Sequential data

Much of this section follows Bishop chapter 13 (posted)

See also Murphy chapters 17 and 18

Sequential data

Sequential data is everywhere.

Examples:
- spoken language (word production)
- written language (sentence level statistics)
- weather
- human movement
- stock market data

Sequential data

Graphical models for such data?

The complexity of the representation seems to increase with time.

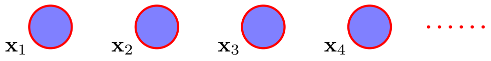
Observations over time tend to depend on the past.

We can simply life by assuming that the distant past does not matter.

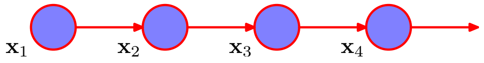
> If we assume that history does not matter other than the immediate previous entity, we have a first order Markov model.
>
> If what happens now depends on two previous entities, we have a second order Markov model.

Markov chains

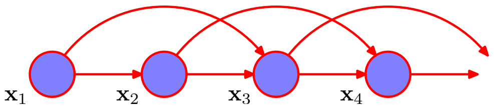
Zeroth order



First order



Second order



Temporal statistical clustering

In sequence data, cluster membership can have temporal (or sequential) structure.

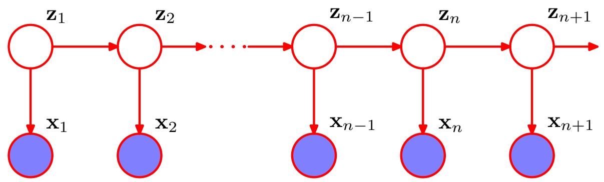
The data comes from the current cluster (as usual), but what is the next cluster?

Example, rain and sleet come from “stormy” and sunshine from “fair weather”.

But now, our hidden cluster variables are what depends on the past. The previous “state” represents all history.

Temporal statistical clustering

Hidden Markov Model (HMM).



The particular state encodes the important part of history.

Temporal statistical clustering

$p(x_n|z_n)$ tells us how to generate points from a cluster.

Once you know your cluster, things are easy.

But the cluster is now changing over time.

Markovian assumptions

As before, if the current state depends on only the previous state, we have a first order Markov model.

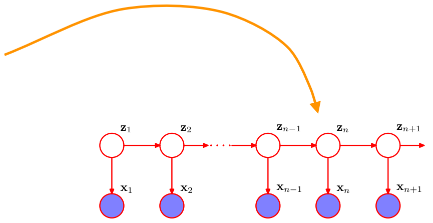
The basic HMM is like a mixture model, with the mix of mixture components being used for the current observations depends on the last mixture component.

Markovian assumptions

As before, if the current state depends on only the previous state, we have a first order Markov model.

The basic HMM is like a mixture model, with the mix of mixture components being used for the current observations depends on the last mixture component.

$$z_{n+1} \perp z_{n-1} \mid z_n$$



Markovian assumptions

Represent each component as a “state”.

Then, for first order Markov models, this leads to the concept of “transition” probabilities.

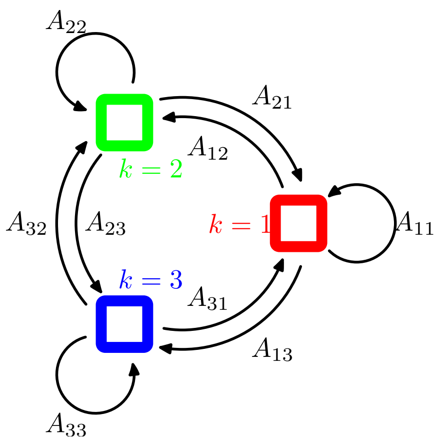
$$A_{jk} \equiv p\left(z_{nk} = 1 \middle| z_{n-1,j} = 1\right)$$

$$0 \le A_{jk} \le 1 \quad \text{and} \quad \sum_k A_{jk} = 1$$

The random variable, z, is a vector over K possible states (e.g., two for stormy vs fair-weather), for each time point.

Transition matrix representation

(Not a graphical model)



Starting state

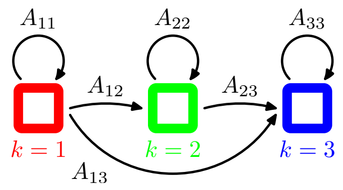
Our HMM will be a generative model, so we need to know how to start.

$$\pi_k \equiv p\left(z_{1k} = 1\right)$$

$$\text{with } 0 \le \pi_k \le 1 \ \text{ and } \ \sum_k \pi_k = 1$$

Left to right HMM

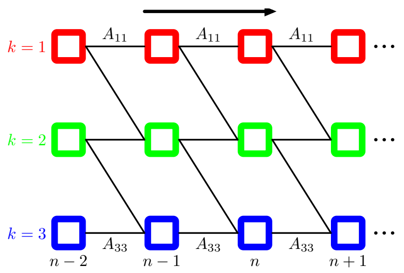
Constrain state number to increase



(State transition diagram)

Left to right HMM

Even more constrained, left to right HMM with single state jumps.



(Lattice diagram)

HMM parameter summary

$\theta = \{\pi, A, \phi\}$

π is probability over initial states

A is transition matrix

ϕ are the data emmission probabilities (e.g., means of Gaussians)

Data distribution from an HMM

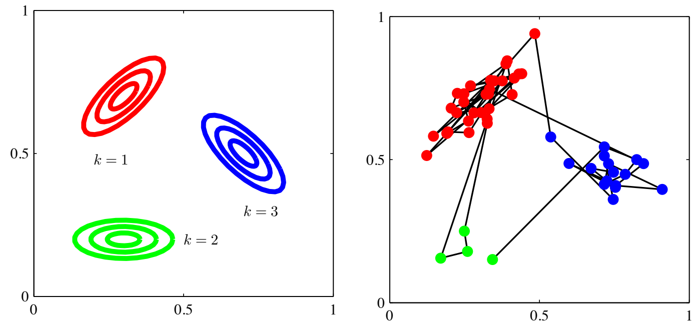
An HMM is specified by: $\theta = \{\pi, A, \phi\}$

$$p\left(X,Z|\theta\right) = p\left(z_1|\pi\right)\left[\prod_{n=2}^{N} p\left(z_n|z_{n-1}, A\right)\right]\prod_{n=1}^{N} p\left(x_n|z_n, \phi\right)$$

(complete data, i.e., we can generate from this).

Here, z_n represents the state (among K) for the n'th data point.

Data distribution from an HMM



Transition probability to another state is 5%

Classic HMM computational problems

1. Given data, what is the HMM (**learning**).
2. Given an HMM, what is the **probability distribution of states for** each time point (z_n in our notation).
3. Given an HMM, what is the most likely **state sequence** for some data?

#2 and #3 seem similar, but to understand the difference consider a three state system about doing problems A, B, and C in order, with B being very easy. So you will spend most of your time in state A and C. State B is least likely state for every time point. But the most likely state sequence must include it.

Learning the HMM (sketch)

If we know the state distributions, and the successive state pair distributions (needed for A), we can compute the parameters.

If we know the parameters, we can compute the state distributions (this is second problem which we need to solve as part of EM).

Blue text highlights differences from the mixture model.

Recall the General EM algorithm

1. Choose initial values for $\theta^{(s=1)}$

(can also do assignments, but then jump to M step).

2. E step: Evalute $p\left(Z\middle|X,\theta^{(s)}\right)$

3. M step: Evalute $\theta^{(s+1)}=\underset{\theta}{\arg\max}\left\{Q\left(\theta^{(s+1)},\theta^{(s)}\right)\right\}$

where $Q\left(\theta^{(s+1)},\theta^{(s)}\right)=\sum_{Z} p\left(Z\middle|X,\theta^{(s)}\right)\log\left(p\left(X,Z\middle|\theta^{(s+1)}\right)\right)$

4. Check for convergence; If not done, goto 2.

★ At each step, our objective function is increases unless it is at a local maximum. It is important to check this is

EM for HMM (sketch)

$$p(X,Z|\theta) = p(z_1|\boldsymbol{\pi})\left[\prod_{n=2}^{N} p(z_n|z_{n-1}, A)\right]\prod_{n=1}^{N} p(x_n|z_n, \phi)$$

$$= \prod_{k=1}^{K} \pi_k^{z_{1,k}}\left[\prod_{n=2}^{N}\prod_{j=1}^{K}\prod_{k=1}^{K} A_{j,k}^{z_{n-1,j} \bullet z_{n,k}}\right]\prod_{n=1}^{N}\prod_{k=1}^{K}\left(p(x_n|\phi_k)\right)^{z_{n,k}}$$

Remember our “indicator variable” notation. Z is a particular assignment of the missing values (i.e., which cluster the HMM was in at each time. For each time point, *n*, one of the values of z_n is one, and the others are zero. So, it “selects” the factor for the particular state at that time.

EM for HMM (sketch)

$$p(X,Z|\theta) = p(z_1|\boldsymbol{\pi})\left[\prod_{n=2}^{N} p(z_n|z_{n-1}, A)\right]\prod_{n=1}^{N} p(x_n|z_n, \phi)$$

$$= \prod_{k=1}^{K} \pi_k^{z_{1,k}}\left[\prod_{n=2}^{N}\prod_{j=1}^{K}\prod_{k=1}^{K} A_{j,k}^{z_{n-1,j} \bullet z_{n,k}}\right]\prod_{n=1}^{N}\prod_{k=1}^{K}\left(p(x_n|\phi_k)\right)^{z_{n,k}}$$

$$\log\left(p(X,Z|\theta)\right) = \sum_{k=1}^{K} z_{1k}\log(\pi_k) + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} z_{n-1,j}z_{n,k}\log\left(A_{j,k}\right) + \sum_{n=1}^{N}\sum_{k}^{K} z_{n,k}\log\left(p(x_n|\phi_k)\right)$$

EM for HMM (sketch)

In the simple clustering case (e.g., GMM), the E step was simple. For HMM it is a bit more involved.

The M step works a lot like the GMM, except we need to deal with successive states. Consider the M step first.

M step for HMM

We assume the E step computed distributions for

The degree each state explains each data point (analogous to GMM responsibilities).

$$\gamma(z_n) = p\left(z_n|X,\theta^{(s)}\right)$$

The degree that the combination of a state, and a previous one explain two data points.

$$\xi(z_{n-1},z_n) = p\left(z_{n-1},z_n|X,\theta^{(s)}\right)$$

“xi”