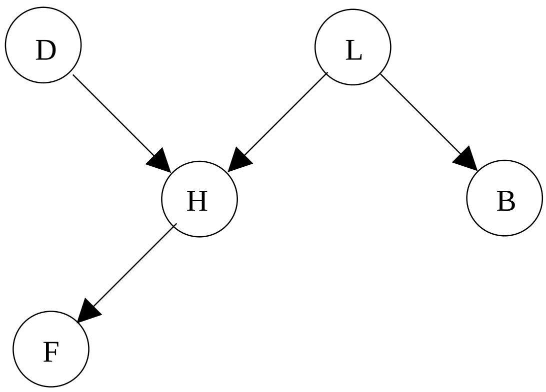
Aufgabe 1 – Probabilistische Inferenz (28 Punkte)

Ein Modell soll es ermöglichen, Fotos zu klassifizieren. Dabei geht man von folgenden Annahmen aus:

- 75% aller Fotos sind überwiegend hell ($L = w$).
- Wenn das Bild überwiegend hell ist ($L = w$), wurde in einem Zehntel der Fälle ein Blitz benutzt ($B = w$). Im anderen Fall sind 50% aller Bilder mit Blitz entstanden.
- Bei 2 von 5 Fotos überwiegen im oberen Drittel Blautöne ($D = w$).
- Ist das Bild hell ($L = w$) und das obere Drittel blau ($D = w$), dann ist zu 90% der Himmel ($H = w$) im Foto. Trifft **beides** nicht zu, ist zu 30% der Himmel im Bild, und wenn genau eine der beiden Voraussetzungen erfüllt ist, gilt es zu 10%.
- Ist der Himmel im Bild ($H = w$), dann handelt es sich zu 80% um ein Panoramafoto ($F = w$), ansonsten trifft dies nur auf ein Fünftel der Bilder zu.

(a) Zeichnen Sie ein Bayes-Netz, das zu diesem Modell passt! Die Wahrscheinlichkeitstabellen brauchen Sie hierfür nicht anzugeben. (6 Punkte)



(b) Wie groß ist die Wahrscheinlichkeit, dass in einem zufälligen Bild der Himmel enthalten ist? (6 Punkte)

$$\begin{aligned}
P(H = w) &= \sum_{x=w,f} \sum_{y=w,f} P(H = w|D = x, L = y)P(D = x)P(L = y) \\
&= 0.9 \cdot 0.75 \cdot 0.4 + 0.3 \cdot 0.25 \cdot 0.6 + 0.1 \cdot 0.75 \cdot 0.6 + 0.1 \cdot 0.25 \cdot 0.4 \\
&= 0.37
\end{aligned}$$

(c) Wie wahrscheinlich ist es, dass ein Foto überwiegend hell ist, wenn der Blitz benutzt wurde? (8 Punkte)

$$\begin{aligned} P(L=w|B=w) &= \frac{P(B=w|L=w)P(L=w)}{\sum_{x=w,f} P(B=w|L=x)P(L=x)} \\ &= \frac{0.1 \cdot 0.75}{0.1 \cdot 0.75 + 0.5 \cdot 0.25} \\ &= 0.375 \end{aligned}$$

(d) Wie sicher kann man sich sein, dass es sich um ein Panoramafoto handelt, wenn das Bild nicht überwiegend hell ist? (8 Punkte)

$$\begin{aligned} & P(F=w|L=f) \\ =\; & \sum_{x=w,f} \sum_{y=w,f} P(F=w|L=f, H=x, D=y) P(H=x, D=y|L=f) \\ =\; & \sum_{x=w,f} \sum_{y=w,f} P(F=w|H=x) P(H=x|D=y, L=f) P(D=y|L=f) \\ =\; & \sum_{x=w,f} \sum_{y=w,f} P(F=w|H=x) P(H=x|D=y, L=f) P(D=y) \\ =\; & 0.8 \cdot (0.1 \cdot 0.4 + 0.3 \cdot 0.6) + 0.2 \cdot (0.9 \cdot 0.4 + 0.7 \cdot 0.6) \\ =\; & 0.332 \end{aligned}$$

Aufgabe 2 – Hidden-Markov-Modell (28 Punkte)

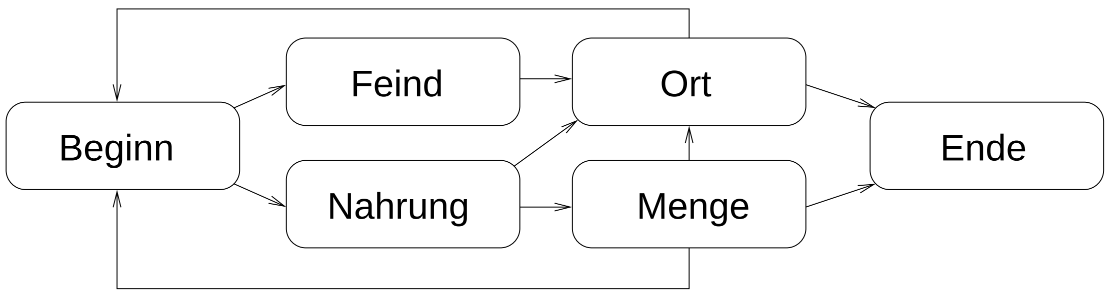
Eine neuentdeckte Chamäleonart nutzt ihre Hautfarbe, um komplexe Botschaften zu kommunizieren. Wir unterscheiden zwischen einer Folge von Segmenten $x_1, x_2, x_3, \ldots$ und der tatsächlich beobachteten Folge von Farben $y_1, y_2, y_3, \ldots$.

x_i	x_{i+1}	$P(x_{i+1}\|x_i)$
Beginn	Feind	0.4
Beginn	Nahrung	0.6
Feind	Ort	1.0
Nahrung	Ort	0.2
Nahrung	Menge	0.8
Ort	Beginn	0.3
Ort	Ende	0.7
Menge	Ort	0.3
Menge	Beginn	0.2
Menge	Ende	0.5

x_i	y_i	$P(y_i\|x_i)$
Beginn	weiß	1.0
Feind	rot	0.6
Feind	blau	0.4
Nahrung	rot	0.7
Nahrung	grün	0.3
Ort	blau	0.8
Ort	orange	0.2
Menge	blau	0.1
Menge	grün	0.9
Ende	schwarz	1.0

Die Markovkette beginnt immer mit $x_1 = Beginn$ und endet mit $x_k = Ende$. Alle nicht angegebenen Wahrscheinlichkeiten $P(x_{i+1}|x_i)$ und $P(y_i|x_i)$ sind Null. Sie können die Segmenttypen mit großen und die Farben mit kleinen Anfangsbuchstaben abkürzen, um Platz zu sparen.

(a) Stellen Sie das Modell für die Ausdrücke in einem Übergangsdiagramm graphisch dar! Sie brauchen keine Wahrscheinlichkeiten einzutragen. (6 Punkte)



(b) Mit welcher Wahrscheinlichkeit tritt die Farbfolge „weiß-rot-blau-orange-schwarz“ in diesem Modell auf? (6 Punkte)

Es gibt nur eine Segmentfolge die diese Farben erzeugen kann:

$$\begin{aligned} P(\text{„w-r-b-o-s“}) &= P(N|B)P(M|N)P(O|M)P(E|O) \\ &\cdot\ P(w|B)P(r|N)P(b|M)P(o|O)P(s|E) \\ &= 0.6 \cdot 0.8 \cdot 0.3 \cdot 0.7 \\ &\cdot\ 1.0 \cdot 0.7 \cdot 0.1 \cdot 0.2 \cdot 1.0 \\ &\approx 0.0014 \end{aligned}$$

(c) Sie beobachten die Folge „weiß-rot-orange-schwarz“. Ist es wahrscheinlicher, dass es um Nahrung oder Feinde geht? Wie sicher ist dies? (8 Punkte)

- Wenn es sich um einen Feind handelt:

$$\begin{aligned} P(x_2 = F|\text{„w-r-o-s“}) &\propto P(F|B)P(O|F)P(E|O) \\ &\cdot\ P(w|B)P(r|F)P(o|O)P(s|E) \\ &= 0.4 \cdot 1.0 \cdot 0.7 \\ &\cdot\ 1.0 \cdot 0.6 \cdot 0.2 \cdot 1.0 \\ &= 0.0336 \end{aligned}$$

- Wenn es sich um Nahrung handelt:

$$\begin{aligned} P(x_2 = N|\text{„w-r-o-s“}) &\propto P(N|B)P(O|N)P(E|O) \\ &\cdot\ P(w|B)P(r|N)P(o|O)P(s|E) \\ &= 0.6 \cdot 0.2 \cdot 0.7 \\ &\cdot\ 1.0 \cdot 0.7 \cdot 0.2 \cdot 1.0 \\ &= 0.01176 \end{aligned}$$

- Normierung:

$$\frac{0.0336}{0.0336 + 0.01176} \approx 0.74$$

- Die Nachricht handelt zu ungefähr 74% von einem Feind.

(d) Wie wahrscheinlich ist eine Nachricht aus 4 Segmenten? (8 Punkte)

- Es gibt 3 Möglichkeiten für eine Nachricht aus 4 Segmenten: Beginn-Feind-Ort-Ende, Beginn-Nahrung-Ort-Ende und Beginn-Nahrung-Menge-Ende.
- Beginn-Feind-Ort-Ende:

$$\begin{aligned} P(BFOE) &= P(F|B)P(O|F)P(E|O) \\ &= 0.4 \cdot 1.0 \cdot 0.7 \\ &= 0.28 \end{aligned}$$

- Beginn-Nahrung-Ort-Ende

$$\begin{aligned} P(BNOE) &= P(N|B)P(O|N)P(E|O) \\ &= 0.6 \cdot 0.2 \cdot 0.7 \\ &= 0.084 \end{aligned}$$

- Beginn-Nahrung-Menge-Ende

$$\begin{aligned} P(BNME) &= P(N|B)P(M|N)P(E|M) \\ &= 0.6 \cdot 0.8 \cdot 0.5 \\ &= 0.24 \end{aligned}$$

- Gesamtwahrscheinlichkeit:

$$\begin{aligned} P(BFOE) + P(BNOE) + P(BNME) &= 0.28 + 0.084 + 0.24 \\ &= 0.604 \end{aligned}$$

- Eine Nachricht mit vier Segmenten tritt mit der Wahrscheinlichkeit 60.4% auf.

Aufgabe 3 – Statistische Lernmethoden (24 Punkte)

Die Berliner S-Bahn ist dafür bekannt, dass die Züge häufig zu spät ankommen. Als einfaches Modell nehmen Sie an, dass jede Zugfahrt unabhängig ist und eine Verspätung mit der Wahrscheinlichkeit p auftreten kann. Allerdings hatten Sie in diesem Monat Glück und bei bisher 10 Fahrten mit der S-Bahn waren die Züge alle pünktlich.

(a) Welche Wahrscheinlichkeitsverteilung hat das Auftreten von mindestens einer Verspätung bei insgesamt n Zugfahrten, wenn die Wahrscheinlichkeit hierfür bei jeder Fahrt p beträgt? (4 Punkte)

$$P(n|p) = 1 - (1-p)^n$$

(b) Die S-Bahn plant ein neues Entschädigungsmodell für Käufer von 4-Fahrten-Karten. Diese sollen 1 Euro zurückerhalten, falls es auf mindestens einer der 4 Fahrten zu einer Verspätung kommt. Wie hoch ist die zu erwartende Entschädigung für eine 4-Fahrten-Karte, wenn Sie $p = 0.1$ annehmen? (4 Punkte)

$$\langle E \rangle = 1 - (1-0.1)^4 = 1 - 0.9^4 = 1 - 0.6561 = 0.3439$$

(c) In der Zeitung lesen Sie, dass zwei von fünf S-Bahn-Zügen verspätet am Ziel ankommen. Sie wollen diese Information als Vorwissen in Form einer Beta-Verteilung $\mathrm{Beta}(p; \alpha, \beta) = B(\alpha, \beta)\, p^{\alpha-1}(1-p)^{\beta-1}$ in ihrer Schätzung von p nutzen. Wie sollten Sie die Hyperparameter α und β wählen? Begründen Sie! (4 Punkte)

- Die Nachricht entspricht einer Beobachtung von 2 Verspätungen bei 5 Fahrten.
- In diesem Fall ist die zugehörige Likelihood durch die Binomialverteilung

  $$P(k=2|n=5,p) = \binom{5}{2} p^2 (1-p)^3$$

  gegeben.
- Um diese Information im neuen Prior zu verwenden, sollte dieser proportional zur oben berechneten Likelihood gewählt werden:

  $$\begin{aligned} & P(p) \propto P(k=2|n=5,p) \\ \iff\ & p^2(1-p)^3 = p^{\alpha-1}(1-p)^{\beta-1} \\ \iff\ & \alpha = 3 \wedge \beta = 4 \end{aligned}$$

  Das entspricht den Hyperparametern $\alpha = 3$ und $\beta = 4$.

(d) Zeigen Sie, dass die Maximum-a-posteriori Hypothese für eine Folge von n pünktlichen Zügen durch $p = (\alpha - 1)/(n + \alpha + \beta - 2)$ gegeben ist! (8 Punkte)

- Der Logarithmus des Posteriors ist durch

$$\begin{aligned}
\log P(p|n) &= \log \left[\frac{P(n|p)\mathrm{Beta}(p;\alpha,\beta)}{P(n)} \right] \\
&= n \log(1-p) + (\alpha - 1) \log p + (\beta - 1) \log(1-p) + \log C \\
&= (\alpha - 1) \log p + (n + \beta - 1) \log(1-p) + \log C
\end{aligned}$$

  mit der Normierungskonstanten $C = B(\alpha, \beta)/P(n)$ gegeben.

- Berechnung der Ableitung:

$$\frac{\mathrm{d}}{\mathrm{d}p} \log P(p|n) = \frac{\alpha - 1}{p} - \frac{n + \beta - 1}{1 - p}$$

- Bedingung für ein Maximum

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}p} \log P(p|n) = 0 \quad &\Longleftrightarrow \quad \frac{\alpha - 1}{p} = \frac{n + \beta - 1}{1 - p} \\
&\Longleftrightarrow \quad (\alpha - 1)(1 - p) = (n + \beta - 1)p \\
&\Longleftrightarrow \quad p = \frac{\alpha - 1}{n + \alpha + \beta - 2}
\end{aligned}$$

(e) Welchen Wert p hat die Maximum-a-posteriori-Hypothese für $n = 10$, wenn Sie die Hyperparameter auf $\alpha = 2$ und $\beta = 6$ setzen? (4 Punkte)

$$p = \frac{\alpha - 1}{n + \alpha + \beta - 2} = \frac{2 - 1}{10 + 2 + 6 - 2} = \frac{1}{16} \approx 0.0625$$

Aufgabe 4 – Neuronales Netz (20 Punkte)

Betrachten Sie ein Perzeptron mit zwei reellwertigen Eingabeneuronen e_1 und e_2, zwei Gewichten w_1 und w_2 sowie einem Bias w_0. Für das lokale Feld h gilt

$$h = w_1 e_1 + w_2 e_2 - w_0$$

und zur Berechnung der Ausgabe a wird die Aktivierungsfunktion

$$a = \operatorname{sgn}(h) = \begin{cases} +1 & \text{für } h > 0 \\ -1 & \text{für } h \leq 0 \end{cases}$$

verwendet. Alle Gewichte und der Bias werden mit 1 initialisiert. Das Perzeptron soll nun die folgenden Beispiele (Soll-Ausgabe y) lernen:

e_1	-2	$+2$	-2	$+4$
e_2	$+1$	$+2$	$+4$	-1
y	-1	$+1$	-1	$+1$

(a) Welche der vier Beispiele werden vom Perzeptron ohne eine Anpassung der Gewichte falsch klassifiziert? (4 Punkte)

Berechnung der Ausgaben ohne Training:

$$\begin{aligned} a_1 &= \operatorname{sgn}((-2) + (+1) - 1) &= \operatorname{sgn}(-2) &= +1 = y_1 \\ a_2 &= \operatorname{sgn}((+2) + (+2) - 1) &= \operatorname{sgn}(+3) &= +1 = y_2 \\ a_3 &= \operatorname{sgn}((-2) + (+4) - 1) &= \operatorname{sgn}(+1) &= +1 \neq y_3 \\ a_4 &= \operatorname{sgn}((+4) + (-1) - 1) &= \operatorname{sgn}(+2) &= +1 = y_4 \end{aligned}$$

Es wird nur Beispiel 3 falsch klassifiziert.

(b) Wie ändern sich die Gewichte, wenn das neuronale Netz gemäß der Perzeptron-Lernregel mit jedem Beispiel einmal trainiert wird? Verwenden Sie $\lambda = 0.4$ als Lernrate und passen Sie auch den Bias w_0 an. (4 Punkte)

- Perzeptron-Lernregel:

$$\left.\begin{array}{ll} a_1 = y_1 & a_2 = y_2 \\ a_3 \neq y_3 & a_4 = y_4 \end{array}\right\} \Longrightarrow \text{Training zunächst nur für Beispiel 3 nötig.}$$

- Für die neuen Gewichte gilt:

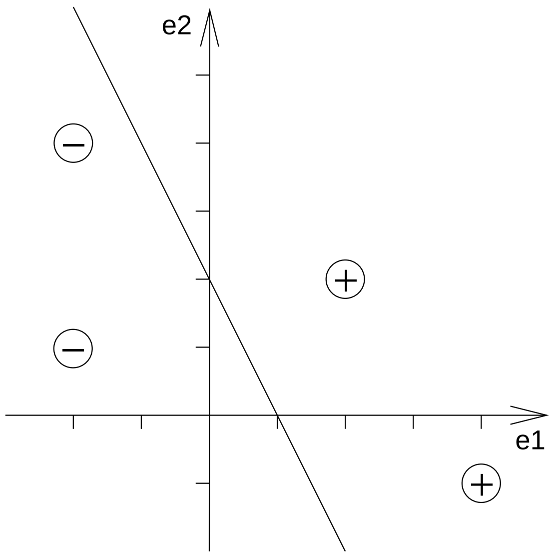
$$\begin{aligned} w_0^+ = 1 + 0.4 \cdot (-1) \cdot (-1) &= +1.4 \\ w_1^+ = 1 + 0.4 \cdot (-1) \cdot (-2) &= +1.8 \\ w_2^+ = 1 + 0.4 \cdot (-1) \cdot (+4) &= -0.6 \end{aligned}$$

- Test für nachfolgende Beispiele:

$$\begin{aligned} a_1 = {} & \operatorname{sgn}(1.8 \cdot (-2) - 0.6 \cdot (+1) - 1.4) &= \operatorname{sgn}(-5.6) &= +1 = y_1 \\ a_2 = {} & \operatorname{sgn}(1.8 \cdot (+2) - 0.6 \cdot (+2) - 1.4) &= \operatorname{sgn}(+1.0) &= +1 = y_2 \\ a_3 = {} & \operatorname{sgn}(1.8 \cdot (-2) - 0.6 \cdot (+4) - 1.4) &= \operatorname{sgn}(-7.4) &= -1 = y_3 \\ a_4 = {} & \operatorname{sgn}(1.8 \cdot (+4) - 0.6 \cdot (-1) - 1.4) &= \operatorname{sgn}(+6.4) &= +1 = y_4 \end{aligned}$$

- Je nach Reihenfolge des Trainings sind diese Tests nur teilweise (zum Beispiel für 1-2-3-4) oder gar nicht (zum Beispiel für 1-2-4-3) erforderlich.

(c) Stellen Sie das Klassifikationsproblem in der e_1-e_2-Ebene graphisch dar und lösen Sie es durch möglichst wenige lineare Entscheidungsgrenzen! (6 Punkte)



(d) Kann ein Perzeptron die Beispiele bei genügend langem Training exakt lernen? Begründen Sie Ihre Antwort! (2 Punkte)

Die positiven und negativen Beispiele können durch eine Gerade voneinander getrennt werden. Also ist die Trainingsmenge *linear separabel* und kann von einem Perzeptron perfekt gelernt werden.

(e) Geben Sie für eine Klassifikationsgrenze, die durch die Punkte $(e_1, e_2) = (0, 2)$ und $(e_1, e_2) = (3, 0)$ verläuft, die zugehörigen Gewichte des Perzeptrons an! (4 Punkte)

An den beiden Punkten gilt $h = 0$:

$$\begin{aligned} w_1 \cdot 0 + w_2 \cdot 2 - w_0 = 0 \quad &\Longrightarrow \quad 2w_2 = w_0 \\ w_1 \cdot 3 + w_2 \cdot 0 - w_0 = 0 \quad &\Longrightarrow \quad 3w_1 = w_0 \\ \text{Wähle:} \quad w_0 = 6 \quad &\Longrightarrow \quad w_1 = 2, w_2 = 3 \end{aligned}$$

Ein geeigneter Gewichtsvektor für das Perzeptron ist $w_1 = 2, w_2 = 3, w_0 = 6$.