Probabilistic Graphical Models

Prof. Joschka Boedecker and Hao Zhu

Department of Computer Science
University of Freiburg

universität freiburg

3. Markov models

- Markov chains
  - Inference
  - Parameter learning
  - Convergence

- Hidden Markov models
  - Inference
  - Decoding
  - Parameter learning
  - Continuous observation space

Outline

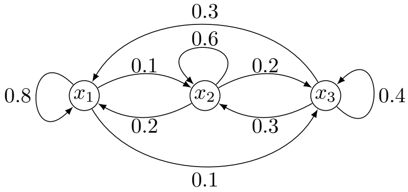
- Markov chains
  - Inference
  - Parameter learning
  - Convergence
- Hidden Markov models
  - Inference
  - Decoding
  - Parameter learning
  - Continuous observation space

Markov chains

Markov chain: a stochastic process $\{X_0, X_1, \ldots, X_t, \ldots\}$ with Markov property

$$\mathbf{P}(x_{t+1} \mid x_0, \ldots, x_t) = \mathbf{P}(x_{t+1} \mid x_t)$$

- X: state space
- time-homogeneous if $\mathbf{P}(X_{t+1} = x_j \mid X_t = x_i)$ is constant for all t



Markov chains

initial state distribution

$$\rho = (\ldots, \mathbf{P}(X_0 = x_i), \ldots), \quad i = 1, \ldots, m$$

- $\rho \in \mathbf{R}^m$; $\rho \succeq 0$; $\rho^T \mathbf{1} = 1$

transition matrix

$$P_{ij} = \mathbf{P}(x_j \mid x_i), \quad \text{for all } x_i, x_j \in X$$

- $P \in \mathbf{R}^{m \times m}$
- $P_{ij} \geq 0$ for all $i = 1, \ldots, m$, $j = 1, \ldots, m$
- $P_{i:}^T \mathbf{1} = 1$ for all $i = 1, \ldots, m$

Inference

given:

- parameters $\Theta = \{\rho, P\}$ of a Markov chain
- observed sequence of states $\zeta = \{X_0 = x_i, X_1 = x_j, X_2 = x_k, \ldots\}$

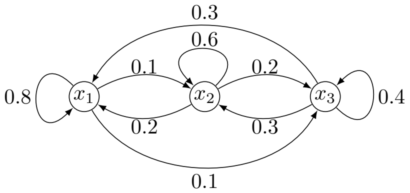
calculate the probability of observing the sequence ζ

$$\mathbf{P}(\zeta \mid \rho, P) = \rho_i P_{ij} P_{jk} \cdots$$

example

$$\rho = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad P = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$$

$$\mathbf{P}(x_1, x_2, x_3, x_1 \mid \rho, P) = \frac{1}{3} \times 0.1 \times 0.2 \times 0.3$$



Parameter learning

given a set of observed state sequences $\mathcal{D} = \{\zeta_1, \zeta_2, \ldots\}$ with $\zeta = \{x_0, \ldots, x_N\}$

estimate $\Theta = \{\rho, P\}$

- learning initial state distribution ρ:

$$\rho_i = \mathbf{E}_{\zeta \sim \mathcal{D}}[\mathbf{P}(X_0 = x_i \mid \zeta)], \quad i = 1, \ldots, m$$

- learning transition matrix P:

$$P_{ij} = \frac{\mathbf{E}_{\zeta \sim \mathcal{D}, t}\left[\mathbf{P}(X_t = x_i, X_{t+1} = x_j \mid \zeta)\right]}{\mathbf{E}_{\zeta \sim \mathcal{D}, t}\left[\mathbf{P}(X_t = x_i \mid \zeta)\right]}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, m$$

  - $t = 0, \ldots, N-1$

Parameter learning

proof

$$\begin{array}{ll}\text{maximize} & l_{\mathcal{D}}(\Theta) = \mathbf{E}_{\zeta\sim\mathcal{D}}\left[\log \mathbf{P}(\zeta \mid \rho, P)\right] \\ \text{subject to} & \rho \succeq 0, \quad \rho^T\mathbf{1} = 1 \\ & P_{ij} \geq 0, \quad P_{i:}^T\mathbf{1} = 1, \quad i = 1, \dots, m, \quad j = 1, \dots, m\end{array}$$

$$\begin{aligned}l_{\mathcal{D}}(\Theta) &= \mathbf{E}_{\zeta\sim\mathcal{D}}\left[\log \mathbf{P}(x_0 \mid \rho) \prod_{t=0}^{N-1} \mathbf{P}(x_{t+1} \mid x_t, P)\right] \\ &= \mathbf{E}_{\zeta\sim\mathcal{D}}\left[\log \mathbf{P}(x_0 \mid \rho)\right] + \mathbf{E}_{\zeta\sim\mathcal{D}}\left[\sum_{t=0}^{N-1} \log \mathbf{P}(x_{t+1} \mid x_t, P)\right] \\ &= \mathbf{E}_{\zeta\sim\mathcal{D}}\left[\sum_{i=1}^{m} I_{x_i}(x_0) \log \rho_i\right] + \mathbf{E}_{\zeta\sim\mathcal{D}}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{t=0}^{N-1} I_{x_i}(x_t) I_{x_j}(x_{t+1}) \log P_{ij}\right]\end{aligned}$$

Parameter learning

example

observed state sequences from a 3-states Markov chain:

- $(x_2, x_2, x_3, x_3, x_3, x_3, x_1)$
- $(x_1, x_3, x_2, x_3, x_3, x_3, x_3)$
- (x_3, x_3, x_2, x_2)
- $(x_2, x_1, x_2, x_2, x_1, x_3, x_1)$

$$\begin{aligned}\rho &= \left(\frac{1}{4}, \frac{2}{4}, \frac{1}{4}\right) \\ &= (0.25, 0.5, 0.25)\end{aligned}$$

$$P = \begin{bmatrix} \frac{0}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \\ \frac{2}{11} & \frac{2}{11} & \frac{7}{11} \end{bmatrix}$$

Convergence

requirements

- irreducible: $\mathbf{P}(x_j \mid x_i) > 0$, for all $x_i, x_j \in X$
- aperiodic: no fixed interval returning back to the same state

convergence: when $t \to \infty$

$$\pi P = \pi$$

$$\lim_{t \to \infty} P^t = \mathbf{1}\pi^T$$

- $\pi \in \mathbf{R}_+^m$: stationary state distribution

Outline

- Markov chains
  - Inference
  - Parameter learning
  - Convergence

- Hidden Markov models
  - Inference
  - Decoding
  - Parameter learning
  - Continuous observation space

Hidden markov models

hidden Markov model (HMM) consists of

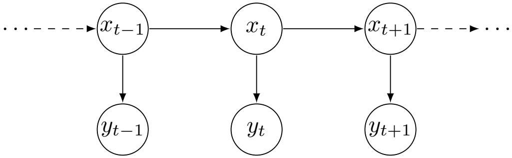
- a Markov chain $\{X_0, X_1, \ldots, X_t, \ldots\}$ with states not observable
- an observable stochastic process $\{Y_0, Y_1, \ldots, Y_t, \ldots\}$ with

$$\mathbf{P}(y_t \mid x_0, \ldots, x_t) = \mathbf{P}(y_t \mid x_t)$$

  - Y: observation space

emission matrix

$$B_{ij} = \mathbf{P}(y_j \mid x_i)$$



- $B \in \mathbf{R}^{m \times n}$
- $B_{ij} \geq 0$ for all $i = 1, \ldots, m$, $j = 1, \ldots, n$
- $B_{i:}^T \mathbf{1} = 1$ for all $i = 1, \ldots, m$

Inference

given:

- parameters $\Theta = \{\rho, P, B\}$ of an HMM
- some sequence of observations $\varphi = \{y_0, \ldots, y_N\}$

calculate the probability of observing the observation sequence φ

direct method

$$\begin{aligned}\mathbf{P}(\varphi \mid \Theta) &= \sum_{\zeta} \mathbf{P}(\varphi, \zeta \mid \Theta) \\ &= \sum_{x_0,\ldots,x_N} \mathbf{P}(x_0 \mid \rho)\mathbf{P}(y_0 \mid x_0, B) \prod_{t=0}^{N-1} \mathbf{P}(x_{t+1} \mid x_t, P)\mathbf{P}(y_{t+1} \mid x_{t+1}, B)\end{aligned}$$

- complexity: $2Nm^N$ (or simply m^N)

Inference

forward algorithm

$$\alpha_t(i) = \mathbf{P}(y_0, \ldots, y_t, X_t = x_i \mid \Theta), \quad i = 1, \ldots, m$$

- $\alpha_t \in \mathbf{R}_+^m$: forward probability
- recursive expression:

$$\alpha_t(i) = \begin{cases} \rho_i \mathbf{P}(y_0 \mid X_0 = x_i, B) & t = 0 \\ \alpha_{t-1}^T P_{:i} \mathbf{P}(y_t \mid X_t = x_i, B) & t > 0, \end{cases} \quad i = 1, \ldots, m$$

- probability of observing the observation sequence φ

$$\mathbf{P}(\varphi \mid \Theta) = \alpha_N^T \mathbf{1}$$

Inference

forward algorithm

---

given hidden Markov model parameters Θ, observation sequence φ.
for $i = 1, \ldots, m$ **do**
  $\alpha_0(i) := \rho_i \mathbf{P}(y_0 \mid X_0 = x_i, B)$.
end for
for $t = 1, \ldots, N$ **do**
  for $i = 1, \ldots, m$ **do**
    $\alpha_t(i) := \alpha_{t-1}^T P_{:i} \mathbf{P}(y_t \mid X_t = x_i, B)$.
  end for
end for
return $\mathbf{P}(\varphi \mid \Theta) = \alpha_N^T \mathbf{1}$.

- complexity: Nm^2 (or simply m^2)

Decoding

given:

- parameters $\Theta = \{\rho, P, B\}$ of an HMM
- some sequence of observations $\varphi = \{y_0, \dots, y_N\}$

find:

- the most probable state x_t^* at time t
- the most probable state sequence $\zeta^* = \{x_0^*, \dots, x_N^*\}$ that generated φ

Optimal state prediction

$$\beta_t(i) = \mathbf{P}(y_{t+1}, \ldots, y_N \mid X_t = x_i, \Theta), \quad i = 1, \ldots, m$$

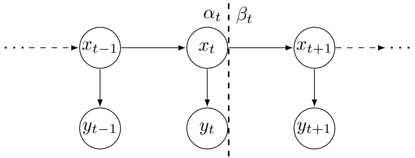
- $\beta_t \in \mathbf{R}^m_+$: backward probability
- recursive expression:

$$\beta_t(i) = \begin{cases} \sum_{j=1}^{m} \beta_{t+1}(j) P_{ij} \mathbf{P}(y_{t+1} \mid X_{t+1} = x_j, B) & t < N \\ 1 & t = N, \end{cases} \qquad i = 1, \ldots, m$$

- probability of observing the observation sequence φ

$$\mathbf{P}(\varphi \mid \Theta) = \alpha_t^T \beta_t$$

for all $t = 1, \ldots, N$



Optimal state prediction

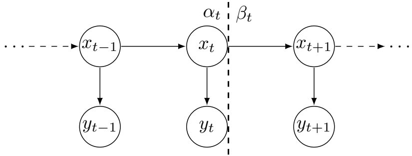
$$\gamma_t(i) = \mathbf{P}(X_t = x_i \mid \varphi, \Theta), \quad i = 1, \ldots, m$$

- $\gamma_t \in \mathbf{R}^m_+$
- in terms of α_t and β_t:

$$\gamma_t(i) = \frac{\mathbf{P}(X_t = x_i, \varphi \mid \Theta)}{\mathbf{P}(\varphi \mid \Theta)} = \frac{\alpha_t(i)\beta_t(i)}{\alpha_t^T \beta_t}, \quad i = 1, \ldots, m$$

- the most probable state x_t^* given φ and Θ

$$x_t^* = \underset{x_i}{\operatorname{argmax}}\, \gamma_t(i)$$



Optimal sequence prediction

$$x_0^*, \ldots, x_N^* = \underset{x_0,\ldots,x_N}{\operatorname{argmax}} \mathbf{P}(x_0, \ldots, x_N \mid \varphi, \Theta)$$

approximate solution

$$\tilde{\zeta}^* = \left\{ \underset{x_i}{\operatorname{argmax}} \gamma_t(i) \,\middle|\, t = 0, \ldots, N \right\}$$

- computationally efficient
- not the global optimal
- does not consider state transitions

Optimal sequence prediction

$$\mathbf{P}(\zeta \mid \varphi, \Theta) = \frac{\mathbf{P}(\zeta, \varphi \mid \Theta)}{\mathbf{P}(\varphi \mid \Theta)} \propto \mathbf{P}(\zeta, \varphi \mid \Theta)$$

Viterbi algorithm

$$\begin{aligned} \delta_t(i) &= \max_{x_0,\ldots,x_{t-1}} \mathbf{P}(x_0, \ldots, x_{t-1}, X_t = x_i, y_0, \ldots, y_t \mid \Theta) \\ &= \max_{j=1,\ldots,m} (\delta_{t-1}(j) P_{ji})\, \mathbf{P}(y_t \mid X_t = x_i, B), \quad i = 1, \ldots, m \end{aligned}$$

- $\delta_t \in \mathbf{R}_+^m$: probability of being in state x_i at time t given that the state subsequence up until $t-1$ is optimal w.r.t. the partial observation sequence $\{y_0, \ldots, y_{t-1}\}$

$$\psi_t(i) = \operatorname*{argmax}_{j=1,\ldots,m} \delta_{t-1}(j) P_{ji}, \quad i = 1, \ldots, m$$

- $\psi_t \in \mathbf{Z}_{++}^m$: the index j of the previous state x_j at time $t-1$ that gives the maximum probability $\delta_{t-1}(j) P_{ji}$, for each state i at time t

Optimal sequence prediction

Viterbi algorithm

given hidden Markov model parameters Θ, observation sequence φ.

```
1. Initialization.
```

for $i = 1, \ldots, m$ **do**

  $\delta_0(i) := \rho_i \mathbf{P}(y_0 \mid X_0 = x_i, B)$.

end for

```
2. Recursion.
```

for $t = 1, \ldots, N$ **do**

  for $i = 1, \ldots, m$ **do**

    $\delta_t(i) := \max_{j=1,\ldots,m} \left(\delta_{t-1}(j) P_{ji}\right) \mathbf{P}(y_t \mid X_t = x_i, B)$.

    $\psi_t(i) := \operatorname{argmax}_{j=1,\ldots,m} \delta_{t-1}(j) P_{ji}$.

  end for

end for

Optimal sequence prediction

`3. Termination.`
$p^* := \max_{i=1,\ldots,m} \delta_N(i)$.
$i_N^* := \operatorname{argmax}_{i=1,\ldots,m} \delta_N(i)$.
$x_N^* := x_{i_N^*}$.
`4. Backtracking.`
for $t = N, \ldots, 1$ **do**
  $i_{t-1}^* := \psi_t(i_t^*)$.
  $x_{t-1}^* := x_{i_{t-1}^*}$.
end for

Parameter learning

expectation-maximization (EM) algorithm

---

1. initialize parameters Θ
2. E-step: calculate some likelihood function w.r.t. Θ
3. M-step: update Θ by maximizing the likelihood function from 2
4. repeat 2–3 until convergence

- can be proved to converge to local optimum
- not guarantee to find the global optimum

Parameter learning

for each EM-iteration

- $\Theta = \{\rho, P, B\}$: the set of estimated parameters from previous iteration
- $\Theta^+ = \{\rho^+, P^+, B^+\}$: the new set of parameters to be updated

$$\rho_i^+ = \mathbf{E}_{\varphi\sim\mathcal{D}}\left[\mathbf{P}(X_0 = x_i \mid \varphi, \Theta)\right], \quad i = 1, \ldots, m$$

$$P_{ij}^+ = \frac{\mathbf{E}_{\varphi\sim\mathcal{D},t}\left[\mathbf{P}(X_t = x_i, X_{t+1} = x_j \mid \varphi, \Theta)\right]}{\mathbf{E}_{\varphi\sim\mathcal{D},t}\left[\mathbf{P}(X_t = x_i \mid \varphi, \Theta)\right]}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, m$$

$$B_{ij}^+ = \frac{\mathbf{E}_{\varphi\sim\mathcal{D},t}\left[\mathbf{P}(X_t = x_i, Y_t = y_j \mid \varphi, \Theta)\right]}{\mathbf{E}_{\varphi\sim\mathcal{D},t}\left[\mathbf{P}(X_t = x_i \mid \varphi, \Theta)\right]}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n$$

- $t = 0, \ldots, N-1$

Parameter learning

Baum-Welch algorithm

$$\begin{aligned}\xi_t(i,j) &= \mathbf{P}(X_t = x_i, X_{t+1} = x_j \mid \varphi, \Theta) \\ &= \frac{\mathbf{P}(X_t = x_i, X_{t+1} = x_j, \varphi \mid \Theta)}{\mathbf{P}(\varphi \mid \Theta)}, \quad i = 1,\ldots,m, \quad j = 1,\ldots,m\end{aligned}$$

- $\xi_t \in \mathbf{R}_+^{m\times m}$:probability of transitioning from state x_i at t to state x_j at $t+1$ given φ
- in terms of α_t and β_t:

$$\xi_t(i,j) = \frac{\alpha_t(i)P(i,j)\mathbf{P}(y_{t+1} \mid X_{t+1} = x_j, B)\beta_{t+1}(j)}{\sum_{i=1}^m \sum_{j=1}^m \alpha_t(i)P(i,j)\mathbf{P}(y_{t+1} \mid X_{t+1} = x_j, B)\beta_{t+1}(j)}$$

  for all $i = 1,\ldots,m$, $j = 1,\ldots,m$

- $\gamma_t(i) = \sum_{j=1}^m \xi_t(i,j)$, for all $i = 1,\ldots,m$

Parameter learning

Baum-Welch algorithm

$$\rho_i^+ = \mathbf{E}_{\varphi\sim\mathcal{D}}[\gamma_0(i)], \quad i = 1, \dots, m$$

$$P_{ij}^+ = \frac{\mathbf{E}_{\varphi\sim\mathcal{D},t}[\xi_t(i,j)]}{\mathbf{E}_{\varphi\sim\mathcal{D},t}[\gamma_t(i)]}, \quad i = 1, \dots, m, \quad j = 1, \dots, m$$

$$B_{ij}^+ = \frac{\mathbf{E}_{\varphi\sim\mathcal{D},t}[\gamma_t(i) I_j(y_t)]}{\mathbf{E}_{\varphi\sim\mathcal{D},t}[\gamma_t(i)]}, \quad i = 1, \dots, m, \quad j = 1, \dots, n$$

- indicator function $I_j(y) = 1$ if the index of $y \in Y$ is equal to j, and 0 otherwise

Continuous observation space

Gaussian hidden Markov models

- $\mathbf{dom}(Y) = \mathbf{R}$
- the emission map from X to Y is given by the Gaussian density function $\mathcal{N}(\mu_i, \sigma_i^2)$:

$$p(y \mid X = x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(y-\mu_i)^2}{2\sigma_i^2}\right), \quad i = 1, \ldots, m$$

  - $\mu_i \in \mathbf{R}$: mean of the Gaussian distribution
  - $\sigma_i^2 \in \mathbf{R}$: variance of the Gaussian distribution