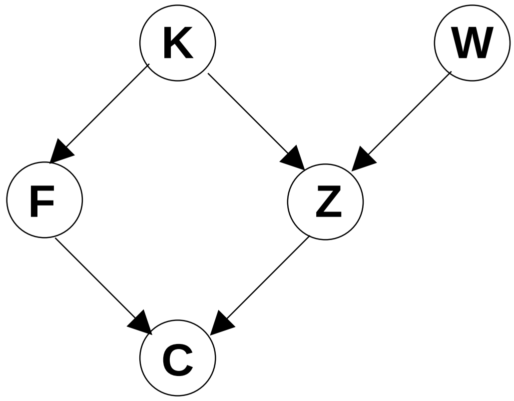
Aufgabe 1 – Probabilistische Inferenz (28 Punkte)

Die BVG will besser auf Ausfälle im S-Bahn-Verkehr eingestellt sein. Sie geht dabei von folgenden Annahmen aus:

- An 20% der Tage ist es besonders kalt ($K = w$), an den restlichen Tagen ist die Temperatur gemäßigt oder warm ($K = f$).
- An kalten Tagen ($K = w$) ist zu 70% mit einem erhöhtem Krankenstand bei S-Bahn-Fahrern zu rechnen ($F = w$). Bei normalen Temperaturen (K=f) dagegen tritt diese Situation nur zu 10% ein.
- An 70% der Tage wurde die Wartung der Züge ($W = w$) nicht ausreichend durchgeführt.
- Ist es kalt und die Wartung ist nicht ausreichend ($W = w \wedge K = w$), sind zu 90% mehr als 1/4 der Züge nicht funktionstüchtig ($Z = w$). Trifft nur genau eine dieser Voraussetzungen zu ($W = w \oplus K = w$), ist die Wahrscheinlichkeit 0.6 und trifft keine zu ($W = f \wedge K = f$) ist die Wahrscheinlichkeit 0.1.
- Tritt ein erhöhter Krankenstand oder ein Ausfall von mehr als 1/4 der Züge ein ($F = w \vee Z = w$), ist zu 90% mit einem S-Bahn-Chaos zu rechnen (C=w). Andernfalls ist die Wahrscheinlichkeit 25%.

(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 es an einem kalten Tag, an dem alle Züge funktionstüchtig sind, zu einem S-Bahn-Chaos kommt? (6 Punkte)

$$\begin{aligned}
P(C = w|Z = f, K = w) &= P(C = w|Z = f, K = w, F = w)P(F = w|K = w) \\
&+ P(C = w|Z = f, K = w, F = f)P(F = f|K = w) \\
&= P(C = w|Z = f, F = w)P(F = w|K = w) \\
&+ P(C = w|Z = f, F = f)P(F = f|K = w) \\
&= 0.9 \cdot 0.7 + 0.25 \cdot 0.3 = 0.63 + 0.075 = 0.705
\end{aligned}$$

(c) Wie hoch ist die Wahrscheinlichkeit, dass es ein kalter Tag ist, wenn es einen erhöhten Krankenstand gibt? (6 Punkte)

$$\begin{aligned}
P(K=w|F=w) &= \frac{P(F=w|K=w)P(K=w)}{P(F=w)} \\
&= \frac{P(F=w|K=w)P(K=w)}{P(F=w|K=w)P(K=w)+P(F=w|K=f)P(K=f)} \\
&= \frac{0.7\cdot 0.2}{0.7\cdot 0.2+0.1\cdot 0.8} = \frac{0.14}{0.14+0.08} = \frac{0.14}{0.22} = \frac{7}{11} \approx 0.636
\end{aligned}$$

(d) Mehr als 1/4 der Züge sind ausgefallen. Wie sicher kann man sich sein, dass nicht ausreichend gewartet wurde? (10 Punkte)

$$\begin{aligned}
P(Z=w|W=w) &= P(Z=w|W=w,K=w)P(K=w) \\
&+ P(Z=w|W=w,K=f)P(K=f) \\
&= 0.9\cdot 0.2+0.6\cdot 0.8 = 0.18+0.48 = 0.66
\end{aligned}$$

$$\begin{aligned}
P(Z=w|W=f) &= P(Z=w|W=f,K=w)P(K=w) \\
&+ P(Z=w|W=f,K=f)P(K=f) \\
&= 0.6\cdot 0.2+0.1\cdot 0.8 = 0.12+0.08 = 0.2
\end{aligned}$$

$$\begin{aligned}
P(W=w|Z=w) &= \frac{P(Z=w|W=w)P(W=w)}{P(Z=w)} \\
&= \frac{P(Z=w|W=w)P(W=w)}{P(Z=w|W=w)P(W=w)+P(Z=w|W=f)P(W=f)} \\
&= \frac{0.66\cdot 0.7}{0.66\cdot 0.7+0.2\cdot 0.3} = \frac{0.462}{0.462+0.060} = \frac{0.462}{0.522} \approx 0.885
\end{aligned}$$

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

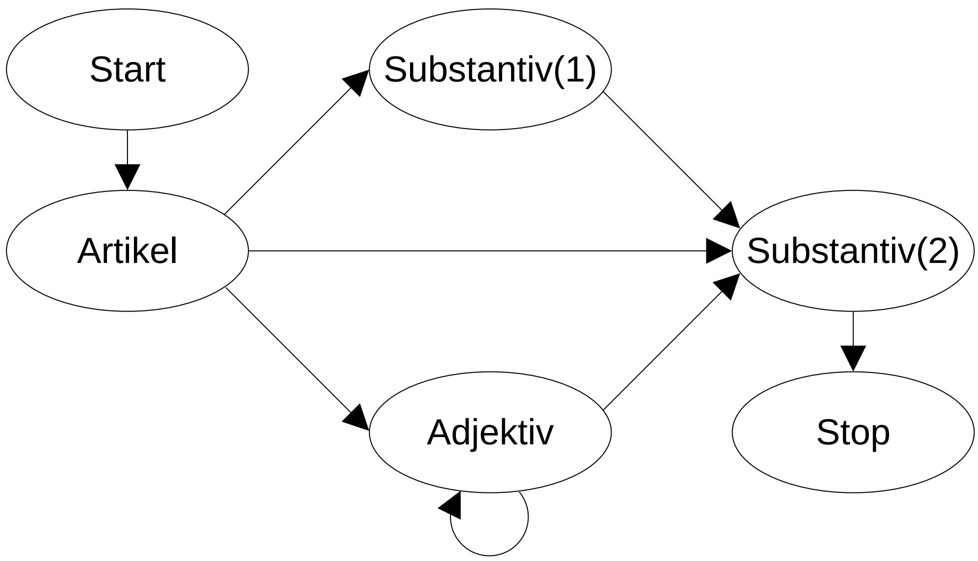
Bei der Spracherkennung werden Wortmodelle eingesetzt, um die Wahrscheinlichkeit von Wortfolgen zu berechnen. Um die Grammatik besser berücksichtigen zu können, wird zwischen einer Folge von Wortarten $x_1, x_2, x_3, \ldots$ und der tatsächlich beobachteten Folge von Wörtern $y_1, y_2, y_3, \ldots$ unterschieden. Ein solches Modell für Nominalgruppen in der englischen Sprache ist in den folgenden zwei Tabellen (stark vereinfacht) angegeben:

x_i	x_{i+1}	$P(x_{i+1}\|x_i)$
Start	Artikel	1.0
Artikel	Adjektiv	0.3
Artikel	Substantiv(1)	0.4
Artikel	Substantiv(2)	0.3
Adjektiv	Adjektiv	0.5
Adjektiv	Substantiv(2)	0.5
Substantiv(1)	Substantiv(2)	1.0
Substantiv(2)	Stop	1.0

x_i	y_i	$P(y_i\|x_i)$
Artikel	the	0.8
Artikel	a	0.2
Adjektiv	small	0.5
Adjektiv	green	0.5
Substantiv(1)	village	0.6
Substantiv(1)	green	0.4
Substantiv(2)	village	0.6
Substantiv(2)	green	0.4

Die Markovkette für die Nominalgruppe beginnt immer mit $x_0 = \mathit{Start}$ und endet mit $x_{k+1} = \mathit{Stop}$. Diesen beiden Zuständen ist keine Ausgabe zugeordnet, da sie Verbindungen zu anderen Teilen des Sprachmodells darstellen. Alle nicht angegebenen Wahrscheinlichkeiten $P(x_{i+1}|x_i)$ und $P(y_i|x_i)$ sind Null.

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



(b) Mit welcher Wahrscheinlichkeit tritt die Nominalgruppe „a small village“ in diesem Modell auf? (6 Punkte)

$$\begin{aligned}
P(\text{„a small village“}) &= P(x_1 = \text{Artikel}|x_0 = \text{Start}) \\
&\cdot\; P(y_1 = \text{„a“}|x_1 = \text{Artikel}) \\
&\cdot\; P(x_2 = \text{Adjektiv}|x_1 = \text{Artikel}) \\
&\cdot\; P(y_2 = \text{„small“}|x_2 = \text{Adjektiv}) \\
&\cdot\; P(x_3 = \text{Substantiv(2)}|x_2 = \text{Adjektiv}) \\
&\cdot\; P(y_3 = \text{„village“}|x_3 = \text{Substantiv(2)}) \\
&\cdot\; P(x_4 = \text{Stop}|x_3 = \text{Substantiv(2)}) \\
&= 1.0 \cdot 0.2 \cdot 0.3 \cdot 0.5 \cdot 0.5 \cdot 0.6 \cdot 1.0 \\
&= 0.009
\end{aligned}$$

(c) Das zweite Wort in der aus drei Wörtern bestehenden Nominalgruppe „the ... green“ wurde nicht erkannt. Wie lautet die wahrscheinlichste Ergänzung? (12 Punkte)

$$\begin{aligned}
P(\text{„the green green“}) &= P(\text{„the green green“}, x_2 = \text{Adjektiv}) \\
&+\; P(\text{„the green green“}, x_2 = \text{Substantiv(1)}) \\
&= 1.0 \cdot 0.8 \cdot 0.3 \cdot 0.5 \cdot 0.5 \cdot 0.4 \cdot 1.0 \\
&+\; 1.0 \cdot 0.8 \cdot 0.4 \cdot 0.4 \cdot 1.0 \cdot 0.4 \cdot 1.0 \\
&= 0.0240 + 0.0512 = 0.0752 \\
P(\text{„the small green“}) &= 1.0 \cdot 0.8 \cdot 0.3 \cdot 0.5 \cdot 0.5 \cdot 0.4 \cdot 1.0 \\
&= 0.0240 \\
P(\text{„the village green“}) &= 1.0 \cdot 0.8 \cdot 0.4 \cdot 0.6 \cdot 1.0 \cdot 0.4 \cdot 1.0 \\
&= 0.0768
\end{aligned}$$

Die Ergänzung der Nominalgruppe zu „the village green“ ist in diesem Sprachmodell am wahrscheinlichsten.

(d) Geben Sie die Wahrscheinlichkeit an, eine Nominalgruppe zu beobachten, die genau k Adjektive ($k = 0, 1, 2, \ldots$) enthält! (4 Punkte)

$$P(k\,\text{Adjektive}) = \begin{cases} 0.7 & \text{für} \quad k = 0 \\ 0.3 \cdot 0.5^k & \text{für} \quad k > 0 \end{cases}$$

Aufgabe 3 – Generatives Modell (24 Punkte)

Ein Freund behauptet, das Ergebnis eines Münzwurfs (Kopf oder Zahl) mittels seiner magischen Kräfte beeinflussen zu können. Zum Beweis wirft er die Münze mehrmals. Der erste Versuch hat das Ergebnis Kopf, aber dann erzielt er 8-mal in Folge das Ergebnis Zahl. Als danach wieder das Ergebnis Kopf kommt, hört er auf. Sie glauben jedoch nicht an übernatürliche Fähigkeiten, sondern daran, dass die Wahrscheinlichkeit für das Ereignis Zahl bei dieser Münze auf $p \neq 0.5$ geändert wurde.

(a) Welche Wahrscheinlichkeitsverteilung hat das Auftreten einer Folge von einen Fehlschlag, k Erfolgen und eines weiteren Fehlschlags in $k+2$ unabhängigen Versuchen, wenn in jedem Versuch die Erfolgswahrscheinlichkeit p beträgt? (4 Punkte)

$$P(k|p) = p^k(1-p)^2$$

(b) Ihr Freund nimmt mit seiner Münze an einem Glücksspiel teil. Sein Einsatz pro Münzwurf beträgt 1 Euro; beim Ereignis Zahl werden 2 Euro ausgezahlt, sonst nichts. Wie hoch ist sein erwarteter Gewinn pro Runde, wenn für die Erfolgswahrscheinlichkeit $p = 0.7$ gilt? (4 Punkte)

$$\langle G \rangle = (2-1) \cdot 0.7 + (0-1) \cdot (1-0.7) - 1 = 0.4$$

(c) Ihr Vorwissen über mögliche Manipulationen an der Münze beschreiben Sie durch die Beta-Verteilung $\mathrm{Beta}(p;\alpha,\beta) = B(\alpha,\beta)\, p^{\alpha-1}(1-p)^{\beta-1}$. $B(\alpha,\beta)$ hängt nicht von p ab. Wie müssen Sie die Hyperparameter α und β wählen, wenn Sie gar kein Vorwissen über p haben, so dass Maximum-a-posteriori- und Maximum-Likelihood-Hypothese übereinstimmen? Begründen Sie! (4 Punkte)

Ohne Vorwissen sollte $\alpha = 1$ und $\beta = 1$ gewählt werden. In diesem Fall gilt

$$\mathrm{Beta}(p;1,1) = 1$$

unabhängig von p, so dass die Maxima von Likelihood $P(k|p)$ und Posterior

$$P(p|k) = CP(k|p)P(p)$$

an der gleichen Stelle liegen. Die Normierungskonstante

$$C = \frac{1}{P(k)}$$

hängt nicht von p ab und hat somit auch keinen Einfluss auf die Lage des Maximums.

(d) Zeigen Sie, dass die Maximum-a-posteriori Hypothese für das Würfelkunststück Ihres Freundes (eine Folge bestehend aus einem Fehlschlag, k Erfolgen und einem weiteren Fehlschlag) durch $p = (k + \alpha - 1)/(k + \alpha + \beta)$ gegeben ist! (8 Punkte)

- Der Logarithmus des Posteriors ist durch

$$\begin{aligned}\log P(p|k) &= \log\left(\frac{P(k|p)\mathrm{Beta}(p;\alpha,\beta)}{P(k)}\right)\\ &= \log C + k\log p + 2\log(1-p)\\ &+ (\alpha-1)\log p + (\beta-1)\log(1-p)\\ &= \log C + (k+\alpha-1)\log p + (1+\beta)\log(1-p)\end{aligned}$$

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

- Berechnung der Ableitung:

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

- Bedingung für ein Maximum

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

(e) Welchen Wert p hat die Maximum-a-posteriori-Hypothese für die Beobachtung $k = 8$, wenn Sie die Hyperparameter auf $\alpha = 3$ und $\beta = 3$ setzen? (4 Punkte)

$$p = \frac{k+\alpha-1}{k+\alpha+\beta} = \frac{8+3-1}{8+3+3} = \frac{10}{14} \approx 0.714$$

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

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

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

e_1	-2	$+1$	$+3$	$+4$
e_2	-1	$+3$	$+4$	-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:

$$\begin{aligned} a_1 &= \text{sgn}(-2-1-1) &= \text{sgn}(-4) &= -1 = y_1 \\ a_2 &= \text{sgn}(1+3-1) &= \text{sgn}(+3) &= +1 = y_2 \\ a_3 &= \text{sgn}(3+4-1) &= \text{sgn}(+6) &= +1 \neq y_3 \\ a_4 &= \text{sgn}(4-1-1) &= \text{sgn}(+2) &= +1 = y_4 \end{aligned}$$

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 $\lambda = 0.2$ als Lernrate und passen Sie auch den Bias w_0 an. (6 Punkte)

Perzeptron-Lernregel:

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

Für die neuen Gewichte gilt:

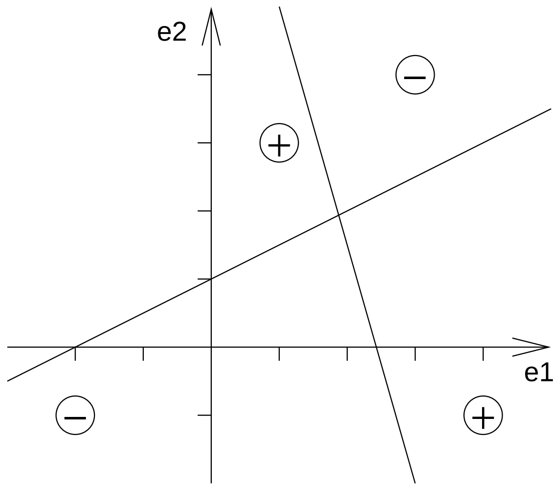
$$\begin{aligned} w_0^+ = 1 + 0.2 \cdot (-1) \cdot (-1) &= 1.2 \\ w_1^+ = 1 + 0.2 \cdot (+3) \cdot (-1) &= 0.4 \\ w_2^+ = 1 + 0.2 \cdot (+4) \cdot (-1) &= 0.2 \end{aligned}$$

Überprüfung von Beispiel 4:

$$a_4^+ = \text{sgn}\,(0.4 \cdot (+4) + 0.2 \cdot (-1) - 1.2) = \text{sgn}(+0.2) = +1 = y_4$$

Es erfolgt keine weitere Anpassung der Gewichte, weil Beispiel 4 immer noch richtig wiedergegeben wird.

(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)

Da die Beispiele nicht linear separabel sind, kann ein Perzeptron die Trainingsmenge nicht korrekt lernen. Es wird immer mindestens ein Beispiel geben, das nicht richtig klassifiziert wird.

(e) Verändern Sie die Lernbarkeit der Trainingsmenge (für ein Perzeptron) durch Hinzufügen oder Weglassen eines Beispiels! (2 Punkte)

Wenn eines der Beispiele weggelassen wird, kann das Perzeptron die restliche Trainingsmenge lernen. Bereits der Startwert der Gewichte führt zu einer korrekten Klassifikation, wenn speziell Beispiel 3 entfernt wird.