State Space Models with Unobservable States

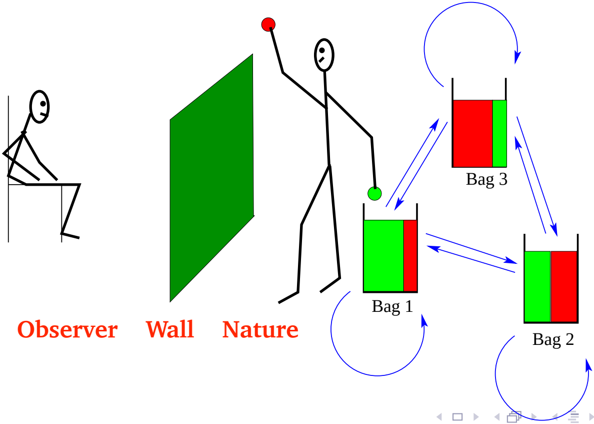
Peter Tino
School of Computer Science
University of Birmingham
UK

A simple example

Hidden Markov Model

Stationary emissions conditional on hidden (unobservable) states.

Hidden states represent basic operating "regimes" of the process.



Temporal structure - Hidden Markov Model

We have M bags of balls of different colors (Red - R, Green - G).

We are standing behind a curtain and at each point in time we select a bag j, draw (with replacement) a ball from it and show the ball to an observer. Color of the ball shown at time t is $C_t \in \{R, G\}$. We do this for T time steps.

The observer can only see the balls, it has no access to the information about how we select the bags.

Nature

Assume: we select bag at time t based only on our selection at the previous time step $t - 1$ (1st-order Markov assumption).

If only we knew ...

If we knew which bags were used at which time steps, things would be very easy! ... just counting

Hidden variables z_t^j: **Z is representing Hidden**
$z_t^j = 1$, iff bag j was used at time t;
$z_t^j = 0$, otherwise.

$$P(bag_j \to bag_k) = \frac{\sum_{t=1}^{T-1} z_t^j \cdot z_{t+1}^k}{\sum_{q=1}^{M} \sum_{t=1}^{T-1} z_t^j \cdot z_{t+1}^q}$$

[state transitions]

$$P(color = c \mid bag_j) = \frac{\sum_{t=1}^{T} z_t^j \cdot \delta(c = C_t)}{\sum_{g \in \{R,G\}} \sum_{t=1}^{T} z_t^j \cdot \delta(g = C_t)}$$

[emissions]

But we don't ...

We need to estimate probabilities for hidden events such as:

- $z_t^j \cdot z_{t+1}^k = 1$
  at time t - bag j, at the next time step - bag k

- $z_t^j \cdot \delta(c = C_t) = 1$
  at time t - bag j, ball of color c

The probability estimates need to be based on observed data $\mathcal{D}$ and our current model of state transition and emission probabilities.

Estimating values of the hidden variables

$$P(z_t^j \cdot z_{t+1}^k = 1 \mid \mathcal{D},\ \text{current model}) = R_t^{j \to k} \quad \textbf{state transition}$$

$$P(z_t^j \cdot \delta(c = C_t) = 1 \mid \mathcal{D},\ \text{current model}) = R_t^{j,c}$$

I will not deal with the crucial question of how to compute those posteriors over hidden variables, given the observed data and current model parameters.

This can be done efficiently - Forward-Backward algorithm.

Re-estimate the model

$$P(bag_j \to bag_k) = \frac{\sum_{t=1}^{T-1} z_t^j \cdot z_{t+1}^k}{\sum_{q=1}^{M} \sum_{t=1}^{T-1} z_t^j \cdot z_{t+1}^q} \quad \to$$

$$P(bag_j \to bag_k) = \frac{\sum_{t=1}^{T-1} R_t^{j \to k}}{\sum_{q=1}^{M} \sum_{t=1}^{T-1} R_t^{j \to q}} \qquad \text{[state transitions]}$$

$$P(color = c \mid bag_j) = \frac{\sum_{t=1}^{T} z_t^j \cdot \delta(c = C(t))}{\sum_{g \in \{R,G\}} \sum_{t=1}^{T} z_t^j \cdot \delta(g = C(t))} \quad \to$$

$$P(color = c \mid bag_j) = \frac{\sum_{t=1}^{T} R_t^{j,c}}{\sum_{g \in \{R,G\}} \sum_{t=1}^{T} R_t^{j,g}} \qquad \text{[emissions]}$$

Let's be more rigorous...

K states, $\mathbf{x}(t) \in \mathcal{X} = \{1, 2, ..., K\}$

Observations $\mathbf{y}(t) \in \mathcal{Y}$

HMM is a parameterised probabilistic model with parameters $\mathbf{w}$:

- initial state probabilities $p(\mathbf{x}(1))$
- transition probabilities $p(\mathbf{x}(t)|\mathbf{x}(t-1))$
- emission probabilities (discrete observations) $p(\mathbf{y}(t)|\mathbf{x}(t))$

HMM as a probabilistic model

Probability assigned to a time series $\mathbf{y}(1..T) = \mathbf{y}(1), \mathbf{y}(2), ..., \mathbf{y}(T)$

$$p(\mathbf{y}(1..T)|\mathbf{w}) = \sum_{\mathbf{x}(1..T) \in \mathcal{X}^T} p(\mathbf{x}(1)) \prod_{t=2}^{T} p(\mathbf{x}(t)|\mathbf{x}(t-1)) \prod_{t=1}^{T} p(\mathbf{y}(t)|\mathbf{x}(t))$$

NOTE: $\mathbf{x}(1..T) \in \mathcal{X}^T$ is hidden (latent) - cannot be directly observed!

Learning models with latent variables

Observed data: $\mathcal{D}$

Log-likelihood of $\mathbf{w}$: $\log p(\mathcal{D}|\mathbf{w})$

Current Model - being trained/learned

Train via Maximum Likelihood:

$$\mathbf{w}_{ML} = \underset{\mathbf{w}}{\operatorname{argmax}} \; \log p(\mathcal{D}|\mathbf{w}).$$

Complete data

Observed data: $\mathcal{D}$

Unobserved data: $\mathcal{Z}$ (unobserved state sequences)

Complete data: $(\mathcal{D}, \mathcal{Z})$

By marginalization ("integrate out the uncertainty in Z"):

$$p(\mathcal{D}|\mathbf{w}) = \sum_{\mathcal{Z}} p(\mathcal{D}, Z = \mathcal{Z}|\mathbf{w})$$

E-M Algorithm Expectation Maximisation

Given the current parameter setting $\mathbf{w}^{old}$ do:

- E-step:
  Estimate $P(Z|\mathcal{D}, \mathbf{w}^{old})$, the posterior distribution over all possible state paths $\mathcal{Z}$, given the observed data $\mathcal{D}$ and current parameter settings $\mathbf{w}^{old}$.

- M-step:
  Obtain new parameter values $\mathbf{w}^{new}$ by maximizing

$$\mathbb{E}_{P(Z|\mathcal{D}, \mathbf{w}^{old})}[\log p(\mathcal{D}, Z|\mathbf{w})].$$

- Set $\mathbf{w}^{old} := \mathbf{w}^{new}$ and go to E-step.

Why hidden states?

Model non-stationarity in the data. The states and be thought of as "stationary regimes", e.g.
- switching models in finance.

Model known expected temporal structures in the data when it is not clear when exactly which structure begins/ends, e.g.
- gene finders operating on DNA sequences
- spoken word recognition from a sequence of sounds
- natural language transcription

Can be extended to hierarchical models to account for e.g. a hierarchy of time scales in the signal (short, medium, long time scale structures)