COMS30035, Machine learning: Sequential Data 1: 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

Textbook

We will follow Chapter 13 of the Bishop book: Bishop, C. M., Pattern recognition and machine learning (2006). Available for free here.

i.i.d. Data

- Up to now, we have mainly consider the data points in our datasets to be *independent and identically distributed* (i.i.d.).
- Independent: the value of one data point does not affect the others. $p(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2)$.
- Identically distributed: all data points have the same distribution. $p(\boldsymbol{x}_i) = p(\boldsymbol{x}_j), \forall i, \forall j$.

Sequential Data

- The i.i.d. assumption means we ignore any ordering of the data points.
- Data points often occur in a sequence, such as words in a sentence, frames in a video, sensor observations over time, stock prices...
- This can be generalised to more than one dimension: pixels in an image, geographical data on a map... (not covered in this lecture).

Modelling Sequential Data

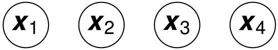
- How have we modelled relationships between data points so far? – Through their features.
- Why can't we take the same approach with sequential relationships and make *time* or *position in the sequence* another feature?
- Because it's what comes before or after that affects this data point – the value of the timestamp or positional index may not tell us anything about itself

Modelling Sequential Data

- Look at the following two texts from Bishop's book, both with a missing word:
  - "later termed Bayes' ____ by Poincarré"
  - "The evaluation of this conditional can be seen as an example of Bayes' ____"
- The first missing word is at position 3, the second is at position 13, but these position indexes don't help to identify the missing word.
- You can guess that the missing word in both cases is "theorem" or maybe "rule", because of the word "Bayes'" right before it.

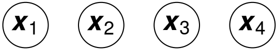
How Can We Model the Dependencies?

i.i.d., $p(\boldsymbol{x}_n | \boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n)$

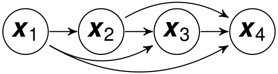


How Can We Model the Dependencies?

i.i.d., $p(\boldsymbol{x}_n | \boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n)$



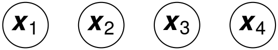
Modelling all connections, $p(\boldsymbol{x}_n | \boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1})$ – *intractable*



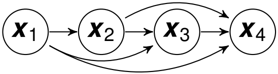
Edwin Simpson
edwin.simpson@bristol.ac.uk

How Can We Model the Dependencies?

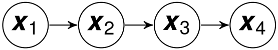
i.i.d., $p(\boldsymbol{x}_n|\boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n)$



Modelling all connections, $p(\boldsymbol{x}_n|\boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1})$ – *intractable*



1st order Markov chain, $p(\boldsymbol{x}_n|\boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})$



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

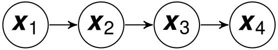
Homogeneous Markov Chains

- *Stationary* distribution: the probability distribution remains the same over time.
- This leads to a *homogeneous* Markov chain.
- E.g., the parameters of the distribution remain the same while the data evolves.
- Contrast with non-stationary distributions that change over time.

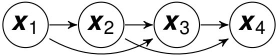
Higher-Order Markov Models

- Sometimes it is necessary to consider earlier observations using a higher-order chain.
- However, the number of parameters increases with the order of the Markov chain, meaning higher-order models are often impractical.

1st order Markov chain, $p(\boldsymbol{x}_n|\boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1})$



2nd order Markov chain, $p(\boldsymbol{x}_n|\boldsymbol{x}_1, ..., \boldsymbol{x}_{n-1}) = p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2})$

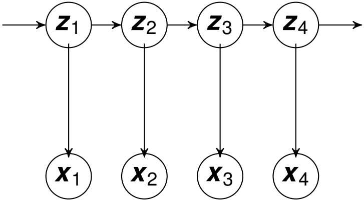


State Space Models

- What if we don't directly observe the states we want to model?
- E.g., the categories of words (nouns, verbs, adjectives, ...).
- E.g., we want to predict the state of the weather (raining, sunny, cloudy, rainfall) from noisy sensor observations (temperature, light meter, rain gauge);
- We encounter the same problem as for classification and regression: the variable we wish to predict is not directly observed.

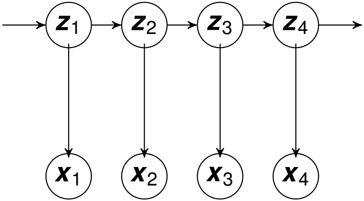
State Space Models

- Introduce latent variables, $\boldsymbol{z}_n$ that form a Markov chain;
- Each observation $\boldsymbol{x}_n$ depends on $\boldsymbol{z}_n$;
- This means we do not need to model the dependencies between observations $\boldsymbol{x}_n$ directly;
- Latent variables model the state of the system, while observations may be of different types, contain noise...



State Space Models

- *Hidden Markov Models (HMMs):* Discrete state z, observations may be continuous or discrete according to any distribution.
- *Linear Dynamical Systems (LDS):* Continuous state z, observations are continuous, both have Gaussian distributions
- The following videos will introduce these two key types of state space model and show how they can be learned in supervised and unsupervised settings.



Now do the quiz!

Please do the quiz for this lecture on Blackboard.