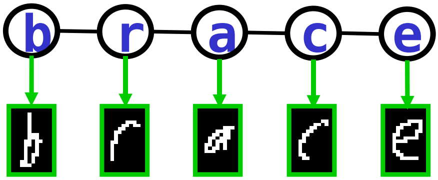
Hidden Markov Models

Eric Moulines[1], Guillaume Obozinski[2]
Florence d'Alché-Buc[1]

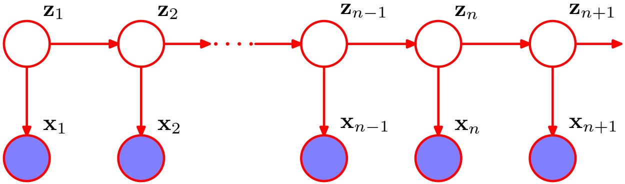
[1]Telecom ParisTech, [2]Ecole des Ponts - ParisTech

Hidden Markov models (HMM)

- Speech recognition
- Natural language processing
- Customer churn detection, trajectory-based activity recognition
- Optical Character Recognition
- Computational biology: biological sequences (proteins, DNA)



Hidden Markov Model(HMM)



$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \prod_{n=2}^{N} p(\mathbf{z}_n | \mathbf{z}_{n-1}) \prod_{n=1} p(\mathbf{x}_n | \mathbf{z}_n)$$

Homogeneous Markov chain

- $\mathbf{z}_n \in \{1, \ldots, K\}$
- Homogeneous Markov chain: $\forall n,\ p(\mathbf{z}_n | \mathbf{z}_{n-1}) = p(\mathbf{z}_2 | \mathbf{z}_1)$
- $\mathbf{x}_n$ emitted symbol ($\{0, 1\}^K$) / observation ($\mathbb{R}^d$)

Hidden Markov Model (HMM)

Parametrization

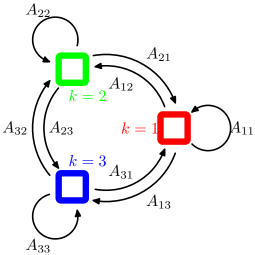
distribution of initial state	$p(\mathbf{z}_1 = k; \pi) = \pi_k$
	with $\sum_{k=1}^{K} \pi_k = 1$
transition matrix	$p(\mathbf{z}_n = j \vert \mathbf{z}_{n-1} = k; A) = A_{jk}$
	with $\sum_{k=1}^{K} A_{jk} = 1$
emission probabilities	$p(\mathbf{x}_n \vert \mathbf{z}_n; \Phi)$
	here a Gaussian distribution
	(in general a mixture model)

Hidden Markov Model (HMM)

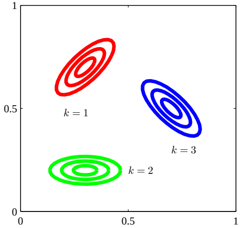
Parametrization

distribution of initial state	$p(\mathbf{z}_1; \pi)$
transition matrix	$p(\mathbf{z}_n \mid \mathbf{z}_{n-1}; A)$
emission probabilities	$p(\mathbf{x}_n \mid \mathbf{z}_n; \Phi)$

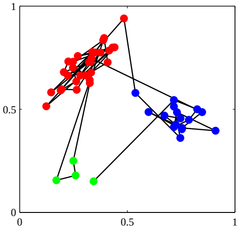
Interpretation



Transitions of $\mathbf{z}_n$



$p(\mathbf{x}_n \mid \mathbf{z}_n)$



Trajectory of $\mathbf{x}_n$

Properties of HMMs (1)

We have:

$$
\begin{aligned}
p(\mathbf{x}_n, \mathbf{z}_n | \mathbf{x}_{n-1}, \mathbf{z}_{n-1}) &= \frac{p(\mathbf{x}_{n-1}, \mathbf{x}_n | \mathbf{z}_{n-1}, \mathbf{z}_n) p(\mathbf{z}_{n-1}, \mathbf{z}_n)}{p(\mathbf{x}_{n-1} | \mathbf{z}_{n-1}) p(\mathbf{z}_{1:(n-1)})} \\
&= p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n | \mathbf{z}_{n-1}),
\end{aligned}
$$

and

$$
\begin{aligned}
p(\mathbf{x}_n, \mathbf{z}_n | \mathbf{x}_{1:n}, \mathbf{z}_{1:n}) &= \frac{p(\mathbf{x}_{1:n} | \mathbf{z}_{1:n}) p(\mathbf{z}_{1:n})}{p(\mathbf{x}_{1:n-1} | \mathbf{z}_{1:n-1}) p(\mathbf{z}_{1:n-1})} \\
&= p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n | \mathbf{z}_{n-1})
\end{aligned}
$$

Properties of HMMs (2)

$$
\begin{aligned}
p(\mathbf{x}_n|\mathbf{x}_{1:n-1}) &= \sum_{k=1}^{K} p(\mathbf{x}_n, \mathbf{z}_n = k|\mathbf{x}_{1:n-1}) \\
&= \sum_{k=1}^{K} p(\mathbf{x}_n|\mathbf{z}_n = k) P(\mathbf{z}_n = k|\mathbf{x}_{1:n-1})
\end{aligned}
$$

The distribution of $\mathbf{x}_n$ conditionally to its past is a mixture model whose weights $P(\mathbf{z}_n = i|\mathbf{x}_{1:n-1})$ depend on the complete past

Learning a HMM

Learning

Given an observed sequence $\mathbf{x}_1, \ldots, \mathbf{x}_N$, find HMM parameter $\boldsymbol{\theta}$ that has the most likely generated the sequence.s

As for a mixture model, the problem does not admit a closed-form solution when maximizing the log-likelihood

Learning a HMM with the Expectation-Maximization algorithm

Iterate:

1. Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$
2. $\hat{\boldsymbol{\theta}}_{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$

With

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) =$$

$$\sum_{k=1}^{K} \gamma_1(k; \boldsymbol{\theta}_t) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1,k=1}^{K} \xi_n(j, k; \boldsymbol{\theta}_t) \log A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_n(k; \boldsymbol{\theta}_t) \log p(\mathbf{x}_n | \Phi_k)$$

$$\gamma_n(k; \boldsymbol{\theta}_t) = P(\mathbf{z}_n = k | \mathbf{x}_{1:N}; \boldsymbol{\theta}_t) \qquad \xi_n(j, k; \boldsymbol{\theta}_t) = P(\mathbf{z}_{n-1} = j, \mathbf{z}_n = k | \mathbf{x}_{1:N}; \boldsymbol{\theta}_t)$$

Expectation step:

Application of the sum-product algorithm

In the context of HMM, the algorithm is known as *forward-backward*.

With $\alpha_1(i) = p(\mathbf{x}_1)\pi_i$ and $\beta_N(i) = 1$:
The following messages are propagated

- forward $\alpha_n(j) = p(\mathbf{x}_n|\mathbf{z}_n = j) \sum_{i=1}^{K} \alpha_{n-1}(i) p(\mathbf{z}_n = i|\mathbf{z}_{n-1} = j)$
- backward $\beta_n(i) = \sum_{j=1}^{K} \beta_{n+1}(j) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1} = j) p(\mathbf{z}_{n+1} = i|\mathbf{z}_n = j)$

they satisfy the properties:

$$\alpha_n(j) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n = j) \qquad \beta_n(i) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n = i)$$

Expectation step ctd':

$$p(\mathbf{z}_n = i | \mathbf{z}_{n-1} = j) = p(\mathbf{z}_{n+1} = i | \mathbf{z}_n = j) = A_{ij}$$

Finally we obtain the marginal probabilities:

$$\gamma_n(i) = \frac{\alpha_n(i)\beta_n(i)}{\sum_{j=1}^{K} \alpha_n(j)\beta_n(j)}$$

et

$$\xi_n(i,j) = \frac{\alpha_n(i) A_{ij} p(\mathbf{x}_{n+1} | \mathbf{z}_n = j)\beta_n(j)}{\sum_{k,\ell} \alpha_n(k) A_{k\ell} p(\mathbf{x}_n | \Phi_\ell)\beta_n(\ell)}$$

Maximization Step

When maximizing w.r.t. π and A one obtains

$$\boxed{\pi_k^{t+1} = \frac{\gamma_1^t(k)}{\sum_{j=1}^K \gamma_1^t(j)}} \qquad \boxed{A_{ij}^{t+1} = \frac{\sum_{n=2}^N \xi_n^t(i,j)}{\sum_{\ell=1}^K \sum_{n=2}^N \xi_n^t(i,\ell)}}$$

Maximization Step

When maximizing w.r.t. π and A one obtains

$$\boxed{\pi_k^{t+1} = \frac{\gamma_1^t(k)}{\sum_{j=1}^K \gamma_1^t(j)}} \qquad \boxed{A_{ij}^{t+1} = \frac{\sum_{n=2}^N \xi_n^t(i,j)}{\sum_{\ell=1}^K \sum_{n=2}^N \xi_n^t(i,\ell)}}$$

If the emissions are Gaussians we have as well:

$$\boldsymbol{\mu}_k^{t+1} = \frac{\sum_{n=1}^N \gamma_n^t(k)\mathbf{x}_n}{\sum_{n=1}^N \gamma_n^t(k)} \qquad \boldsymbol{\Sigma}_k^{t+1} = \frac{\sum_{n=1}^N \gamma_n^t(k)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^N \gamma_n^t(k)}$$