6.867 Machine learning and neural networks

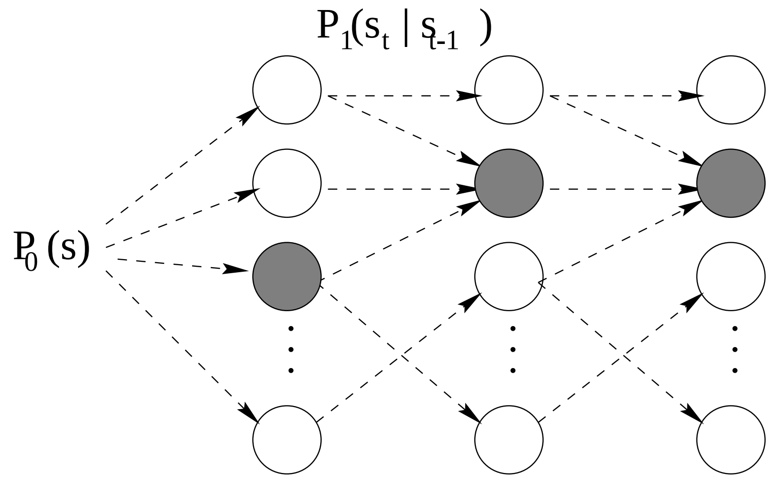
Tommi Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Lecture 16: Markov and hidden Markov models

Topics

- Markov models
  - motivation, definition
  - prediction, estimation
- Hidden markov models
  - definition, examples
  - forward-backward algorithm
  - estimation via EM

Review: Markov models

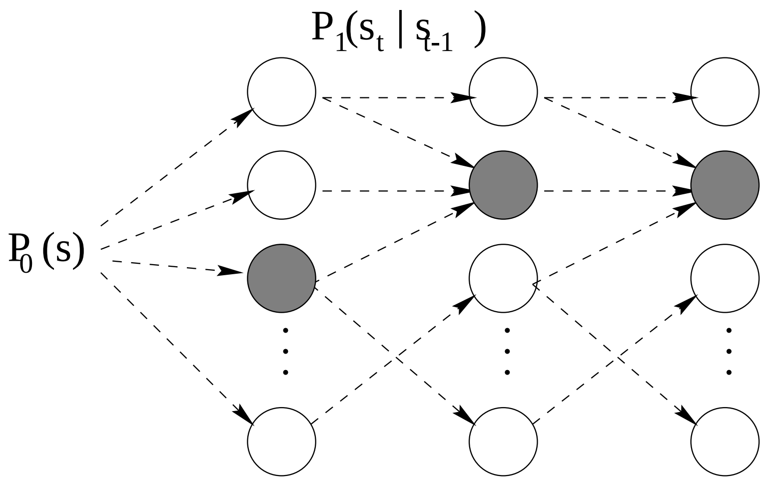


- The initial state s_0 is drawn form $P_0(s_0)$.
- There are a set of possible transtions from each state. These are marked with dashed arrows and correspond to transitions for which $P_1(s'|s_t) > 0$.
- Given the current state s_t we draw the next state s_{t+1} from the one step transition probabilities $P_1(s_{t+1}|s_t)$

$$s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow \ldots$$

- This is a *homogeneous Markov chain* where the transition probability does not change with time t

Properties of Markov chains



$$s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow \ldots$$

- If after some finite k transitions from any state i can lead to any other state j, the markov chain is *ergodic*:

$$P(s_{t+k} = j | s_t = i) > 0 \text{ for all } i, j \text{ and sufficiently large } k$$

(is the markov chain in the figure ergodic?)

Markov chains

- Problems we have to solve
  1. Prediction
  2. Estimation

- **Prediction**: Given that the system is in state $s_t = i$ at time t, what is the probability distribution over the possible states s_{t+k} at time $t + k$?

$$
\begin{aligned}
& P_1(s_{t+1}|s_t = i) \\
P_2(s_{t+2}|s_t = i) &= \sum_{s_{t+1}} P_1(s_{t+1}|s_t = i)\, P_1(s_{t+2}|s_{t+1}) \\
P_3(s_{t+3}|s_t = i) &= \sum_{s_{t+2}} P_2(s_{t+2}|s_t = i)\, P_1(s_{t+3}|s_{t+2}) \\
& \cdots \\
P_k(s_{t+k}|s_t = i) &= \sum_{s_{t+k-1}} P_{k-1}(s_{t+k-1}|s_t = i)\, P_1(s_{t+k}|s_{t+k-1})
\end{aligned}
$$

where $P_k(s'|s)$ is the k-step transition probability matrix.

Markov chain: estimation

- We need to estimate the initial state distribution $P_0(s_0)$ and the transition probabilities $P_1(s'|s)$
- Estimation from L observed sequences of different lengths

$$
\begin{array}{l}
s_0^{(1)} \to s_1^{(1)} \to s_2^{(1)} \to \ldots \to s_{n_1}^{(1)} \\
\ldots \\
s_0^{(L)} \to s_1^{(L)} \to s_2^{(L)} \to \ldots \to s_{n_L}^{(L)}
\end{array}
$$

Maximum likelihood estimates (observed fractions)

$$\hat{P}_0(s_0 = i) \; = \; \frac{1}{L} \sum_{l=1}^{L} \delta(s_0^{(l)}, i)$$

where $\delta(x, y) = 1$ if $x = y$ and zero otherwise

Markov chain: estimation

$$
\begin{array}{l}
s_0^{(1)} \rightarrow s_1^{(1)} \rightarrow s_2^{(1)} \rightarrow \ldots \rightarrow s_{n_1}^{(1)} \\
\ldots \\
s_0^{(L)} \rightarrow s_1^{(L)} \rightarrow s_2^{(L)} \rightarrow \ldots \rightarrow s_{n_L}^{(L)}
\end{array}
$$

- The transition probabilities are obtained as observed fractions of transitions out of a specific state

  Joint estimate over successive states

$$\hat{P}_{s,s'}(s=i, s'=j) \;=\; \frac{1}{(\sum_{l=1}^{L} n_l)} \sum_{l=1}^{L} \sum_{t=0}^{n_l-1} \delta(s_t^{(l)}, i)\delta(s_{t+1}^{(l)}, j)$$

  and the transition probability estimates

$$\hat{P}_1(s'=j|s=i) \;=\; \frac{\hat{P}_{s,s'}(s=i, s'=j)}{\sum_k \hat{P}_{s,s'}(s=i, s'=k)}$$

Markov chain: estimation

- Can we simply estimate Markov chains from a single long sequence?

$$s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow \ldots \rightarrow s_n$$

  - What about the initial state distribution $\hat{P}_0(s_0)$?
  - Ergodicity?

Topics

- Hidden markov models
  - definition, examples
  - forward-backward algorithm
  - estimation via EM

Hidden Markov models

- A hidden Markov model (HMM) is model where we generate a sequence of outputs in addition to the Markov state sequence

$$\begin{array}{ccccccc} s_0 & \rightarrow & s_1 & \rightarrow & s_2 & \rightarrow & \ldots \\ \downarrow & & \downarrow & & \downarrow & & \\ O_0 & & O_1 & & O_2 & & \end{array}$$

  A HMM is defined by

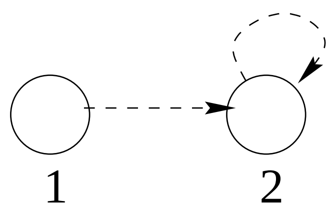
  1. number of states m
  2. initial state distribution $P_0(s_0)$
  3. state transition model $P_1(s_{t+1}|s_t)$
  4. output model $P_o(O_t|s_t)$ (discrete or continuous)

- This is a *latent variable model* in the sense that we will only observe the outputs $\{O_0, O_1, \ldots, O_n\}$; the state sequence remains "hidden"

HMM example

- Two states 1 and 2; observations are tosses of unbiased coins

$$P_0(s=1) = 0.5, \;\; P_0(s=2) = 0.5$$
$$P_1(s'=1|s=1) = 0, \;\; P_1(s'=2|s=1) = 1$$
$$P_1(s'=1|s=2) = 0, \;\; P_1(s'=2|s=2) = 1$$
$$P_o(O=\text{heads}|s=1) = 0.5, \;\; P_o(O=\text{tails}|s=1) = 0.5$$
$$P_o(O=\text{heads}|s=2) = 0.5, \;\; P_o(O=\text{tails}|s=2) = 0.5$$

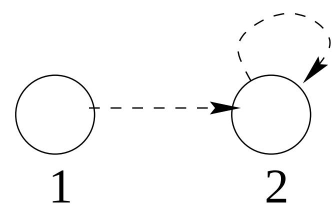


- This model is *unidentifiable* in the sense that the particular hidden state Markov chain has no effect on the observations

HMM example: biased coins

- Two states 1 and 2; outputs are tosses of *biased* coins

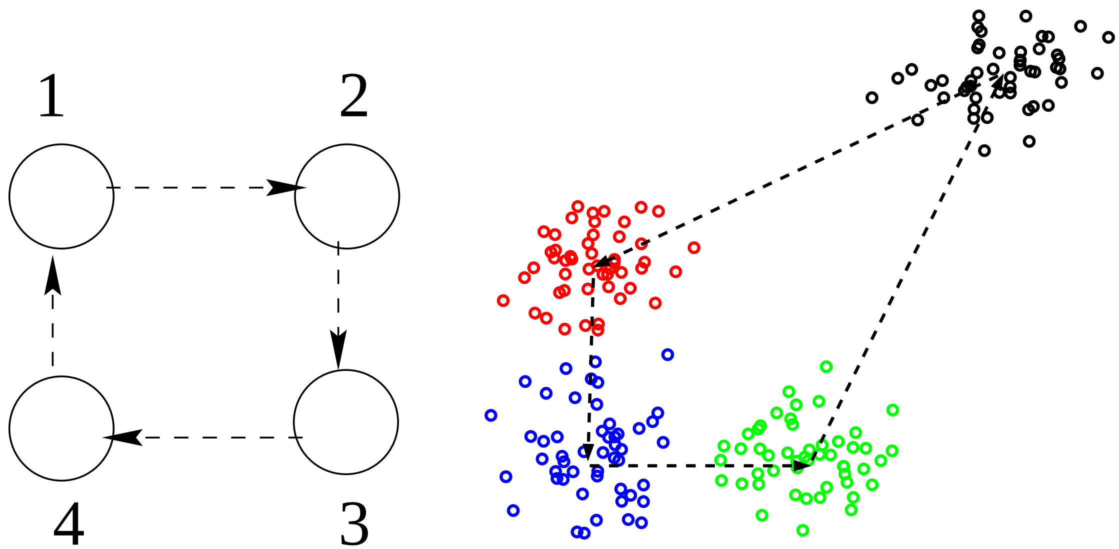
$$P_0(s = 1) = 0.5, \;\; P_0(s = 2) = 0.5$$
$$P_1(s' = 1|s = 1) = 0, \;\; P_1(s' = 2|s = 1) = 1$$
$$P_1(s' = 1|s = 2) = 0, \;\; P_1(s' = 2|s = 2) = 1$$
$$P_o(O = \text{heads}|s = 1) = 0.25, \;\; P_o(O = \text{tails}|s = 1) = 0.75$$
$$P_o(O = \text{heads}|s = 2) = 0.75, \;\; P_o(O = \text{tails}|s = 2) = 0.25$$



- What type of output sequences do we get from this HMM model?

HMM example

- Continuous output model: $O = [x_1, x_2]$, $P_o(O|s)$ is a Gaussian with mean and covariance depending on the underlying state s. Each state is initially equally likely.



- How does this compare to a mixture of four Gaussians model?

HMMs in practice

- HMMs have been widely used in various contexts
- Speech recognition (single word recognition)
  - words correspond to sequences of observations
  - we estimate a HMM for each word
  - the output model is a mixture of Gaussians over spectral features
- Biosequence analysis
  - a single HMM model for each type of protein (sequence of amino acids)
  - gene identification (parsing the genome)

  etc.

- HMMs are closely related to Kalman filters