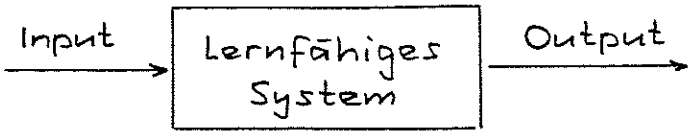
II. LERNEN AUS BEISPIELEN



$\Longrightarrow$ Lernen von Input/Output - Relationen aufgrund von Beispielen

z.B.: - Klassifizierung (Buchstabenerkennung, Diagnose, etc)

- Vorhersage, Modellierung (kontinuierliche Zusammenhänge)

- Strategien, Verhaltensweisen (Spiele, Steuerung, Regelung)

- Lernbeispiele:

a) $I^{\smile} \rightarrow O^{\smile}$: Inputbeispiele mit zugehörigem Output [→ Lernen mit Lehrer]

b) $I^{\smile} \rightarrow B(O^{\smile})$: Inputbeispiele mit zugehöriger Bewertung des Outputs

→ z.B. gut/schlecht

→ ev. nur indirekte Bewertung $[O \rightarrow A, B(A)]$

c) nur $I^{\smile}$: selbständige Konstruktion einer Inputklassifizierung

→ Merkmal-Erkennung (ohne Lehrer!)

II.1 Induktive Lernverfahren

Beispiel: (Quinlan)

Grösse	Haare	Augen		Klasse
k	b	bl		+
g	b	br		−
g	r	bl		+
k	d	bl	⟶	−
g	d	bl		−
g	b	bl		+
g	d	br		−
k	b	br		−
(k	d	br		⋮)
k	r	bl		2
k	r	br		⋮
g	r	br		!

- Lernen = *Identifizieren einer Klassifizierungsregel (eines Klassifizierungskonzepts)*

  [z.B. logische Regel od. Klassif.verfahren]

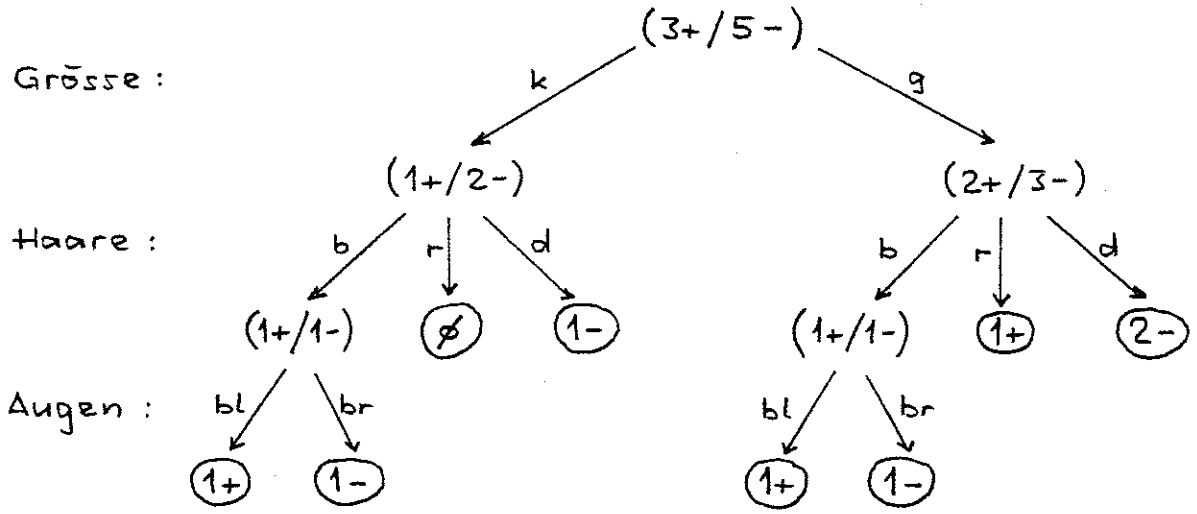
- ID3 - Verfahren:

  J.R. Quinlan, in "Machine Learning", ed. by R.S. Michalski, J.G. Carbonell, and T.M. Mitchell; Springer-Verlag 1984

  - Induktionsalgorithmus (→ if/then Regeln)
  - Konstruktion des "einfachsten" Entscheidungsbaums, der alle Lernbeispiele richtig klassifiziert.
  - ID3 (und entsprechende Weiterentwicklungen) werden oft in kommerziellen AI-Systemen verwendet.

- **Entscheidungsbäume** :

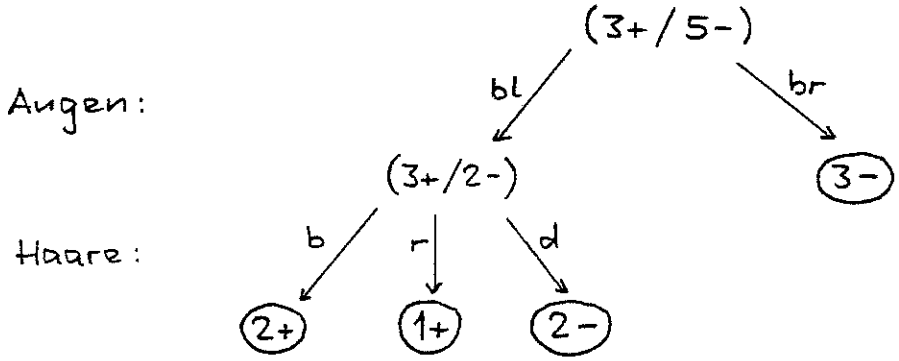
8 Lernbeispiele (3 für Klasse "+" / 5 für "−"):



→ *Komplizierte Regel:*

- Klasse "+" if [k AND b AND bl]
  OR [g AND (r OR (b AND bl))]

- Leere Blätter (keine Beispiele)
  [→ zufällige Klassifizierung]



→ *Einfache Regel:*

- Klasse "+" if [bl AND (b OR r)]

"Einfachster Entscheidungsbaum":

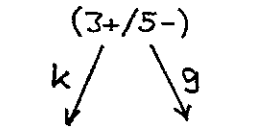
ID3: Informationstheoretische Methode:

- Entropie $E = -p^+ \log p^+ - p^- \log p^-$ $(p^+ + p^- = 1)$

  [oft $\log = \log_2$, da dann $E = 1$ für $p^+ = p^- = \frac{1}{2}$, d.h. für maximale Ignoranz]

- Entropie = Mass für fehlende Information
  → Informationszunahme = Entropieabnahme

- ID3: Maximale Informationszunahme!



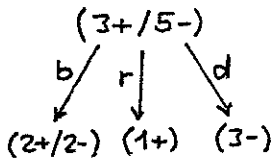
Entropie vor erstem Entscheidungsschritt:

$$E = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0.954$$

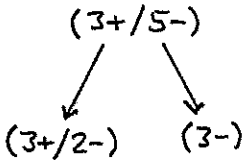
Mittlere Entropie nach Entscheidungsschritt:

$$E = p_k E_k + p_g E_g$$
$$= \frac{3}{8}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right] +$$
$$+ \frac{5}{8}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right] = 0.951$$

→ $\Delta I = -\Delta E = 0.003$ (Grösse)

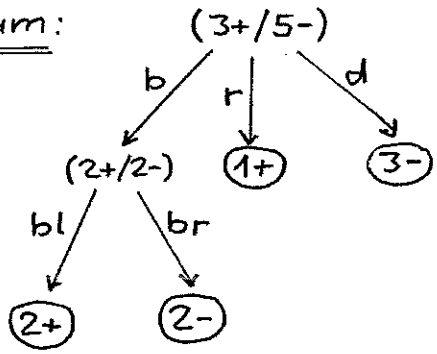


→ $\Delta I = 0.454$ (Haare)



→ $\Delta I = 0.347$ (Augen)

→ **ID3-Baum:**



⟸ 1. Haarfarbe

⟸ 2. Augenfarbe

- Probleme mit induktiven Lernverfahren:

  - Widersprüchliche Beispiele können nicht direkt verarbeitet werden.

  - Empfindlichkeit gegenüber Auswahl der Beispiele.
    (z.B. rote Haare → "+", obwohl nur 1 Beispiel!)

  - Iterative Verarbeitung der Lernbeispiele ist nicht möglich (oder sehr aufwendig).

  - Verallgemeinerung: Nicht definiert bei "leeren Blättern"!

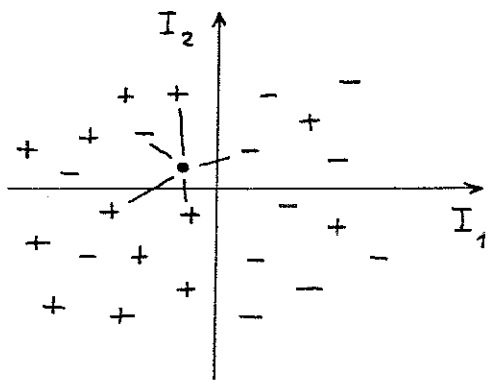
  - Kontinuierliche Inputs: Müssen zuerst diskretisiert werden!

II.2 KNN - Methoden

(KNN = K Nearest Neighbors)

- Definition eines Distanzmasses!
- Mehrheit der K "ähnlichsten" Lernbeispiele entscheidet über Klassenzuteilung

Beispiel: - kontinuierliche Inputs
- K=5



⇒ Input • wird der Klasse "+" zugeordnet

II.3 Bayes'sche Klassifikatoren

- Inputvektor (Evidenzvektor): $\underline{e} = (e_1, e_2, \ldots, e_d)$
- Klassen: $C_i \quad (i = 1, \ldots, N)$

→ Entscheidungsregel:

Bei gegebener Evidenz $\underline{e}$ wird diejenige Klasse C_i gewählt, für die gilt:

$$P(C_i|\underline{e}) > P(C_j|\underline{e}) \quad \text{für alle } j \neq i$$

wobei: $P(C/\underline{e})$ = Wahrscheinlichkeit der Klasse C bei gegebener Evidenz $\underline{e}$

Bayes Rule:

$$P(C|\underline{e}) = \frac{P(\underline{e}|C)\,P(C)}{P(\underline{e})}$$

→ Entscheidungsregel:

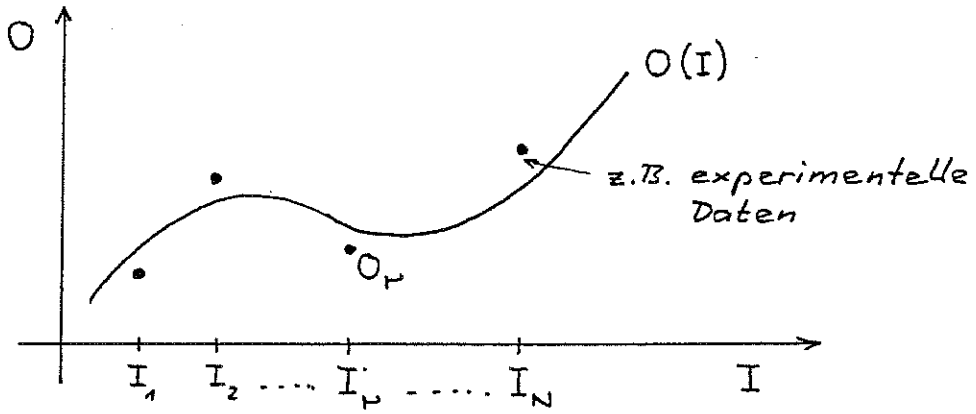
$$P(\underline{e}|C_i)\,P(C_i) > P(\underline{e}|C_j)\,P(C_j) \quad \text{für alle } j \neq i$$

- $P(\underline{e}/C_i)$ und $P(C_i)$, $i = 1, \ldots, N$, werden aus den vorliegenden Daten abgeschätzt (→ LERNPROZESS!)
- Wenn die Evidenzen e_j kontinuierlich sind, bildet man zuerst e_j-Intervalle!
- Oft wird die vereinfachende Annahme gemacht, die Evidenzen e_j seien statistisch unabhängig:

$$\rightarrow P(\underline{e}/C_i) = \prod_{j=1}^{N} P(e_j/C_i)$$

II.4 Regressions-Verfahren

→ z.B. "Kurvenfitten"
(bzw. entsprechende Verfahren für mehrdimensionale Input-Räume)



→ "Lernen" einer Funktion (eines Zusammenhangs) $O(I)$

(i) Wahl einer Familie $O(I;\underline{a})$ von Funktionen
(z.B. $O = a_0 + a_1 I + a_2 I^2$)

(ii) Bestimme Parameter $\underline{a}$ so, dass

$$\sum_\mu \left[O(I_\mu;\underline{a}) - O_\mu \right]^2 = \min$$

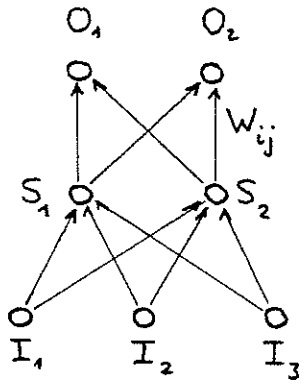
[Minimierung des quadratischen Fehlers über alle Lernbeispiele!]

(iii) Ev. Regularisierungsterm:
(Glattheitsforderung)

$$\sum_\mu \left[O(I_\mu;\underline{a}) - O_\mu \right]^2 + \lambda \int dI \left[\frac{d^2 O}{dI^2} \right]^2 = \min$$

λ: Regularisierungsparameter

II.5 Lernen in neuronalen Netzwerken



$$O_i = f\left(\sum_j W_{ij}\, S_j\right)$$

$$S_j = f\left(\sum_k W_{jk}\, I_k\right)$$

$$\rightarrow \quad \underline{O} = \underline{O}\left(\underline{I}, \underline{\underline{W}}\right)$$

→ LERNEN = ADAPTIEREN DER GEWICHTE $\underline{\underline{W}}$ AUFGRUND VON BEISPIELEN

A) "Supervised Learning" (Lernen mit einem Lehrer)

1) Wähle einen Satz von Lernbeispielen:

→ Paare $\underline{I}^\nu / \underline{D}^\nu$, $\nu = 1, \ldots, N$

[$\underline{D}^\nu$ = "gewünschter" (vorgegebener) Output]

2) Definiere ein Mass für den Output-Fehler:

→ $F^\nu = F\left(\underline{O}^\nu(\underline{\underline{W}}) , D^\nu\right)$

[z.B. $F^\nu = (\underline{D}^\nu - \underline{O}^\nu)^2$]

3) Anpassung der Gewichte W_{ij} , sodass

$$\sum_\nu F^\nu = \min$$

→ LERNEN = OPTIMIEREN

→ LERNVERFAHREN = OPTIMIERUNGSVERFAHREN

→ Neuronale Netzwerke mit "supervised Learning" sind theoretisch gesehen nichts anderes als eine Oberklasse von Regressionsmodellen.

→ Wesentlicher Vorteil:

Neuronale Netzwerke sind "universelle Approximatoren", während bei herkömmlichen Regressionsmodellen die Struktur der Regressionsfunktion im Voraus festgelegt werden muss (z.B. Polynom n-ten Grades).

B) Lernen mit Bewertung

("Lernen mit einem Kritiker")

→ In den Lernbeispielen ist der gewünschte Output nicht explizite vorgegeben, sondern nur eine Bewertung der möglichen Outputs:

$B(\underline{I}^{\nu}, \underline{O})$: kann eine skalare oder auch nur eine binäre Grösse sein (z.B. gut/schlecht)

→ Optimierung der Bewertung

C) "Unsupervised Learning" (kein Lehrer)

→ Nur Input $\underline{I}^{\nu}$ der Lernbeispiele vorgegeben

→ Automatische Klassifizierung der Inputbeispiele

→ Lernprinzipien, Lernverfahren ?

(z.B. Hopfield-Netzwerke, Kohonen-Netzwerke !)

II.6 Vergleich mit konventionellen Lernsystemen

Vorteile neuronaler Netzwerke:

- Einfache Implementierbarkeit
- Kleiner Speicherplatzbedarf
- Schnelle Klassifizierung
- Robustheit
- Adaptierbarkeit
- Diskrete und kontinuierliche Inputs und Outputs

Nachteile:
- Lange Lernzeit
- Interpretation der Lösungen schwierig

II.7 Lernen und Verallgemeinern

- Lernen: $\sum_{\mu}^{\text{Lernbsp.}} F^{\mu} = \min$

  (F = Mass für Outputfehler)

- Verallgemeinern:

$$\sum_{\mu}^{\text{Testbsp.}} F^{\mu} = \min$$

(Testbeispiele nicht aus Lern-Set!)

- Definition, Bewertung der Verallgemeinerungsfähigkeit?

- Abhängigkeit der Verallgemeinerungsfähigkeit von der Netzwerkarchitektur, etc.?

ANHANG:

VERGLEICH VERSCHIEDENER LERNSTRATEGIEN
(Bernasconi & Gustafson, 1994/1998)

Verallgemeinerung beim Quinlan-Problem:

Objekt:	Klassifizierung:					andere	
(9)	−	−	−	−	−	⋮	
(10)	+	+	−	+	+	⋮	
(11)	−	+	−	−	+	⋮	
(12)	−	+	−	+	−	⋮	
ID3	0	100	0	0	0	0	%
Neuronale Netzwerke	71.4	12.1	9.4	4.2	2.9	0	%
Menschen	73.2	14.6	2.4	4.9	0	4.8	%

Beobachtungen:

- Menschen und neuronale Netzwerke bevorzugen eine andere Klassifizierungsregel als das ID3-Verfahren.
- Die ID3-Strategie hat eine Tendenz, die Bedeutung von Daten zu überschätzen, die statistisch nicht signifikant sind.
- Neuronale Netzwerke erfassen die statistische Information in den Daten wesentlich effizienter als ID3.