Zufallsvariablen

Stochastische Variable oder *Zufallsvariable* X:

Abbildung eines Ereignisraumes S auf reelle Zahlen.

Das Ereignis $\{E_j \mid X(E_j) = x_j\}$ wird mit $X = x_j$ bezeichnet.

diskrete Zufallsvariablen

Sei P eine Wahrscheinlichkeitsverteilung über S, X eine Zufallsvariable über S, die die Werte $x_1, .., x_n$ annimmt.

Dann ist die Abbildung

$$f : S \to [0, 1] \text{ mit } f(x_i) = P(X = x_i)$$

die *Wahrscheinlichkeitsverteilung* von X.

Erwartungswert einer diskreten Zufallsvariablen (entspricht arithmetischem Mittel):

$$\mu_X = E(X) = \sum_{x \in S} x \cdot f(x) = \sum_{i=1}^{n} x_i \cdot f(x_i)$$

Varianz σ^2:

$$\sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot f(x_i)$$

Zufallsvariablen

stetige Zufallsvariablen

Die Wahrscheinlichkeit, dass der Wert einer stetigen Zufallsvariablen X genau eine reelle Zahl x annimmt, geht gegen 0.

Deshalb wird die summierte Wahrscheinlichkeit von $X \leq x$ betrachtet:

$$F(x) = P(\{u \mid X(u) \leq x\}) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

$F : R \rightarrow [0, 1]$ heisst auch Verteilungsfunktion

(und f ist die Ableitung davon).

Erwartungswert:

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

Varianz σ^2:

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx$$

Zufallsvariablen

mehrdimensionale Zufallsvariablen

Sei S ein Ereignisraum, P eine W'keitsverteilung darüber, X und Y Zufallsvariablen darüber mit den Werten $x_1, .., x_n$ bzw. $y_1, .., y_m$.

Dann ist die Abbildung

$$(x_i, y_j) \rightarrow P(X = x_i, Y = y_j)$$

die *gemeinsame Wahrscheinlichkeitsverteilung* oder -funktion von X und Y.

Gilt für alle (x_i, y_j):

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j),$$

dann sind X und Y *unabhängig*.

spezielle W'keitsverteilungen

Binomialverteilung

Bernoulli-Experimente:

Experimente mit nur zwei Ausgängen.

Bernoulli-Variable:

Zufallsvariable, bei der der eine Ausgang den Wert 0, der andere den Wert 1 erhält.

Bernoullische Formel:

Sei p die Wahrscheinlichkeit für den Ausgang 1 eines B.-Experiments. Dann ist die W'keit, dass bei n Ausführungen k-mal der Ausgang 1 eintritt:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Sei X eine Zufallsvariable, die die Werte 0, 1, 2, ..., n annehmen kann, mit $X(k) = \binom{n}{k} p^k (1-p)^{n-k} = B_{n,p}(k)$, heisst X *binomialverteilt* mit Parametern n und p.

Es gilt: $E(X) = np, v(X) = np(1-p)$.

Binomialverteilung ist abhängig von der Anzahl der Versuche! Betrachtung von $n \to \infty$: Verschiebung des Erwartungswerts auf 0, Streckung von k um $1/\sigma$ und von $B_{n,p}(k)$ um σ.

Gauss-Funktionen, Normalverteilung

die eulersche Zahl: $e = \sum_{n=0}^{\infty} \frac{1}{n!}$

Gauss-Funktion: $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}x^2}$

Graph von φ ist eine Glockenkurve...

Näherung für Binomialverteilung für $n > \frac{9}{p(1-p)}$

$$B_{n,p}(k) \approx \frac{1}{\sigma}\varphi\left(\frac{k-\mu}{\sigma}\right)$$

mit $\mu = np$ und $\sigma = \sqrt{np(1-p)}$

Gauss'sche Summenfunktion: $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$

Sei X $B_{n,p}$-verteilt, dann gilt für genügend grosses n:

$$P(X \leq k) \approx \Phi(\frac{k-\mu}{\sigma}).$$

Normalverteilung

Zufallsvariable X mit $P(X \leq x) = \Phi(\frac{x-\mu}{\sigma}),$

für alle reellen Zahlen x heisst X *normalverteilt* mit Erwartungswert μ und Varianz σ^2 ($N(\mu, \sigma)$-verteilt).

Stochastische Prozesse

Ein *stochastischer* oder *Zufallsprozess* ist eine Folge von Zufallsvariablen $X_1, X_2, ...$ über demselben Ereignisraum.

Die möglichen Ausgänge heissen auch *Zustände* des Prozesses, der Prozess ist im Zustand x_t zum Zeitpunkt t.

Die X_i sind nicht zwingend unabhängig voneinander!

Zufallsprozesse können über diskrete oder stetige Zeitparameter und über diskrete oder stetige Zufallsvariablen betrachtet werden, hier aber nur diskrete Zeitschritte und endliche Ausgangsmengen.

Vollständige Charakterisierung eines Zufallsprozesses:

Wahrscheinlichkeit $P(X_1 = x_j)$ für alle Ausgänge x_j für den Anfangszustand.

für jeden folgenden Zustand $X_{t+1}, t = 1, 2, ...$ die bedingten W'keiten $P(X_{t+1} = x_{i_{t+1}} \mid X_1 = x_{i_1}, .., X_t = x_{i_t})$

N-Gram-Modelle

Annahme: nur die letzten $n-1$ Wörter haben Einfluss auf die Wahrscheinlichkeit des nächsten. Gebräuchlich ist $n = 3$: Trigram-Modelle.

Wahrscheinlichkeit für ein Wort w_n nach der Wortfolge $w_{1,n-1}$:

$$P(w_n \mid w_{1,n-1}) = P(w_n \mid w_{n-2}, w_{n-1})$$

Die Wahrscheinlichkeit für eine Wortfolge $w_{1,n}$ beträgt dann:

$$\begin{aligned} P(w_{1,n}) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)..P(w_n \mid w_{1,n-1}) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)..P(w_n \mid w_{n-2,n-1}) \\ &= P(w_1)P(w_2 \mid w_1)\prod_{i=3}^{n} P(w_i \mid w_{i-2,i-1}) \end{aligned}$$

$$P(w_{1,n}) = \prod_{i=3}^{n} P(w_i \mid w_{i-2,i-1})$$

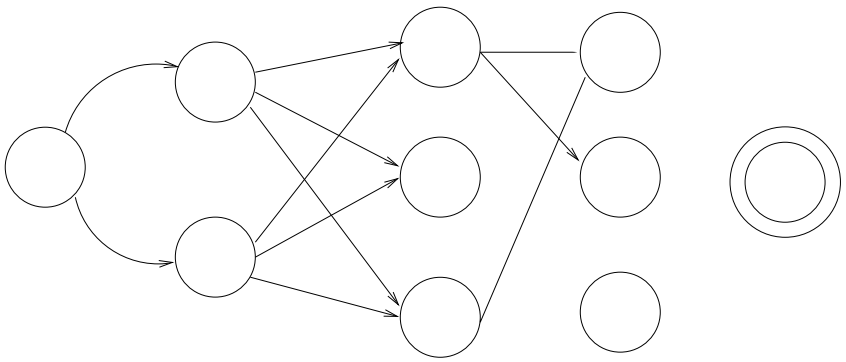
Markov-Kette

Eine *Markov-Kette* ist ein Zufallsprozess, bei dem die Wahrscheinlichkeit des nächsten Zustands nur vom aktuellen abhängt.

Die Markov-Eigenschaft ist also:

$$P(X_{t+1} = x_{i_{t+1}} \mid X_1 = x_{i_1}, .., X_t = x_{i_t}) \\ = P(X_{t+1} = x_{i_{t+1}} \mid X_t = x_{i_t})$$

Beispiel:



Stochastische Matrix

Sei eine (*endliche*) Markov-Kette mit n Zuständen gegeben.

Die *Übergangswahrscheinlichkeiten* von Zustand s_i in s_j, d.h. $P(X_{t+1} = s_j \mid X_t = s_i) = p_{ij}$ können in einer Übergangsmatrix dargestellt werden:

$$\mathbf{P} = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \cdots & \cdots & \cdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix}, 0 \leq p_{ij} \leq 1, \sum_{j=1}^{n} p_{ij} = 1 \text{ für } i = 1, 2, .., n$$

Ein Vektor $\mathbf{v} = [v_1, .., v_n]$ mit $1 \geq v_i \geq 0$ und $\Sigma_{i=1}^{n} v_i = 1$ heisst Wahrscheinlichkeitsvektor, und kann z.B. für den ersten Zustand einer Markov-Kette gelten. Dann gilt: $v_i = P(X_i = s_i), i = 1, .., n$.

Der initiale Wahrscheinlichkeitsvektor zusammen mit der Übergangsmatrix bestimmen eine Markov-Kette vollständig, d.h. die Wahrscheinlichkeiten, dass sich der Prozess an einem best. Zeitpunkt t in einem best. Zustand s_i befindet, können daraus errechnet werden:

$$[p^{(t)}(s_1), .., p^{(t)}(s_n)] = \mathbf{v}\mathbf{P}^{t-1}$$

Beispielmatrix

Für das Beispiel zur Markov-Kette kann man folgende Zustandsübergangs-Wahrscheinlichkeits-Matrix aufstellen:

$$\begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.34 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.34 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0.33 & 0.34 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01 & 0.01 & 0.98 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.49 & 0.49 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Matrix-Multiplikation

Wir brauchen nur quadratische, d.h. $(n \times n)$ Matrizen.

Seien $\mathbf{A}$ und $\mathbf{B}$ $(n \times n)$ Matrizen mit den Elementen a_{ij} und b_{ij}, $i, j = 1, .., n$, i die Zeile, j die Spalte.

Dann ist das Produkt $\mathbf{A} \cdot \mathbf{B}$ definiert als $(n \times n)$ Matrix $\mathbf{C}$ mit

$$c_{ij} = \sum_{k=1}^{n} (a_{ik} \cdot b_{kj}).$$

(Das ist das Produkt aus dem i-ten Zeilenvektor $z_i = [a_{i1}, .., a_{in}]$ und dem j-ten Spaltenvektor $s_j = [a_{j1}, .., a_{jn}]$)

Produkt aus $(n \times n)$ Matrix $\mathbf{A}$ und Vektor x (n-stellig):

$$\mathbf{A}x = \begin{bmatrix} z_1 x \\ \vdots \\ z_n x \end{bmatrix} = \begin{bmatrix} \Sigma_{k=1}^{n} a_{1k} x_k \\ \vdots \\ \Sigma_{k=1}^{n} a_{nk} x_k \end{bmatrix}$$

Markov-Modelle

Sei jeder Zustand einer Markov-Kette mit einer endlichen Menge von *Signalen* verbunden.

Nach jedem Zustandsübergang wird eines der zum aktuellen Zustand gehörenden Signale ausgegeben. Die Zufallsvariable η_t repräsentiert dieses Signal zum Zeitpunkt t.

Ein *Markov-Modell* (erster Ordnung) besteht aus:

- einer endliche Menge von Zuständen $\omega = \{s_1, .., s_n\}$
- einem Signal-Alphabet $\Sigma = \{\sigma_1, .., \sigma_m\}$
- einer $(n \times n)$-Zustandsübergangs-Matrix $\mathbf{P} = [p_{ij}]$ mit $p_{ij} = P(X_{t+1} = s_j \mid X_t = si)$
- einer $(n \times m)$-Signal-Matrix $\mathbf{A} = [a_{ij}]$ mit der Wahrscheinlichkeit $a_{ij} = p(\eta_t = \sigma_j \mid X_t = s_i)$ für jedes Zustands-Signal-Paar, dass σ_j im Zustand s_i ausgegeben wird.
- und einem initialer Vektor $\mathbf{v} = [v_1, .., v_n]$ mit $v_i = P(X_1 = s_i)$

Sei $p^{(t)}(\sigma_j)$ die Wahrscheinlichkeit, dass zur Zeit t das Signal σ_j ausgegeben wird. Der Vektor

$$[p^{(t)}(\sigma_1), .., p^{(t)}(\sigma_m)] = \mathbf{v}\mathbf{P}^{t-1}\mathbf{A}$$

enthält diese Wahrscheinlichkeiten für alle $\sigma \in \Sigma$.

Hidden Markov Models, HMMs

Wenn keine Beobachtung der Zustände möglich ist, sondern nur die Signale beobachtet werden können, liegt ein *Hidden Markov Model* (HMM) vor.

Sei $\mathbf{O} \in \Sigma^*$ eine Folge von beobachteten Signalen und $\mathbf{S} \in S^*$ die unbekannte Folge von Zuständen.

Die beste Schätzung für **S** ist die Folge mit dem grössten Wert für $P(\mathbf{S} \mid \mathbf{O})$

Laut Bayes'schem Satz gilt:

$$P(\mathbf{S} \mid \mathbf{O}) = \frac{P(\mathbf{O} \mid \mathbf{S}) \cdot P(\mathbf{S})}{P(\mathbf{O})}$$

und da $P(\mathbf{O})$ nicht von $\mathbf{S}$ abhängt, können wir auch $P(\mathbf{O} \mid \mathbf{S}) \cdot P(\mathbf{S})$ maximieren.

$P(\mathbf{O} \mid \mathbf{S})$ heisst Signalmodell, $P(\mathbf{S})$ Sprachmodell.

Anwendungen für HMMs

1. Schätzung der Wahrscheinlichkeit einer Signalfolge (Identifikation einer Sprache), $P(\mathbf{O})$
2. Bestimmung der wahrscheinlichsten Zustandsfolge, die zu einer Signalfolge geführt hat:

   Tagging

   Signale: Wörter eines Eingabetextes

   Zustände: Mengen von Wortarten

   Aufgabe: finde die wahrscheinlichste Folge von Wortartmengen, die den Wörtern zugeordnet werden können.

   Spracherkennung

   Signale: (Repräsentation der) akustischen Signale

   Zustände: mögliche Wörter

   Aufgabe: finde die wahrscheinlichste Folge von Wörtern, die die akustischen Signale hervorgerufen haben

3. Bestimmung der Parameter $\mathbf{P}, \mathbf{A}, \mathbf{v}$

1. $P(\mathbf{O})$

Sei $\mathbf{O} = (\sigma_{k_1}, .., \sigma_{k_T}), \mathbf{S} = (s_{i_1}, .., s_{i_T})$.

Dann:

$$\begin{aligned} P(\mathbf{O} \mid \mathbf{S}) &= \Pi_{t=1}^{T} P(\eta_t = \sigma_{k_t} \mid X_t = s_{i_t}) = \Pi_{t=1}^{T} a_{i_t k_t} \\ P(\mathbf{S}) &= v_{i_1} \cdot \Pi_{t=1}^{T} p_{i_{t-1} i_t} \\ P(\mathbf{O} \cap \mathbf{S}) &= P(\mathbf{O} \mid \mathbf{S}) \cdot P(\mathbf{S}) \\ &= (\Pi_{t=1}^{T} a_{i_t k_t}) \cdot (\Pi_{t=1}^{T} p_{i_{t-1} i_t}) \\ &= a_{i_1 k_1} \cdot v_{i_1} \cdot \Pi_{t=2}^{T} p_{i_{t-1} i_t} a_{i_t k_t} \end{aligned}$$

und:

$$P(\mathbf{O}) = \sum_{\mathbf{S}} P(\mathbf{O} \cap \mathbf{S})$$

und das ist viel zu aufwendig!

wie aufwendig?

$O(2Tn^T)$

Der Vorwärts-Algorithmus

Vorwärts-Variablen:

$$\alpha_t(i) = P(\mathbf{O}_{\leq t}; X_t = s_i)$$
$$= P(\eta_1 = \sigma_{k_1}, .., \eta_t = \sigma_{k_t}; X_t = s_i).$$

$$P(\mathbf{O}) = \sum_{i=1}^{n} P(\eta_1 = \sigma_{k_1}, .., \eta_T = \sigma_{k_T}; X_T = s_i) = \sum_{i=1}^{n} \alpha_T(i)$$

$$\alpha_1(i) = a_{ik_1} \cdot v_i \text{ und } \alpha_{t+1}(i) = (\sum_{i=1}^{n} \alpha_t(i) \cdot p_{ij}) \cdot a_{jk_{t+1}}$$

Begründung (Markov-Annahme im zweiten Schritt):

$$P(\mathbf{O}_{\leq t+1}; X_{t+1} = s_i) =$$
$$= \Sigma_{i=1}^{n} P(\mathbf{O}_{\leq t}; X_t = s_i)$$
$$\cdot P(\eta_{t+1} = \sigma_{k_{t+1}}; X_{t+1} = s_j \mid \mathbf{O}_{\leq t}; X_t = s_i)$$
$$= \Sigma_{i=1}^{n} P(\mathbf{O}_{\leq t}; X_t = s_i)$$
$$\cdot P(\eta_{t+1} = \sigma_{k_{t+1}} \mid X_{t+1} = s_j) \cdot P(X_{t+1} = s_j \mid X_t = s_i)$$

Aufwand: $O(n^2 T)$

Der Rückwärts-Algorithmus

Rückwärts-Variablen:

$$\beta_t(i) = P(\mathbf{O}_{>t}; X_t = s_i)$$
$$= P(\eta_{t+1} = \sigma_{k_{t+1}}, .., \eta_T = \sigma_{k_T}; X_t = s_i).$$

$$P(\mathbf{O}) = \sum_{i=1}^{n} P(\eta_1 = \sigma_{k_1}, X_1 = s_i) \cdot P(\eta_2, .., \eta_T = \sigma_{k_T}; X_1 = s_i)$$
$$= \sum_{i=1}^{n} a_{ik_1} \cdot v_i \cdot \beta_1(i)$$

Definiere $\beta_T(i) = 1$ für $i = 1, .., n$.

$$\beta_t(i) = \sum_{j=1}^{n} p_{ij} \cdot a_{jk_{t+1}} \cdot \beta_{t+1}(j)$$

weil:

$$P(\mathbf{O}_{>t} \mid X_{t+1} = s_i) =$$

$$= \Sigma_{j=1}^{n} P(\mathbf{O}_{>t}; X_{t+1} = s_j \mid X_t = s_i)$$
$$= \Sigma_{j=1}^{n} P(\mathbf{O}_{>t} \mid X_t = s_i; X_{t+1} = s_j)$$
$$\cdot P(X_{t+1} = s_j \mid X_t = s_i)$$

$$= \Sigma_{j=1}^{n} P(\eta_{t+1} = \sigma_{k_{t+1}} \mid X_{t+1} = s_j) \cdot P(\mathbf{O}_{>t} \mid X_{t+1} = s_j)$$
$$\cdot P(X_{t+1} = s_j \mid X_t = s_i)$$

Der Vorwärts-Rückwärts-Algorithmus

$$\begin{aligned} P(\mathbf{O}) &= \Sigma_{i=1}^{n} P(\mathbf{O}; X_t = s_i) \\ &= \Sigma_{i=1}^{n} P(\mathbf{O}_{\leq t}; X_t = s_i) \cdot P(\mathbf{O}_{>t} \mid \mathbf{O}_{\leq t}; X_t = s_i) \\ &= \Sigma_{i=1}^{n} P(\mathbf{O}_{\leq t}; X_t = s_i) \cdot P(\mathbf{O}_{>t} \mid X_t = s_i) \\ &= \Sigma_{i=1}^{n} \alpha_t(i)\beta_t(i) \end{aligned}$$

Wahrscheinlichkeit, zum Zeitpunkt t im Zustand s_i zu sein, wenn **O** die gesamte beobachtete Sequenz von Zeit 1 bis T ist:

Vorwärts-Rückwärts-Variablen:

$$\gamma_t(i) = P(X_t = s_i \mid \mathbf{O}) = \frac{P(\mathbf{O}; X_t = s_i)}{P(\mathbf{O})} = \frac{\alpha_t(i)\beta_t(i)}{\Sigma_{i=1}^{n} \alpha_t(i)\beta_t(i)}$$