COMS30035, Machine learning: Sequential Data 2: Hidden Markov Models

Edwin Simpson
edwin.simpson@bristol.ac.uk

Department of Computer Science, SCEEM
University of Bristol

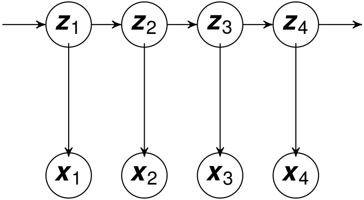
November 5, 2020

Agenda

- Markov Models
- Hidden Markov Models
- EM for HMMs
- Linear Dynamical Systems

Hidden Markov Models (HMMs)

- A state space model
- $\boldsymbol{z}_n$ are latent (unobserved) discrete state variables.
- $\boldsymbol{x}_n$ are observations, which may be discrete or continuous values depending on the application.



Uses of HMMs: Sequence Labelling for Text

- *Sequence labelling*, i.e., classifying data points in a sequence.
- E.g., classifying words in a text document into grammatical categories such as "noun", "verb", "adjective", etc.
- This is called part-of-speech (POS) tagging and is used by natural language understanding systems, e.g., to extract facts and events from text data.

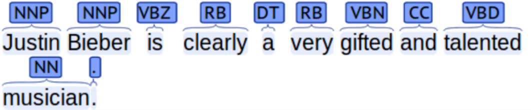
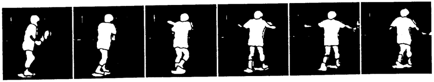


Image from "Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno", Yimam et al., 2014, ACL System Demonstrations.

Uses of HMMs: Human Action Recognition

- Observations: sequence of images (video frames) of a person playing tennis.
- Latent states: the actions being taken:
  - Backhand volley;
  - Forehand volley;
  - Forehand stroke;
  - Smash;
  - Serve.
- Why use an HMM? Actions typically follow a temporal sequence.


a)backhand volley

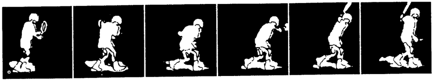

b)backhand stroke

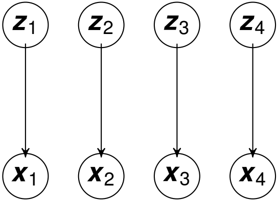
Image from Yamato, J., Ohya, J., Ishii, K. (1992). Recognizing human action in time-sequential images using hidden markov model. In CVPR (Vol. 92, pp. 379-385).

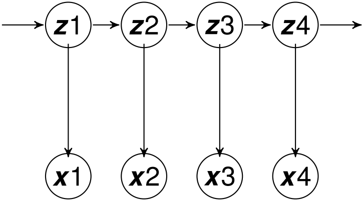
Uses of HMMs: In General

- HMMs can be used with different goals in mind:
  - Inferring the latent states (sequence labelling);
  - Predicting the next latent state;
  - Predicting the next observation;
- They can also be used with different levels of supervision:
  - Supervised: the latent states are given in the training set.
  - Unsupervised: no labels for the latent states, so the model seeks an assignment that best explains the observations given the model.
  - Semi-supervised: some labels are given, but the model is learned over both labelled and unlabelled data. Avoid overfitting to a very small labelled dataset while identifying latent states that follow the desired labelling scheme.

HMM is an Extension to Mixture Models

- Recall the latent variables, $\boldsymbol{z}_n$, in a mixture model, which identify the component responsible for an observation.
- These are also discrete variables, like latent states $\boldsymbol{z}_n$ in an HMM.
- In a mixture model, latent variables are i.i.d. rather than Markovian.





Anatomy of the HMM

- The probabilistic model of the HMM is made up of two main parts:
- The *transition* distribution, which can be represented as a *transition matrix* and models the dependencies between the latent states;
- The *emission* distributions, which model the observations given each latent state value.

Transition Matrix

- The probability of $\boldsymbol{z}_n$ depends on the previous state: $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1})$.
- Given K labels (state values), we can write all the values of $p(\boldsymbol{z}_n = k|\boldsymbol{z}_{n-1} = l)$ in a *transition matrix*, $\boldsymbol{A}$.
  - Rows correspond to values of the previous state, $\boldsymbol{z}_{n-1}$.
  - Columns are values of the current state, $\boldsymbol{z}_n$.

$p(\boldsymbol{z}_n\|\boldsymbol{z}_{n-1}, \boldsymbol{A})$		$\boldsymbol{z}_n$ 1	2	3
$\boldsymbol{z}_{n-1}$	1	0.5	0.1	0.4
	2	0.3	0.1	0.6
	3	0.01	0.19	0.8

- A vector of probabilities, π is used for $\boldsymbol{z}_1$, since it has no predecessor.
- Can you draw a transition matrix for a mixture model?

Emission Distributions

- Distribution over the observed variables, $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \phi)$, where ϕ are parameters of the distributions, for example:
  - Real-valued observations may use Gaussian emissions;
  - If there are multiple observations, we may use a multivariate Gaussian;
  - Discrete observations may use a categorical distribution.
- For each observation there are K values of $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \phi)$, one for each possible value of the unobserved $\boldsymbol{z}_n$.

The Complete HMM Model

- The complete HMM can be defined by the joint distribution over observations and latent states:

$$p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{A}, \pi, \phi) = p(\boldsymbol{z}_1 | \pi) \prod_{n=2}^{N} p(\boldsymbol{z}_n | \boldsymbol{z}_{n-1}, \boldsymbol{A}) \prod_{n=1}^{N} p(\boldsymbol{x}_n | \boldsymbol{z}_n, \phi) \quad (1)$$

- $\boldsymbol{A}$, π and ϕ are parameters that must be learned or marginalised.
- Generative model: think of generating each of the state variables $\boldsymbol{z}_n$ in turn, then generating the observation $\boldsymbol{x}_n$ for each generated state.

Now do the quiz!

Please do the quiz for this lecture on Blackboard.
Next, we will see how to learn an HMM using the EM algorithm.