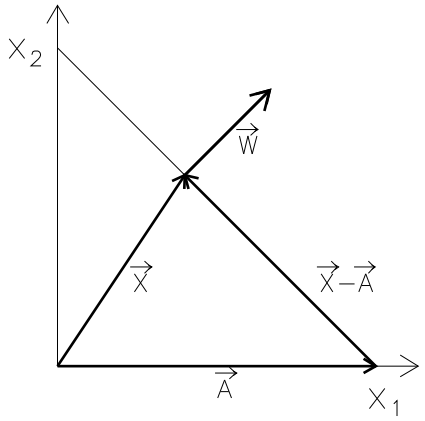


Abbildung 9.17: Zur Darstellung der Hesseschen Normalform der Geradengleichung.

Dasselbe gilt aber nicht für die Klasseneinteilung durch $\vec{w}\vec{x} < 0$ und $\vec{w}\vec{x} > 0$, weil durch (9.25) die Klassen gerade vertauscht werden. Wir wollen uns deshalb die Bedeutung der Orientierung von $\vec{w}$ genauer klar machen.

Für die folgenden Überlegungen wollen wir die Gewichte und Vektoren für einen 2-dimensionalen Musterraum betrachten:

$$\begin{aligned} \vec{X} &= (x_1, x_2) \\ \vec{W} &= (w_1, w_2) \end{aligned}$$

(die großen Buchstaben sollen von den Vektoren $\vec{x}$ und $\vec{w}$ unterscheiden, die ja mit den 0-Komponenten die Schwellen enthalten). Dann ist die Gleichung der Hyperebene:

$$\vec{W}\vec{X} = -w_0,$$

so dass auch für einen festen Ortsvektor $\vec{A}$ eines Punktes auf der Geraden gilt:

$$\vec{W}\vec{A} = -w_0$$

und damit:

$$\vec{W}(\vec{X} - \vec{A}) = 0 \tag{9.26}$$

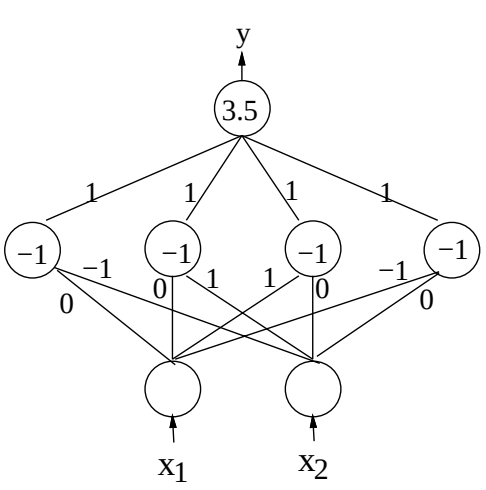
Das heißt, $\vec{W}$ steht senkrecht auf $\vec{X} - \vec{A}$ und damit senkrecht auf der Geraden, weil $\vec{X} - \vec{A}$ die Richtung der Geraden hat (Abb. 9.17). Durch die Wahl des Vorzeichens der Gewichte wird damit eine Orientierung der Normalen auf der Hyperebene festgelegt. Gleichung (9.26) ist die Hessesche Normalform der Geradengleichung (wobei genau genommen $\vec{W}$ zu normieren wäre).

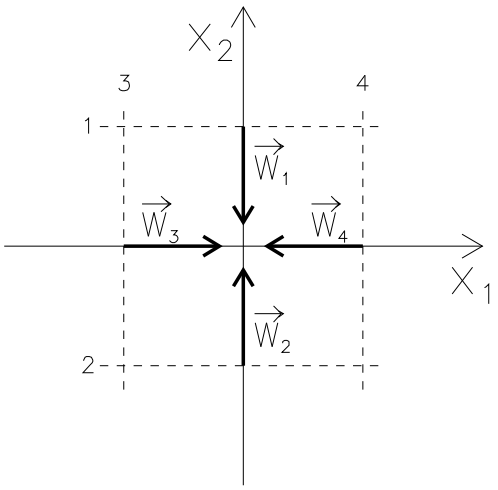
Musterklassifizierung mit einem Dreilagen-Perzeptron:

Die Punkte in dem Quadrat $[-1 < x < +1; -1 < y < +1]$ sollen zur Musterklasse A gehören (Abb. 9.18). Um diese Klasse zu separieren, sind dann 4 verdeckte Knoten notwendig, die jeweils eine Begrenzungsgerade festlegen (siehe Tabelle in Abb. 9.18). Wenn man die Vorzeichen so wählt, dass die Gewichtsvektoren alle in das Volumeninnere zeigen (Abb. 9.18), dann lassen sich die Ausgänge der verdeckten Knoten alle mit positiven Gewichten kombinieren, um die Klasse A zu selektieren.

Θ-Funktion als Übertragungsfunktion: Benutzt man die Θ-Funktion als Übertragungsfunktion dann wird mit den Gewichten und Schwellen in Abb. 9.18 das Quadrat exakt herausgeschnitten.

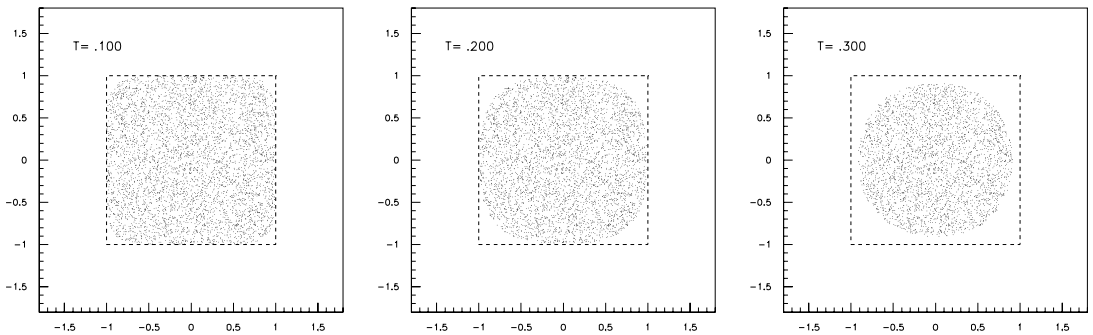
Sigmoide Übertragungsfunktion: Bei Verwendung von sigmoiden Funktionen als Übertragungsfunktion werden in der ersten verdeckten Lage die trennenden Hyperebenen immer noch scharf definiert. Im Gegensatz zu der 0-1-Entscheidung ('links' oder 'rechts' von der Hyperebene) der Θ-Funktion erhält man hier jedoch ein kontinuierliches Maß für den Abstand von der Hyperebene. Erst bei der gewichteten Summe dieser Abstände in der nächsten Stufe spielt die relative Größe der Abstände eine Rolle. In dieser Summe kann nämlich ein kleiner Abstand von einer Hyperebene einen großen Abstand von einer anderen Ebene kompensieren. Das





i	Geraden-Gl.	w_{i0}	w_{i1}	w_{i2}	w'_i
1	$-x_2 + 1 = 0$	1	0	-1	1
2	$x_2 + 1 = 0$	1	0	1	1
3	$x_1 + 1 = 0$	1	1	0	1
4	$-x_1 + 1 = 0$	1	-1	0	1

Abbildung 9.18: Oben: a) Netzwerk mit Gewichten und Schwellen zur Selektion der Punkte innerhalb des in b) gezeigten Quadrates. Unten: Definition der Geraden und Gewichtsvektoren für das Netzwerk in der Abbildung. Der Index i steht sowohl für einen versteckten Knoten als auch für die zu diesem Knoten gehörige Gerade.



a) b) c)

Abbildung 9.19: Durch das Netz in Abb. 9.18 selektierte Punktmenge bei Benutzung einer sigmoiden Schwellenfunktion mit Temperaturparameter a) $T = 0.1$, b) $T = 0.2$, c) $T = 0.3$.

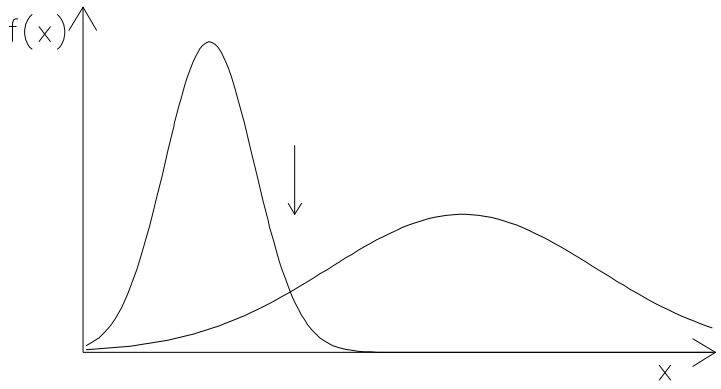


Abbildung 9.20: Beispiel für überlappende Verteilungen im Musterraum.

führt zu Abrundungen von Ecken bei der Klassifikation und erlaubt im Allgemeinen die Konturen des Klassenvolumens besser zu approximieren.

In Abb. 9.19 wird gezeigt, wie sich die Kontur der selektierten Punktmenge verändert, wenn man im obigen Beispiel des Quadrates statt der Θ-Funktion die 'logistische Funktion' (9.19) mit dem Temparaturparameter $T = 1$ benutzt.

An diesem Beispiel läßt sich der Einfluß des Parameters T gut verdeutlichen: Für $T \to 0$ nähert man sich der Θ- Funktion an und damit nähert sich das ausgeschnittene Volumen mehr dem Quadrat; für $T \to \infty$ wird das Volumen abgerundeter. Trotz dieses starken Einflusses ist ein variabler T-Parameter eigentlich überflüssig: die Wirkung von T kann durch geeignete Normierung der Gewichte ebenso erreicht werden (große Gewichte ergeben scharfe Grenzen und umgekehrt). In der Lernphase kann es sich andererseits als nützlich erweisen, mit einem T-Parameter das Lernverhalten zu steuern.

9.4.6 Lernen

Die Lernstrategie:

Für Feed-Forward-Netze sind Lernstrategien entwickelt worden, bei denen das Netz mit Hilfe eines Trainingsdatensatzes lernt, die richtige Antwort zu geben. Während des Trainings kann das Netz seine Antwort mit der richtigen vergleichen; das ist also die Situation 'Lernen mit Lehrer' (supervised learning). Wenn wir Muster in Klassen einteilen wollen, erwarten wir für einen Mustervektor $\vec{x}$ folgende Antworten y_j:

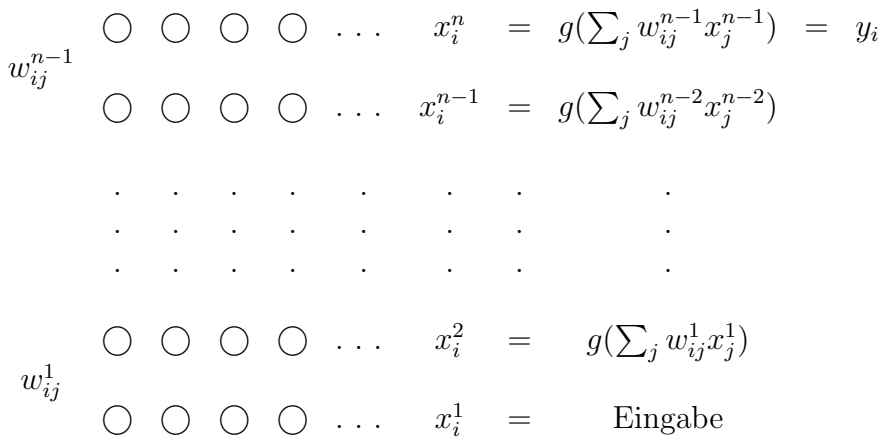
$$\vec{x} \;\to\; y_j \quad \begin{array}{rcl} = & 1 & \text{wenn } \vec{x} \text{ in Klasse } j \\ = & 0 & \text{sonst} \end{array}$$

Dieses Lernziel ist sofort einsichtig, wenn die Klassen disjunkt sind. Wir wollen es aber auch beibehalten, wenn die Klassen sich überlappen wie im Fall der beiden Gauß-Verteilungen in Abb. 9.20. Wenn die Fläche unter den Kurven ein Maß für die Häufigkeit des Auftretens von Mustern der jeweiligen Klasse ist, dann ist die optimale Trennung dort, wo beide Wahrscheinlichkeiten gleich sind, d.h. der Schnittpunkt beider Kurven ('Bayes-Diskriminante'). Wir werden sehen, dass ein wohl-trainiertes Netz diesen optimalen Grenzfall erreichen kann.

Wie gut das Netz gelernt hat, wird mit einem dem Netz unbekannten Datensatz getestet, d.h. man prüft, ob das Netz das Gelernte auf unbekannte Daten übertragen, ob es 'generalisieren' kann.

Lernalgorithmen:

Wir betrachten ein Feed-Forward-Netz mit n Lagen, die Ausgangsaktivitäten der k-ten Lage seien durch den Vektor $\vec{x}^k$ gegeben, die Gewichte zwischen der k-ten Lage und dem i-ten Knoten in der k+1-ten Lage sei $\vec{w}_i^k$. Das Netz hat dann folgende Struktur:



Der Trainingsdatensatz enthalte N Mustervektoren, für jedes Muster p ($p = 1, \ldots, N$) und für jeden Ausgangsknoten i sei die richtige Antwort $\hat{y}_i^{(p)}$ bekannt, die mit der Antwort $y_i^{(p)}$ des Netzes verglichen werden kann. Als Maß für die Optimierung des Netzwerkes definieren wir die Fehlerfunktion (l ist die Zahl der Ausgangsknoten)

$$E = \frac{1}{2} \sum_{p=1}^{N} \sum_{i=1}^{l} (y_i^{(p)} - \hat{y}_i^{(p)})^2 \tag{9.27}$$

Die Fehlerfunktion soll durch Variation der Gewichte w_{ij}^k minimiert werden, es muß also gelten:

$$\frac{\partial E}{\partial w_{ij}^k} = 0 \qquad k = 1, \ldots n-1 \tag{9.28}$$

Da E nicht-linear von den Gewichten abhängt, kann das Gleichungssystem (9.28) im allgemeinen nur iterativ gelöst werden. Wir wählen das für solche Optimierungsprobleme geläufige Gradientenabstiegs-Verfahren (Abb. 9.21) um das (globale) Minimum zu suchen. Es sei hier bemerkt, dass es bei multi-dimensionalen Problemen im allgemeinen sehr schwierig ist, das globale Minimum zu finden. Für unsere Anwendungen ist es aber in der Regel nicht wichtig, ob das Netz tatsächlich das globale Minimum gefunden hat, wenn es nur ein relativ gutes gefunden hat.

Die Fehlerfunktion soll also entlang des negativen Gradienten im Gewichtsraum schrittweise verkleinert werden. Dazu korrigieren wir jedes Gewicht w_{ij}^k entsprechend:

$$\Delta w_{ij}^k = -\eta \frac{\partial E}{\partial w_{ij}^k} \tag{9.29}$$

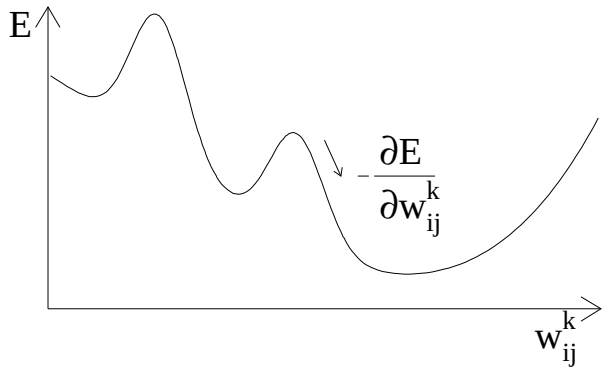


Abbildung 9.21: Beispiel für den Verlauf einer Fehlerfunktion im Gewichtsraum.

Wenn der Lernparameter η genügend klein ist (damit es keine Oszillationen um das Minimum gibt), kann die Korrektur nach jedem angebotenen Muster p erfolgen:

$$\Delta w_{ij}^k \;=\; -\eta \frac{\partial E^{(p)}}{\partial w_{ij}^k}$$

Dann stellt jedes Muster bereits einen Iterationsschritt dar; in der Regel ist dieses Verfahren schneller, als wenn man vor jeder Gewichtskorrektur erst über alle N Muster mittelt. Aus Stabilitätsgründen kann es allerdings manchmal vorteilhaft sein über eine kleine Zahl m von Mustern zu mitteln (m≈10).

Eine effiziente Methode, die Gewichtskorrekturen für die verschiedenen Lagen zu berechnen, ist der Backpropagation-Algorithmus, den wir allerdings hier aus Zeitgründen nicht näher besprechen.

Training:

Im folgenden sollen einige Begriffe, die beim Training von FF-Netzen auftreten, erläutert werden:

Trainingsdatensatz: Der Trainingsdatensatz enthält N Muster, die jeweils den Eingabevektor $\vec{x}^{(p)}$ und die erwartete Antwort $\vec{y}^{(p)}$ enthalten:

$$(\vec{x}^{(p)}, \hat{\vec{y}}^{(p)}),\ p = 1, \ldots, N \tag{9.30}$$

Lernzyklen: Im allgemeinen muß das Lernen relativ langsam erfolgen ($\eta < 1$), damit das Minimum sicher gefunden werden kann. Um zum Minimum zu kommen, muß der Trainingsdatensatz in der Regel wiederholt dargeboten werden (Lernzyklen).

Konvergenzkontrolle: Die **Konvergenz** des Verfahrens wird nach jedem Zyklus (oder nach q Zyklen) getestet durch Auswertung der Fehlerfunktion E (oder meistens E/N) oder der **Effizienz** der Selektion für jede Klasse i:

$$\epsilon_i \;=\; \frac{N_i^{net}}{N_i^{in}} \tag{9.31}$$

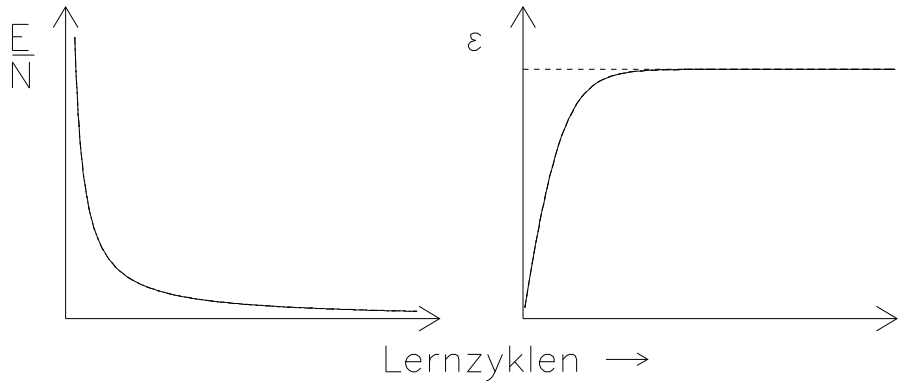


Abbildung 9.22: Kontrolle der Konvergenz: typische Verläufe der Fehlerfunktion (links) und der Effizienz (rechts).

Dabei ist N_i^{net} die Anzahl der Muster, die vom Netz richtig in die i-te Klasse eingeordnet werden, und N_i^{in} die Anzahl der dem Netz angebotenen Muster der Klasse i. Die Effizienz sollte in einen Sättigungswert übergehen, der je nach Überlapp der Klassen zwischen 50% und 100% liegen sollte (100% kann nur für disjunkte Klassen erwartet werden). Abbildung 9.22 zeigt das erwartete Verhalten der Fehlerfunktion und der Effizienz.

Generalisierung: Die Bewährungsprobe für ein Netz ist schließlich der Nachweis, dass es das Gelernte auf einen ihm unbekannten Testdatensatz anwenden kann. Geprüft wird auch hier die Fehlerfunktion und die Effizienzen für die verschiedenen Klassen. Im allgemeinen sind die Effizienzen etwas niedriger und die Fehlerfunktion etwas größer als für die Trainingsdaten. Bei zu großer Diskrepanz ist zu prüfen, ob das Netz durch 'Overtraining' zu stark an die Trainingsdaten angepaßt ist. Das ist dann auch ein Hinweis, dass das Netz wahrscheinlich zuviele Freiheitsgrade hat.

Praktische Regeln zum Netzwerktraining:

Wahl von 'intelligenten' Variablen: Um gute Resultate mit Neuronalen Netzen zu erzielen, ist es in der Regel wichtig, die benutzten Variablen geschickt auszuwählen und eventuell vorzuverarbeiten.

Kontrolle von Lerngeschwindigkeit und Konvergenzverhalten: Es gibt viele verschiedene Methoden, um das Lernen, das häufig sehr zeitaufwendig sein kann, effektiver zu machen. Dazu gehört die dynamische Anpassung des Lernparameters an die Variation der Fehlerfunktion mit den Gewichten. Statistische Schwankungen im Trainigsdatensatz können durch Hinzufügen eines "Trägheitsterms", der proportional zur Gewichtsänderung im vorhergehenden Schritt ist, gedämpft werden:

$$\Delta w_{ij}^k(t+1) \;=\; -\eta \frac{\partial E}{\partial w_{ij}^k}(t) + \alpha\, \Delta w_{ij}^k(t). \tag{9.32}$$

Dabei ist der Trägheitsparameter α auf das Problem abzustimmen.

Beschränkung der Komplexität eines Netzes:

Wieviele Lagen sind notwendig? Mit 2 Lagen können linear separierbare Probleme behandelt werden (siehe Lösungen der AND-, OR-Probleme mit dem Perzeptron).

Mindestens 3 Lagen werden gebraucht, wenn das Problem nicht linear separierbar ist (zum Beispiel, wenn eine Klasse in zwei disjunkten Bereichen, getrennt durch eine andere Klasse, liegen; siehe XOR-Problem). Ohne Beweis sei angegeben: Mit einem 3-Lagen-Netz kann

- jede kontinuierliche Funktion $y = f(\vec{x})$ approximiert werden,
- jede Boolsche Funktion $y = f(x_1, \ldots, x_n)$, mit $y,\ x_i = 1$ oder 0, dargestellt werden.

Wieviele Knoten pro Lage? Ein geschlossenes Volumen in n Dimensionen kann im allgemeinen durch n+1 Hyperebenen (oder weniger, wenn es zu einer oder mehreren Seiten offen ist,) eingeschlossen werden. Mehr als n+1 Hyperebenen pro geschlossenem, zu selektierendem Volumen liefert mehr Freiheit, den Konturen zu folgen (für das Quadrat ist offensichtlich n+2=4 eine bessere Wahl der Anzahl der Hyperebenen). Wir halten also fest:

- In der Regel sind mindestens n+1 Knoten in der ersten versteckten Lage notwendig.
- Die Zahl der Knoten in der zweiten versteckten Lage hängt von der Komplexität des Problems ab, insbesondere von der Anzahl der nicht-zusammenhängenden Volumina. Es ist wahrscheinlich nicht mehr als ein Knoten pro Volumen notwendig.
- Es sollten so wenig Knoten wie möglich definiert werden, um die Generalisierungsfähigkeit des Systems sicherzustellen.

Entfernen und Generieren von Verbindungen und Knoten: Um die Komplexität des Netzes so gering wie möglich zu halten, sind Techniken entwickelt worden, die erlauben, unwichtige Verbindungen und Knoten zu erkennen und zu entfernen oder auch notwendige Verbindungen und Knoten zu generieren.

Selbstgenerierung der Netz-Architektur: Bei diesem Vorgehen beginnt man zunächst mit einem sehr einfachen Netz und baut dann sukzessiv neue Verbindungen, Knoten und Lagen auf, deren Notwendigkeit dann wieder durch das Verhalten der Fehlerfunktion, der Konvergenz etc. geprüft werden kann.

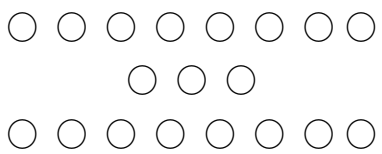
Tabelle 9.1: Vorzeichen der für das Encoder-Problem gefundenen Gewichte w_{ij} in der ersten Schicht.

i $j \rightarrow$	1	2	3	4	5	6	7	8
1	-	+	-	+	+	+	-	-
2	+	-	-	-	+	+	+	-
3	+	-	+	+	-	+	-	-

9.4.7 Typische Anwendungen für Feed-Forward-Netze

Beispiel für ein binäres Netz: 8-Bit-Encoder:

Wir trainieren ein (8-3-8)-Netz



mit 8 Mustervektoren $\vec{x}^p = (x_1^p, \ldots, x_8^p),\ p = 1, \ldots, 8$, und den erwarteten Netzantworten $\hat{\vec{y}}^p = (\hat{y}_1^p, \ldots, \hat{y}_8^p),\ p = 1, \ldots, 8$, denen folgende Binärwerte zugeordnet werden:

$$\begin{aligned} x_i^p &= \delta_{ip} \\ \hat{y}_i^p &= \delta_{ip} \end{aligned}$$

Wir erwarten also das gleiche Muster am Eingang und Ausgang. Wie schafft es das Netz diese Information durch das Nadelöhr von nur 3 Knoten in der versteckten Lage zu transportieren?

Das Netz wurde mit einem PC-Programm (NNSIMU) trainiert. Die Gewichte in der ersten Schicht ergaben sich alle zu etwa $|w_{ij}| \approx 5$. Das Interessante an den Gewichten ist eigentlich nur ihr Vorzeichen, siehe Tab. 9.1. Das Vorzeichen von w_{ij} gibt in diesem Fall direkt die Aktivität des i-ten versteckten Knotens an, wenn das j-te Muster anliegt. Aus der Tabelle erkennt man sofort, dass das Netz den Binärcode 'entdeckt' hat: die redundanten 8-Bit-Sequenzen sind in 3-Bit-Binärzahlen umgewandelt worden.

Funktionsapproximation:

Wie bereits in Abschnitt 9.4.6 ausgeführt, kann mit einem 3-lagigen Netz jede kontinuierliche Funktion,

$$\vec{x} = (x_1, \ldots, x_n) \rightarrow y = f(\vec{x}),$$

approximiert werden.

In Abb. 9.23 ist das Ergebnis eines Trainings der Funktion

$$y = \sin x, \quad 0 < x < \pi$$

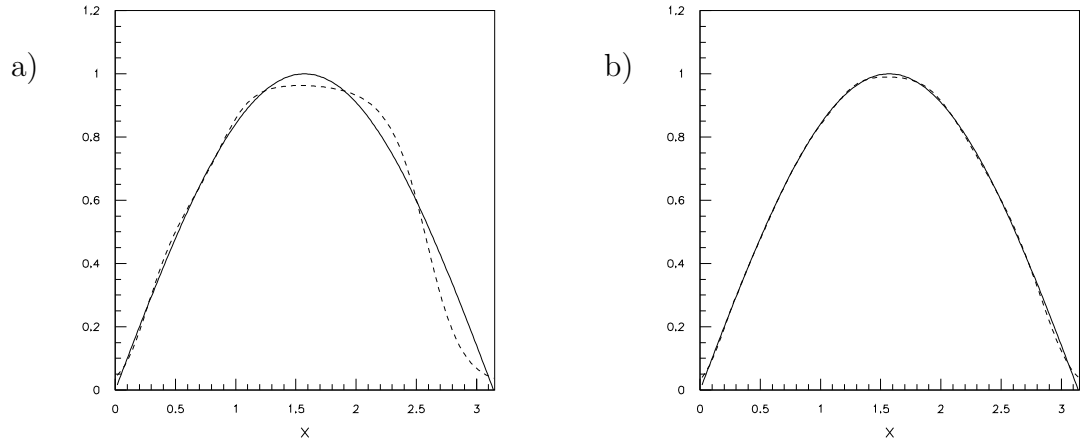


Abbildung 9.23: Approximation einer Sinus-Funktion durch ein (1-8-1)-Netz. Trainingszeiten: a) einige Sekunden, b) etwa 8 Stunden.

gezeigt. Trainiert wurde ein (1-8-1)-Netz mit 200 Musterpaaren (x, y), äquidistant verteilt auf der x-Achse. Nach einigen Lernzyklen, entsprechend einer Rechenzeit von einigen Sekunden, ergab sich die Approximation in Abb. 9.23a. Erst nach etwa 8 Stunden wurde die ausgezeichnete Reproduktion des Sinus durch das Netz in Abb. 9.23b erzielt (diese extrem lange Zeit für ein doch relativ einfaches Problem zeigt eigentlich nur, dass das benutzte Programm nicht sehr effektiv war).

In Abb. 9.24 sind einige Zwischenwerte des Netzes als Funktion von x dargestellt. Es läßt sich gut erkennen, wie daraus die Sinus-Funktion zusammengebaut wird. Außerdem wird durch einige fast verschwindende Aktivitäten nahegelegt, dass Knoten in der versteckten Lage (zum Beispiel der 6. und 8. Knoten) überflüssig sein könnten, die in einem nächsten Schritt entfernt werden könnten.

Klassifikationsprobleme:

Das Problem, Muster in verschiedene Klassen einzuordnen, tritt in unterschiedlichsten Zusammenhängen auf, zum Beispiel:

- Einteilung in disjunkte Klassen: als Beispiele mit kontinuierlichen Musterräumen hatten wir das Quadrat behandelt (siehe Abb. 9.18); Beispiele für diskrete Musterräume sind die Boolschen Funktionen (AND, OR, XOR, ...).

- Die Muster verschiedener Klassen können im allgemeinen auch in Verteilungen liegen, die sich überlappen. Ein einfaches Beispiel sind die überlappenden Gauß-Verteilungen in Abb. 9.20 (mehr dazu im nächsten Abschnitt).

  Gemeinsam ist diesen Fragestellungen, dass von einem bestimmten Muster nicht unbedingt gesagt werden kann, in welcher Klasse es liegt. Im allgemeinen kann nur eine Wahrscheinlichkeit angegeben werden, einer bestimmten Klasse anzugehören. Was die optimale Trennung ist und wie ein NN entscheidet, wird im nächsten Abschnitt besprochen.

- Mustererkennung: Eine der großen Herausforderungen für die Neuroinformatik ist die Verarbeitung und das Erkennen von visuellen, auditiven oder anderen kognitiven Mustern. Von den bisherigen Beispielen unterscheidet sich diese

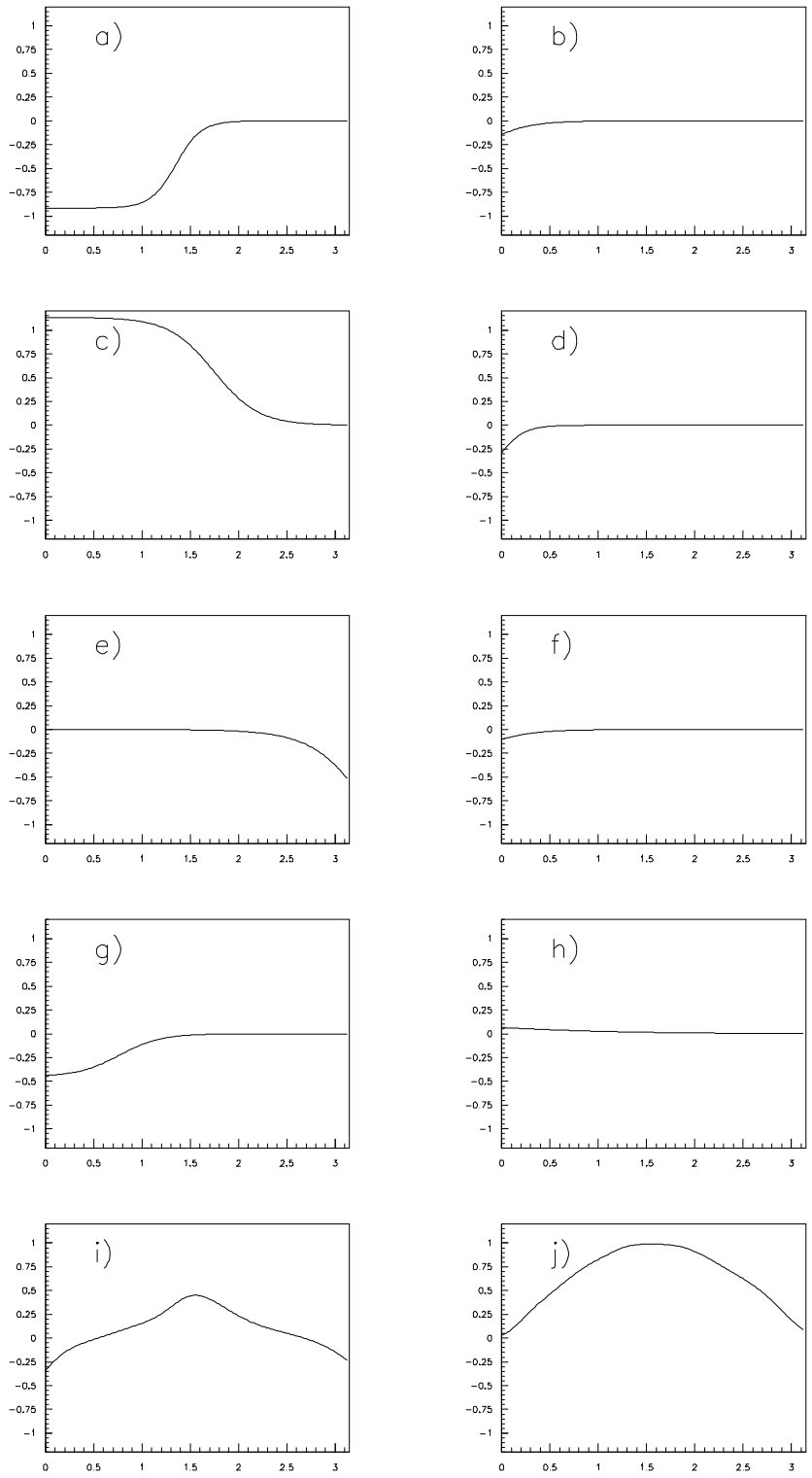


Abbildung 9.24: Für das in Abb. 9.23b benutzte Netz sind als Funktion von x dargestellt: a) bis h) die 8 gewichteten Ausgänge der versteckten Knoten $v_i = w_i' g(z_i)$; i) die Aktivität des Ausgangsknotens $z' = \sum_{i=1,\ldots,8} v_i$, j) das Ausgangssignal $y = g(z')$.

Problemstellung im wesentlichen durch ihre sehr viel größere Komplexität. Ein Bild beispielsweise muß in sehr viele Pixel unterteilt werden, die als Eingabe für das Netz dienen; die Netze werden damit sehr umfangreich. Ein besonderes Problem ist auch die Dynamik, durch die neben der räumlichen auch die zeitliche Dimension ins Spiel kommt. Besonders wichtige Eigenschaften der Netze sind Fehlertoleranz und Rauschunterdrückung.

9.4.8 BP-Lernen und der Bayes-Diskriminator

Die Bayes-Diskriminante:

Es seien Musterklassen $C_i,\ (i = 1, \ldots, m)$, gegeben. Der Bayes-Diskriminator ordnet einen Mustervektor $\vec{x}$ in diejenige Klasse C_i ein, für die die folgende Bayes-Diskriminanten-Funktion maximal ist:

$$P(C_i|\vec{x}) = \frac{p(\vec{x}|C_i)\, P(C_i)}{\sum_{j=1}^{m} p(\vec{x}|C_j)\, P(C_j)} \tag{9.33}$$

Dabei ist

$P(C_i\|\vec{x})$	(a posteriori) Wahrscheinlichkeit, dass $\vec{x}$ in Klasse C_i ist,
$P(C_i)$	(a priori) Wahrscheinlichkeit für Klasse C_i,
$p(\vec{x}\|C_i)$	Wahrscheinlichkeitsverteilung für $\vec{x}$, wenn es in Klasse C_i liegt.

Die Wahrscheinlichkeiten sind normiert:

$$\sum_i P(C_i) \;=\; 1; \qquad \int_{\Omega^n} p(\vec{x}|C_i)\, d^n x$$

Es ist wichtig zu beachten, dass Ω^n das 'beobachtete' Volumen ist, d.h. im allgemeinen ist die tatsächliche Verteilung noch mit einer Akzeptanzfunktion η zu korrigieren:

$$p(\vec{x}|C_i) \;\rightarrow\; p(\vec{x}|C_i)\, \eta(\vec{x}|C_i)$$

Beispiel: Bei der Teilchenidentifikation durch Flugzeitmessung (TOF) wird der Impuls p und die Geschwindigkeit β gemessen. Daraus läßt sich das Quadrat der Masse ('TOF-Masse') bestimmen:

$$m^2_{TOF} \;=\; p^2(\frac{1}{\beta^2} - 1)$$

Die verschiedenen Klassen entsprechen den Teilchensorten Pion, Kaon und Proton $(C_i,\ i = \pi, K, p)$, die mit der Häufigkeit $P(C_i)$ auftreten. Unter der Annahme, dass m^2_{TOF} für eine Teilchensorte i Gauß-verteilt ist um die tatsächliche Masse m_i^2 des Teilchens, ergibt sich für die Verteilung von m^2_{TOF} unter der Hypothese i:

$$p(m^2_{TOF}|C_i) \;=\; \frac{1}{\sqrt{2\pi}\sigma_i} \exp \frac{-(m^2_{TOF} - m_i^2)^2}{2\sigma_i^2}$$

Ein typisches Meßergebnis ist in Abb. 9.25 gezeigt. Die Entscheidung wird dann für das Teilchen gefällt, für das die Diskriminanten-Funktion in (9.33) maximal ist.

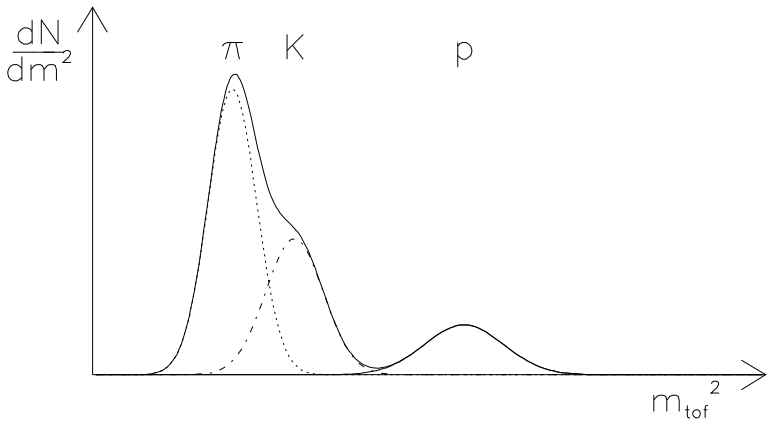


Abbildung 9.25: Typische Verteilung der Massenquadrate, berechnet aus einer Flugzeitmessung für Pionen , Kaonen und Protonen.

Approximation des Bayes-Diskriminators mit neuronalen Netzen:

Ein Netz sei auf die Trennung der beiden Klassen C_1 und C_2 trainiert worden, so dass die erwarteten Netzantworten jeweils sind:

$$\begin{aligned} \hat{y} &= 1 \quad \text{für } \vec{x} \text{ in } C_1 \\ \hat{y} &= 0 \quad \text{für } \vec{x} \text{ in } C_2 \end{aligned}$$

Dann berechnet sich der Erwartungswert der mittleren quadratischen Abweichungen der Netzantworten von den erwarteten Antworten:

$$E = \frac{1}{2} \int d\vec{x} \left[\alpha_1 p_1(\vec{x})(y(\vec{x}) - 1)^2 + \alpha_2 p_2(\vec{x})(y(\vec{x}))^2 \right] \tag{9.34}$$

Das Integral geht über den gesamten Musterraum; die α_i sind die Häufigkeiten, mit denen die Klassen C_i auftreten; die $p_i(\vec{x})$ sind die Wahrscheinlichkeitsverteilungen der Muster $\vec{x}$, wenn sie jeweils einer der beiden Klassen angehören. Mit den Definitionen aus dem vorigen Abschnitt gilt dann also:

$$\begin{aligned} \alpha_i &= P(C_i) \\ p_i(\vec{x}) &= p(\vec{x}|C_i) \end{aligned} \tag{9.35}$$

Bei überlappenden Verteilungen können in der Fehlerfunktion (9.34) die Fehleranteile beider Klassen ungleich Null sein. Dann wird das Minimum nicht mehr unbedingt für $y = 0$ oder 1 erreicht, sondern es gibt eine optimale Wahl des Netzes für y, die sich an jeder Stelle des Musterraumes aus folgender Bedingung herleiten läßt:

$$\frac{\partial E}{\partial y} = \alpha_1 p_1(\vec{x})(y(\vec{x}) - 1) + \alpha_2 p_2(\vec{x}) y(\vec{x}) = 0 \tag{9.36}$$

Die Auflösung nach y ergibt:

$$y(\vec{x}) = \frac{\alpha_1 p_1(\vec{x})}{\alpha_1 p_1(\vec{x}) + \alpha_2 p_2(\vec{x})} \tag{9.37}$$

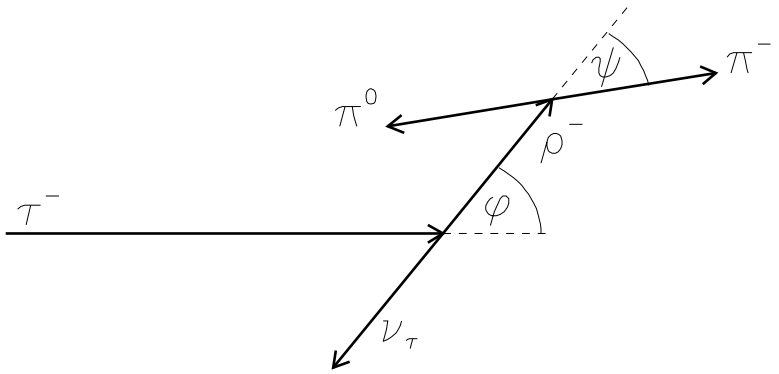


Abbildung 9.26: Darstellung der Zerfallswinkel in Reaktion (9.39).

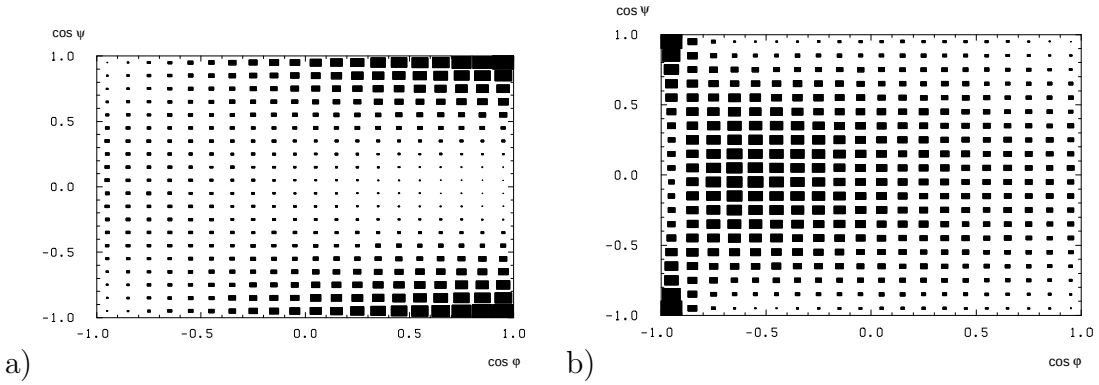


Abbildung 9.27: Winkelverteilung nach (9.40) für τ- Zerfälle im Helizitätszustand $+1$ (a) oder -1 (b).

Die Verallgemeinerung auf m Klassen lautet:

$$y_i(\vec{x}) \;=\; \frac{\alpha_i p_i(\vec{x})}{\sum_{j=1}^{m} \alpha_j p_j(\vec{x})} \tag{9.38}$$

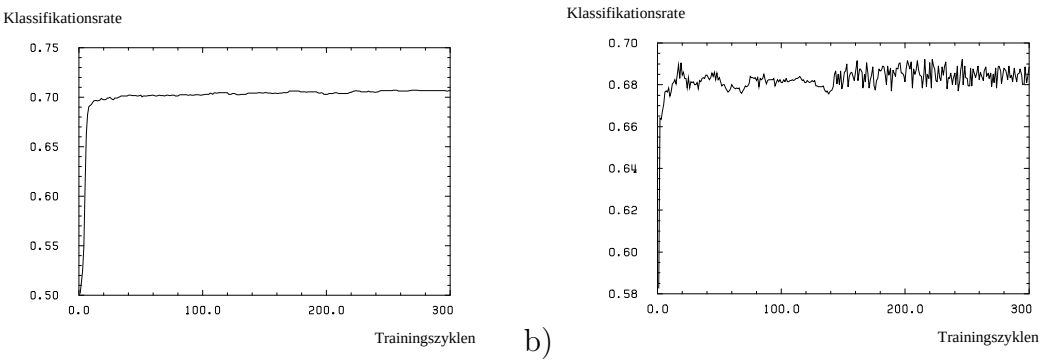
Das maximale y_i bestimmt, in welche Klasse das Muster einzuordnen ist. Bei zwei Klassen ist der Übergang offensichtlich gerade da, wo die beiden Wahrscheinlichkeiten gleich sind:

$$\alpha_1 p_1 \;=\; \alpha_2 p_2 \;\Longrightarrow\; y \;=\; 0.5$$

Im anschließenden Beispiel werden wir sehen, dass ein Netzwerk die optimale Lösung (9.38) approximieren kann.

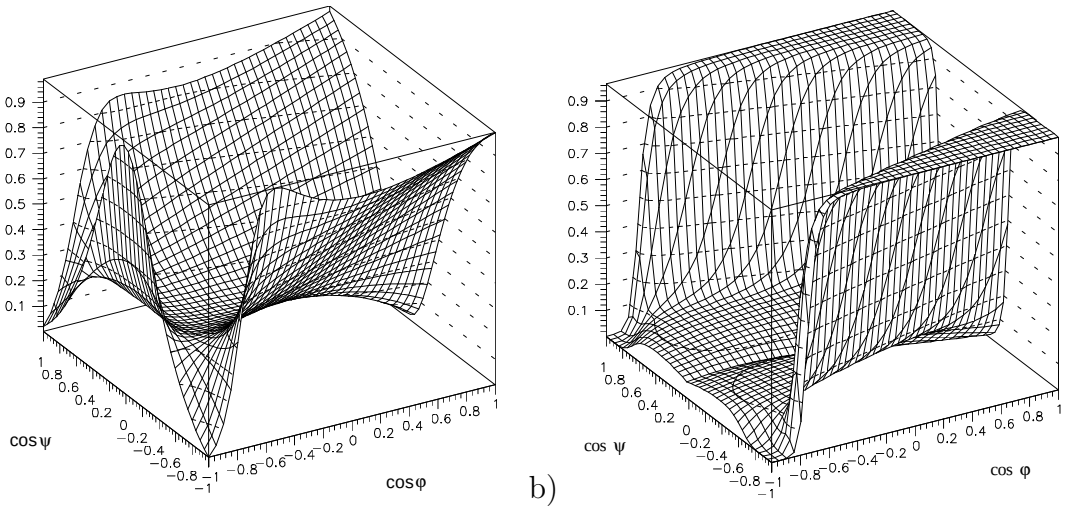
Beispiel für die Approximation des Bayes-Diskriminators durch ein Netz:

Als Beispiel für die Trennung von Klassen mit unterschiedlichen, aber überlappenden Verteilungen nehmen wir die Zerfallswinkelverteilungen von τ-Leptonen in den beiden möglichen Helizitätszuständen $h = +1$ und $h = -1$ (das τ-Lepton hat Spin 1/2; die Helizität ist der auf ± 1 normierte Erwartungswert der Projektion des Spins

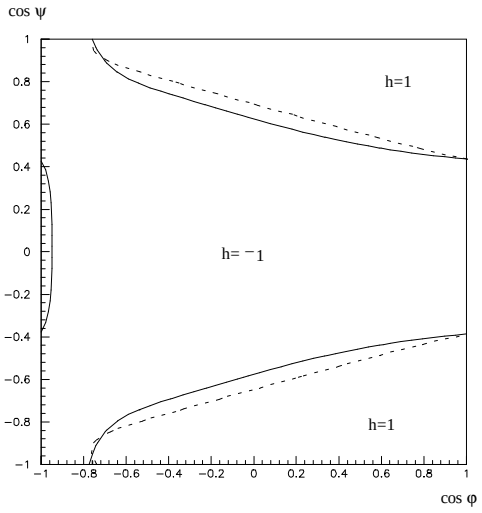


a) b)

Abbildung 9.28: Effizienzen für die Zuordnung des richtigen Helizitätszustandes. Das Netz wurde mit den Lernparametern a) $\eta = 0.001$, $\alpha = 0.9$ und b) $\eta = 0.1$, $\alpha = 0.9$ trainiert.



a) b)



c)

Abbildung 9.29: a) Bayes-Diskriminanten-Funktion aufgetragen über der ($\cos\phi$,-$\cos\psi$)-Ebene; b) dasselbe für den Ausgang y des Netzes. c) Klassifikationsgrenzen für die beiden Helizitäten (volle Linie: Bayes, gepunktete Linie: Netz).

eines Teilchens auf seine Flugrichtung). Wir nehmen an, die τ's seien in einem reinen Helizitätszustand ($h = \pm 1$) produziert worden.

Ein Zerfall, in dem sich die Spininformation im Endzustand gut messen läßt, ist der Zerfall des τ's in ein ρ-Meson mit Spin 1 und ein Neutrino mit Spin 1/2. Während das Neutrino nicht nachzuweisen ist, läßt sich die ρ-Spineinstellung über den ρ-Zerfall in zwei Pionen analysieren:

$$\tau \;\rightarrow\; \rho^- \nu_\tau \;\rightarrow\; \pi^- \pi^0 \nu_\tau \tag{9.39}$$

Die meßbaren Winkel sind der Winkel ϕ zwischen dem ρ und der Laborrichtung des τ (im Ruhesystem des τ) und der Winkel ψ zwischen dem π^- und dem ρ (im ρ Ruhesystem), siehe Abb. 9.26. Die beiden Winkelverteilungen sind Funktionen von $\cos\phi$ und $\cos\psi$:

$$\begin{aligned} P_{+1} &= \cos^2\psi \left[\cos\eta\cos\frac{\phi}{2} + \frac{m_\rho}{m_\tau}\sin\eta\sin\frac{\phi}{2}\right]^2 \\ &\quad + \frac{\sin^2\psi}{2}\left[\left(\sin\eta\cos\frac{\phi}{2} - \frac{m_\rho}{m_\tau}\cos\eta\sin\frac{\phi}{2}\right)^2 + \left(\frac{m_\rho}{m_\tau}\right)^2 \sin^2\frac{\phi}{2}\right] \\ P_{-1} &= \cos^2\psi \left[\cos\eta\sin\frac{\phi}{2} - \frac{m_\rho}{m_\tau}\sin\eta\cos\frac{\phi}{2}\right]^2 \\ &\quad + \frac{\sin^2\psi}{2}\left[\left(\sin\eta\sin\frac{\phi}{2} - \frac{m_\rho}{m_\tau}\cos\eta\cos\frac{\phi}{2}\right)^2 + \left(\frac{m_\rho}{m_\tau}\right)^2 \cos^2\frac{\phi}{2}\right] \end{aligned} \tag{9.40}$$

Dabei ist

$$\cos\eta \;=\; \frac{m_\tau^2 - m_\rho^2 + (m_\tau^2 + m_\rho^2)\cos\phi}{m_\tau^2 + m_\rho^2 + (m_\tau^2 - m_\rho^2)\cos\phi}$$

Abbildung 9.27 zeigt die sich ergebenden zwei-dimensionalen Verteilungen für die beiden Helizitäten.

Mit diesen Verteilungen wurde ein 3-lagiges FF-Netz darauf trainiert, die beiden Helizitäten zu unterscheiden. Die Netzkonfiguration war 2-8-1; der Trainingsdatensatz bestand aus 1000 Ereignissen, gleichviel von jeder Helizität. Abbildung 9.28 zeigt die Effizienz (Anzahl der richtig erkannten Ereignisse zur Gesamtzahl) in Abhängigkeit vom Lernzyklus für einen Testdatensatz. Mit dem Lernparameter $\eta = 0.001$ und dem Trägheitsparameter $\alpha = 0.9$ wird nach 300 Trainingszyklen eine Effizienz von nahezu 71% erreicht. Das kann verglichen werden mit der theoretisch berechenbaren Effizienz bei Benutzung des Bayes-Diskriminators, die sich zu 71.7% ergibt.

In Abb. 9.29 wird gezeigt, dass die Bayes-Diskriminanten-Funktion (Abb. 9.29a) von dem Ausgang y des Netzes (Abb. 9.29b) approximiert wird. Nach einem Schnitt bei $y = 0.5$ ergeben sich die Klassentrennungen, wie in Abb. 9.29c gezeigt. Ob noch eine bessere Approximation der Bayes-Trennung möglich ist, hängt neben einer ausreichenden Netzgröße auch von der Statistik des Trainingsdatensatzes ab. Es ist verständlich, dass zum Beispiel der kleine Zipfel bei $(-1, 0)$ von dem Netz nur dann richtig eingeordnet werden kann, wenn in diesem kleinen Bereich Ereignisse liegen.