Fontes de informação sem memória

- Numa *fonte de informação sem memória*, não há correlação entre os símbolos emitidos em instantes diferentes.
- Em cada instante, há uma distribuição de probabilidade descrevendo a probabilidade de ser emitido cada símbolo nesse instante.
- Se essa distribuição for independente do instante, a distribuição diz-se *estacionária*.
- Por exemplo, atirar repetidas vezes uma moeda ao ar. Trata-se de uma fonte de informação sem memória com $P(H) = P(T) = 1/2$.
  Calculamos anteriormente $H(P) = 1$ (bit), ou seja a quantidade informação é 1 bit/símbolo.

- Alternando atirar moeda ao ar com lançar um dado com quatro das faces indicando H e as outras duas T, obtemos um processo gerador de sequências aleatórias sem memória mas não estacionário pois as distribuições de probabilidades alternam entre $P(H) = P(T) = 1/2$ e $P(H) = 2/3$, $P(T) = 1/3$.

  Aqui a entropia alterna entre 1 bit/símbolo e $\simeq 0.918$ bits/símbolo pelo que a quantidade de informação na sequência é a média das duas, nomeadamente $\simeq 0.959$ bits/símbolo.
- Em geral, a probabilidade de ser emitido um dado símbolo num certo instante, depende dos símbolos emitidos anteriormente, ou seja a fonte de informação tem memória do que acaba de ser emitido,
  sendo um modelo útil considerar o caso em que essa memória se reduz a um número de passos anteriores fixado.

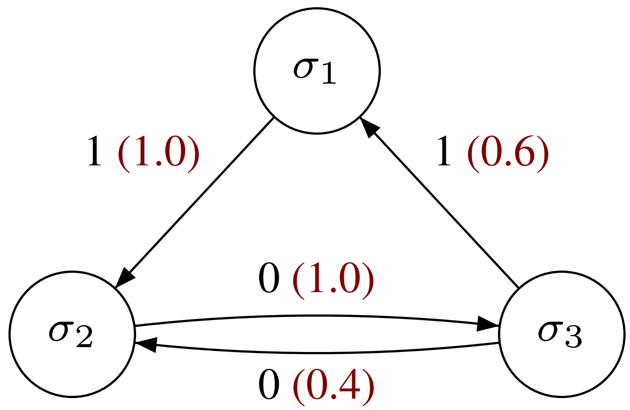
Fontes de Markov

Uma *fonte de Markov* consiste de:

- um alfabeto A,
- um conjunto de estados Σ,
- um conjunto de transições entre estados,
- para cada transição entre estados, uma etiqueta dada por um símbolo de A;
- uma distribuição de probabilidades nos estados, que representa a probabilidade de começar uma sucessão num dado estado,
- uma distribuição de probabilidades nas transições: $P(j|i)$ representa a probabilidade de seguir a transição do estado σ_i para o estado σ_j.

Nota: assume-se implicitamente que as distribuições são estacionárias.

Exemplo



- Para gerar uma sequência, começamos com probabilidade $P(\sigma_i)$ no estado σ_i e seguimos cada transição $\sigma_i \to \sigma_j$ com a probabilidade $P(j|i)$, emitindo nesse caso o símbolo que etiqueta essa transição.

n-gramas

- Num *n-grama*, são dadas as probabilidades de ocorrência das sequências de símbolos (*palavras*) de comprimento n.
- Uma *fonte de Markov de ordem m* é uma fonte de Markov tal que:
  - os estados são palavras w de comprimento $|w| = m$;
  - se há uma transição da palavra u para a palavra v, então v é obtida de u cortando a primeira letra e acrescentando uma letra no fim, i.e., $u = aw$ e $v = wb$;

  A probabilidade da transição $aw \to wb$ representa-se então por $P(b|aw)$.

Dados um m-grama e um $m+1$-grama, temos uma fonte de Markov associada:

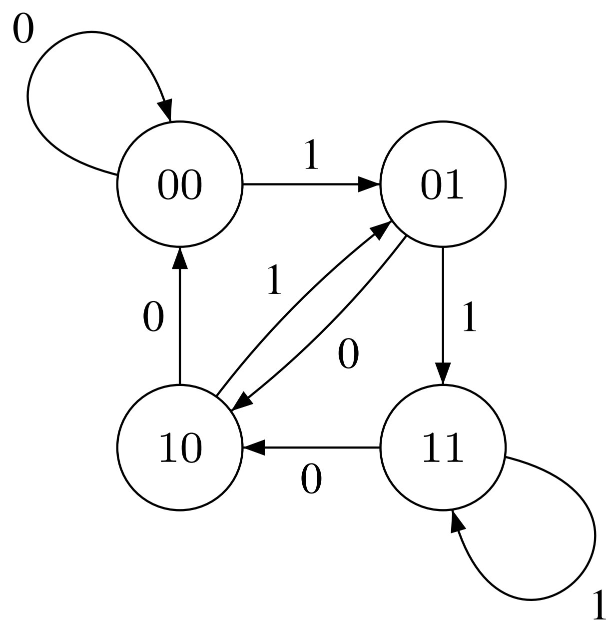
- os estados são todas as palavras $w = s_1 s_2 \cdots s_m$, de comprimento m, com probabilidades $P(s_1, s_2, \ldots, s_m)$ dadas pelo m-grama;
- as transições são todas as da forma $aw \to wb$, onde $|aw| = |wb| = m$, com etiqueta b e probabilidade

$$P(s|s_1, \ldots, s_m) = \frac{P(s_1, \ldots, s_m, s)}{P(s_1, \ldots, s_m)}$$

dadas pelo $m+1$-grama, sendo $a = s_1$, $w = s_2 \cdots s_m$, $b = s$.

Exemplo

O seguinte diagrama, juntamente com um 2-grama e um 3-grama, representa uma fonte de Markov de ordem 2:



Matriz de transição de probabilidades

- Para um conjunto de estados $\Sigma = \{\sigma_1, \ldots, \sigma_N\}$, a *matriz de transição* de probabilidades é a matriz $N \times N$ dada por

$$\Pi = [P(i|j)]_{i,j=1,\ldots,N} = \begin{bmatrix} P(1|1) & P(1|2) & \cdots & P(1|N) \\ P(2|1) & P(2|2) & \cdots & P(2|N) \\ \vdots & \vdots & \ddots & \vdots \\ P(N|1) & P(N|2) & \cdots & P(N|N) \end{bmatrix}$$

- Representando por w_i^t a probabilidade da fonte de Markov se encontrar no estado i no instante t e fazendo

$$W^t = \begin{bmatrix} w_1^t \\ w_2^t \\ \vdots \\ w_N^t \end{bmatrix}$$

obtemos $W^{t+1} = \Pi W^t$ e, por indução, $W^t = \Pi^t W^0$.

- Note-se que em todas as matrizes acima as colunas representam distribuições de probabilidade e, portanto as somas das entradas em cada coluna são 1 (para todo o t).

Distribuições estacionárias

- Pode-se perguntar qual é a evolução ao longo do tempo da distribuição W^t, nomeadamente qual é o seu comportamento assintótico.
- Por exemplo, considere-se uma fonte de Markov com matriz de transição $\Pi = \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix}$ e distribuição inicial $W^0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$.

  Fazendo as contas, obtém-se, sucessivamente para $W^1, W^2, \ldots$:

$$\begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \begin{bmatrix} 0.36 \\ 0.64 \end{bmatrix}, \begin{bmatrix} 0.3444 \\ 0.656 \end{bmatrix}, \begin{bmatrix} 0.3376 \\ 0.6624 \end{bmatrix}, \ldots$$

- Facilmente se adivinha que $W^t \rightarrow \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}$ matriz que denotamos por W_s^0.
- Note-se que

$$W_s^1 = \Pi W_s^0 = \begin{bmatrix} 0.6 & 0.2 \\ 0.4 & 0.8 \end{bmatrix} \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 2/3 \end{bmatrix}$$

ou seja $W_s^t = W_s^0$ para todo o t, o que significa que a distribuição W_s^0 permanece inalterada ao longo do tempo.

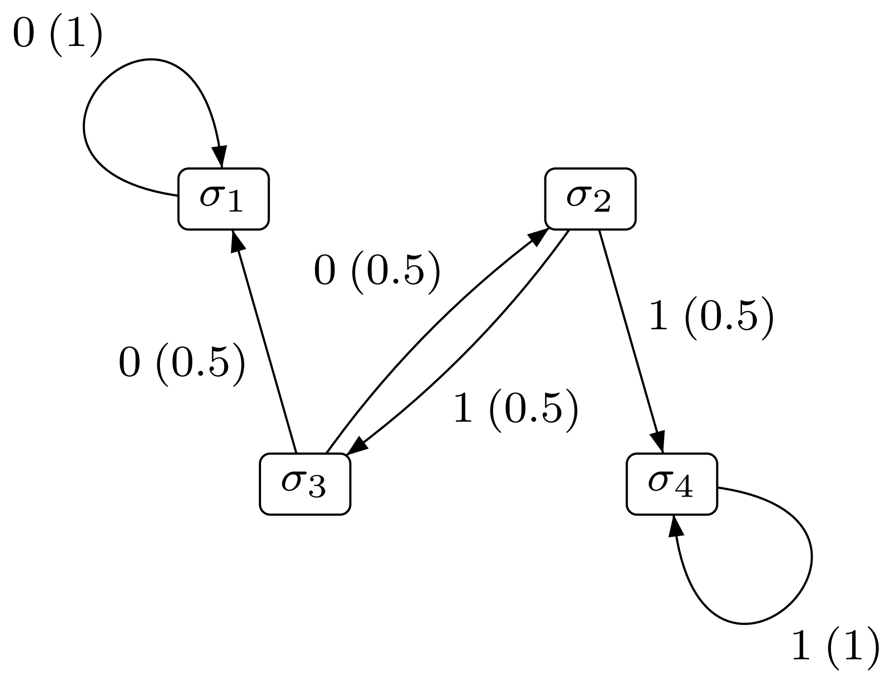
Distribuição estacionária

- Uma *distribuição estacionária* W sobre os estados de uma fonte de Markov com matriz de transição Π é uma distribuição de probabilidade W tal que $\Pi W = W$.
  Por outras palavras, é um vetor próprio para o valor próprio 1 da matriz Π cujas entradas são não negativas e têm soma 1.

Exemplo

- Uma fonte de Markov pode ter mais do que uma distribuição estacionária. Por exemplo, a fonte de Markov com 4 estados e matriz de transição

$$\Pi = \begin{bmatrix} 1 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 1 \end{bmatrix}$$



- As distribuições estacionárias desta fonte de Markov são $(1, 0, 0, 0)$ e $(0, 0, 0, 1)$.

- A matriz Π duma fonte de Markov tem entradas não negativas. Um tal matriz real diz-se uma *matriz não negativa* e diz-se uma *matriz positiva* se todas as entradas forem positivas.
- Uma matriz não negativa A diz-se *regular* ou *primitiva* se existir k tal que A^k é uma matriz positiva.
- Note-se que, para uma *matriz estocástica* (matriz quadrada cujas colunas são distribuições de probabilidades), o vetor constante $(1, \ldots, 1)$ é um vetor próprio à esquerda pertencente ao valor próprio 1. Logo uma tal matriz também admite algum vetor próprio à direita pertencente ao valor próprio 1.

- Por outro lado, sendo $\|A\| = \max_i \sum_j |a_{i,j}|$, é fácil verificar que se tem $\|Ax\| \leqslant \|A\| \, \|x\|$. Combinando com a igualdade $\|\lambda x\| = |\lambda| \, \|x\|$, decorre que, se x é um vetor próprio pertencente ao valor próprio λ, então $|\lambda| \leqslant \|A\|$.
- Aplicando o argumento anterior à transposta duma matriz estocástica A, concluímos que todos os seus valores próprios (em $\mathbb{C}$) têm módulo quando muito 1.

Teorema 1.8 (Perron-Frobenius) *Seja A uma matriz real não negativa. Então A tem um valor próprio real positivo λ_{PF} tal que:*

(a) existem vetores próprios reais não negativos à esquerda e à direita;

(b) $|\lambda| \leqslant \lambda_{PF}$ para todo o valor próprio λ de A.

No caso da matriz A ser regular, λ_{PF} é um valor próprio simples, que admite vetores próprios à esquerda e à direita positivos, e a igualdade em (b) só se verifica para $\lambda = \lambda_{PF}$.

- Já observámos atrás que, para uma matriz estocástica, $\lambda_{PF} = 1$. Usando o Teorema de Perron-Frobenius, deduzimos que uma tal matriz admite alguma distribuição estacionária.
- No caso duma matriz estocástica regular, o Teorema de Perron-Frobenius garante que há uma única distribuição estacionária, sendo as suas componentes positivas.

- Uma fonte de Markov diz-se *ergódica* se, para toda a sucessão de símbolos $s_1 s_2 \ldots$ por ela gerada e para toda a palavra u, a frequência de ocorrência de u nos segmentos iniciais finitos $s_1 s_2 \ldots s_n$ da sucessão ($\frac{1}{n}\#\{i \leqslant n - |u| + 1 : \; s_i s_{i+1} \ldots s_{i+|u|-1} = u\}$) tende para um dado valor $m(u)$, quando $n \to \infty$, valor esse que é independente da sucessão gerada pela fonte.
- ***Exercício:*** mostre que a fonte do exemplo da página 60 não é ergódica.

Entropia das fontes da Markov

- A *entropia* do i-ésimo estado de uma fonte de Markov é a entropia da distribuição de probabilidades no conjunto das transições a partir desse estado.

$$H(P_i) = -\sum_{j=1}^{N} P(j|i)\log(P(j|i)).$$

- Uma fonte de Markov diz-se *unifilar* se, para cada estado, transições distintas que saem desse estado têm etiquetas distintas.

- Suponhamos que temos uma fonte de Markov unifilar M que possui alguma distribuição estacionária $W = (w_1, \ldots, w_N)$ nos estados.
  A *entropia* de M (relativamente a W) é o valor esperado da entropia dos estados para a distribuição de probabilidades W, ou seja, é dada por

$$H(M) = \sum_{i=1}^{N} w_i H(P_i) = -\sum_{i=1}^{N} \sum_{j=1}^{N} w_i P(j|i) \log(P(j|i)).$$

Sucessões de símbolos

Os teoremas seguintes, que não iremos provar, pelo menos por agora, são válidos para fontes de Markov ergódicas.

Teorema 1.9 *Dados $\varepsilon > 0$ e $\delta > 0$, existe algum $N_0 > 0$ tal que todas as sequências de comprimento $N \geqslant N_0$ são de um dos seguintes tipos:*

- *um conjunto de sequências cuja probabilidade global é menor do que ε;*
- *as restantes sequências, para as quais a probabilidade p satisfaz a desigualdade*

$$\left| \frac{\log(1/p)}{N} - H \right| < \delta$$

onde H é a entropia da fonte.

Para um alfabeto A, seja A^N o conjunto de todas as palavras de comprimento N formadas por letras de A.

Teorema 1.10 *Seja M uma fonte de Markov com alfabeto $A = \{a_1, a_2, \ldots, a_n\}$ e entropia H.*

Para cada $s \in A^N$ seja $P(s)$ a probabilidade da palavra s ser emitida pela fonte. Defina-se

$$G_N = -\frac{1}{N} \sum_{s \in A^N} P(s) \log(P(s)),$$

a entropia por símbolo das palavras de N símbolos. Então G_N é uma função decrescente de N e $\lim_{N \to \infty} G_N = H$.

Teorema 1.11 *Seja M uma fonte de Markov com alfabeto $A = \{a_1, a_2, \ldots, a_n\}$ e entropia H.*
Para cada $s \in A^{N-1}$, seja $P(sa_i)$ a probabilidade da fonte emitir a palavra sa_i e seja $P(a_i|s)$ a probabilidade condicional da fonte emitir o símbolo a_i depois de ter acabado de emitir a palavra s.
Seja

$$F_N = -\sum_{s \in A^{N-1}} \sum_{i=1}^{N} P(sa_i) \log(P(a_i|s)),$$

a entropia condicional do próximo símbolo conhecidos os últimos $N-1$ símbolos.
Então F_N é uma função decrescente de N e $\lim_{N\to\infty} F_N = H$.

Teorema 1.12 *Sendo F_N e G_N definidas como acima, tem-se*

$$F_N = NG_N - (N-1)G_{N-1},$$

$$G_N = \frac{1}{N}\sum_{I=1}^{N} F_I,$$

$$F_N \leqslant G_N.$$

- Estes resultados mostram que o comportamento estatístico de palavras emitidas pela fonte de Markov, de comprimento suficientemente grande, fornecem aproximações arbitrariamente boas da entropia.
- A sucessão de aproximações F_N (que satisfaz $F_N \leqslant G_N$) fornece melhores aproximações que a sucessão G_N.
- No caso da dependência da emissão de um dado símbolo se restringir a, quando muito, os últimos $N_D - 1$ símbolos (memória limitada),
  então $F_{N_D} = H$.
  (Pois $P(a_i|ts) = P(a_i|s)$ se $|s| = N_D - 1$.)