Hidden Markov models

David M. Blei

March 12, 2012

Markov chains

- We now consider data that are not IID. In particular, we consider sequential data.
- The idea behind a *Markov chain* is that the distribution of data at time t depends on the data at time $t-1$.
- What kind of data can be modeled sequentially?
  - Language
  - Genetic data
  - Gesture data
  - Location data
- Now this data is often more complicated than simply sequential. Sequential models make the data dependent while limiting the complexity of computing about it.
- A Markov chain of T variables has this joint distribution

$$p(x_1,\ldots,x_T) = p(x_1)\prod_{t=2}^{T} p(x_t \mid x_{t-1}) \tag{1}$$

- The parameters are
  - A transition matrix A
  - An initial distribution π
- Assume that x_t is discrete and takes values from K items.

  - (Draw the intitial distribution as a K-vector.)
  - (Draw the transition matrix as a $K \times K$ matrix.)

- (Draw the graphical model.)

- A Markov chain can be summarized with this independence:

$$x_{t+1} \perp\!\!\!\perp x_{t-1} \,|\, x_t \tag{2}$$

  "The future is independent of the past given the present."

- (Show this independence on the graphical model.)

Cool examples

- Cool example: Google
  - PageRank is based on a Markov chain.
  - Search for a query.
  - Return web results. (Dots on the board.)
  - These results are part of a network. (Edges on the board.)
  - Imagine a surfer traversing that network. This is a MC.
  - The "stationary distribution" determines the PageRank.
  - (Sketch the idea of a stationary distribution.)
  - (Note: It relates to the eigenvectors of the transition probabilities.)

- Another cool example : n-grams
  - Consider x_t to be a word
  - What does the transition matrix measure?
  - The probability of each word following another.
  - This is the *bigram* distribution of English.
  - Suppose the next word depended on the last 5 words
    * This is called a 5-gram distribution.
    * What happens to the transition "matrix"?
    * Recall smoothing. Why is it more of a problem here?
  - Claude Shannon discussed bigrams in 1948 and *sampled* from the distribution:
    1. Pick up a book and opened a random page.

2. Pick a random word.
3. Pick up another book and opened a random page.
4. Read that book until you find the word from the last book.
5. Mark the next word.
6. Repeat.

– We'll come back to these.

Hidden Markov models

- Here is a sequence of observations
  - (Draw 1d HMM observations.)
  - (Constrain the transition to move between two states from each.)

- This is contrived, but it's a *hidden Markov model*.

- The idea is that a sequence of latent class labels are drawn from a Markov chain. The data arise, as in a mixture model, from components associated with each latent class.

- (Draw the graphical model with $z_{1:T}$ and $x_{1:T}$.)

- The generative process is
  1. Draw $z_1 \sim \pi$
  2. For $t \in \{2, \ldots, T\}$
     (a) Draw $z_t \mid z_{t-1} \sim A_{z_t}$.
     (b) Draw $x_t \mid z_t \sim \theta_{z_t}$.

- Through the latent classes, the data exhibit a sequential structure.

- Discuss the parameters
  - Transition probabilities move from latent class to latent class. In HMMs nomenclature, the latent class variables are called *states*.
  - Emission probabilities are the K data-generating distributions, e.g., Gaussian or multinomial or something else.
  - Initial probabilities give the distribution over initial states with which to launch the chain.

- Examples. (Discuss each a little bit.)

  - Speech recognition
  - Part of speech tagging
  - Handwriting recognition
  - DNA analysis
  - Gesture recognition, e.g., Piazza post and bee dances.

- Discuss speech recognition in detail
  - Speech recognition is a success story for the HMM.
  - Latent states are words
  - Observations are the signal. (Segmenting is a separate problem.)
  - The transition distribution is a bigram distribution of words.
  - The observation probabilities reflect the distribution of signal for each possible word.
  - The intuition is that when I say "I'm going to [unintelligable] Michigan" and United Airlines only flies to Detroit, then knowing "Michigan" tells me a lot (via the bigram distribution) about which city I said.
  - Independent inferences wouldn't capture this.
  - Note: This is not exactly HMMs, which are purely unsupervised methods. Rather, this is like "supervised sequential classification."

- Discuss the joint distribution in detail. (See written notes.)

- Discus the close relationship to mixture models.

Computing with HMMs

- With these examples in mind, let's discuss how to compute about HMMs.

- As usual—cf classification, mixtures—there are two tasks: prediction and estimation.

- To estimate an HMM we use EM. (Why? Because there are hidden variables.) Prediction is part of the EM procedure.

- Here, things won't be as simple as in the other cases. The hidden variables are *dependent* on each other. Managing dependencies well a theme in computation with graphical models.

- (Switch to written notes.)

Bigger picture

- HMMs illustrate the modularity of graphical models.
- But, note that the complexity of inference is affected by the structure of the joint distribution. (More on this is in COS513.)
- HMMs also illustrate how EM is a general purpose strategy for fitting hidden variable models with maximum likelilhood.

Extensions of HMMs

(See written notes.)