MSA101/MVE187 2018 Lecture 13

Petter Mostad

Chalmers University

October 16, 2018

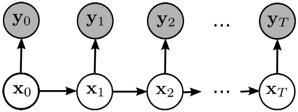
Overview

Today: Focus is on finding the maximum aposteriori (MAP) in graphical models.

- A basic example: The Viterbi algorithm, for HMMs.
- A more advanced example: The Baum-Welch algorithm, for estimating parameters in HMMs. An example of the EM-algorithm!
- The ideas of the two examples above can be generalized into handling graphical models where the graph does not contain loops. (Note that small loops can be handled conceptually by joining together several variables into one).
- When the graph contains large loops, the problem may be very hard (NP) and approximate solutions may be necessary.

The Viterbi algorithm

We start with an HMM where the x_i have a finite state space $\{1, \ldots, k\}$:

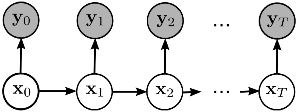


Objective: Compute the vector $x_0, \ldots, x_T$ which gives the highest posterior, for given fixed values of the y_i. IDEA: Compute and store, recursively, for $i = 0, \ldots, T$, the following:

- For $j = 1, \ldots, k$:
  - The vector $(\hat{x}_1, \ldots, \hat{x}_i)$ maximizing $\pi(\hat{x}_1, \ldots, \hat{x}_i, y_1, \ldots, y_i)$ with $\hat{x}_i = j$. NOTE: Only the value of $\hat{x}_{i-1}$ needs to be stored!
  - The value of this maximum.
- Because of indepenencies, the first $i-1$ values of $(\hat{x}_1, \ldots, \hat{x}_i)$ will always correspond to those considered at the $i-1$'th step.
- At any point, $(\hat{x}_1, \ldots, \hat{x}_i)$ can be reconstructed tracing backwards through stored information.
- The recursion step consists in considering all possible combinations of x_{i-1} and x_i.

The Baum-Welch algorithm

We start with an HMM where all the nodes have a finite state spaces



but where some of the parameters of the distributions $\pi(X_0)$, $\pi(X_i \mid X_{i-1})$, and $\pi(Y_i \mid X_i)$ are unknown. Objective: Given fixed values for the y_i, find maximum likelihood estimates for the parameters in the model.

- Note: By adding nodes representing the unknown parameters, and assuming flat priors, the problem becomes that of computing a MAP.
- Idea: Use the EM algorithm, with the values of the x_i as the augmented data.
- The E step of the EM algorithm is computed using the Forward-Backward algorithm.

The Baum-Welch algorithm: Example

For simplicity we assume each X_i can have values $1, \ldots, M$. Let

$$\theta = (q, p) = ((q_1, \ldots, q_M), (p_{11}, \ldots, p_{MM}))$$

be the parameters we want to estimate, where

$$\begin{aligned} q_j &= \Pr(X_0 = j) \\ p_{jk} &= \Pr(X_i = k \mid X_{i-1} = j) \end{aligned}$$

The full loglikelihood given θ becomes

$$\begin{aligned} & \log\left(\pi(x_0, \ldots, x_T, y_0, \ldots, y_T \mid \theta)\right) \\ = \; & \log\left(\pi(x_0 \mid \theta) \prod_{i=1}^{T} \pi(x_i \mid x_{i-1}, \theta) \prod_{i=0}^{T} \pi(y_i \mid x_i)\right) \\ = \; & \log \pi(x_0 \mid \theta) + \sum_{i=1}^{T} \log \pi(x_i \mid x_{i-1}, \theta) + \sum_{i=0}^{T} \log \pi(y_i \mid x_i) \\ = \; & C + \sum_{j=1}^{M} I(x_0 = j) \log q_j + \sum_{i=1}^{T} \sum_{j=1}^{M} \sum_{k=1}^{M} I(x_{i-1} = j) I(x_i = k) \log p_{jk} \end{aligned}$$

The Baum-Welch algorithm continued

- In the E step, we would like to compute the expectation of the full loglikelihood under the distribution $\pi(x_0, \dots, x_T \mid y_0, \dots, y_T, \theta^{old})$ for some set of parameters θ^{old}.
- Thus we need to compute the expectations $\mathrm{E}\left[I(x_0 = j)\right]$ and $\mathrm{E}\left[I(x_{i-1} = j) I(x_i = k)\right]$ under this distribution.
- Fixing θ^{old}, we can use the Forward-Backward algorithm to compute the densities $\pi(x_i \mid y_0, \dots, y_i)$ and $\pi(y_{i+1}, \dots, y_T \mid x_i)$. Further we have that

$$\begin{aligned}
& \pi(x_i, x_{i+1} \mid y_0, \dots, y_T) \\
\propto \quad & \pi(y_{i+1}, \dots, y_T \mid x_i, x_{i+1}) \pi(x_i, x_{i+1} \mid y_0, \dots, y_i) \\
\propto \quad & \pi(y_{i+2}, \dots, y_T \mid x_{i+1}) \pi(y_{i+1} \mid x_{i+1}) \pi(x_{i+1} \mid x_i) \pi(x_i \mid y_0, \dots, y_i)
\end{aligned}$$

making it possible to compute the joint posterior for x_i and x_{i+1} from these densities.

The Baum-Welch algorithm continued

The algorithm can now be summed up as

- Choose starting parameters θ^{old}.
- Run the Forward-Backward algorithm on the Markov model with parameters θ^{old} to compute the numbers $\mathrm{E}\left[I(x_0 = j)\right]$ and $\mathrm{E}\left[I(x_{i-1} = j)I(x_i = k)\right]$.
- Find the θ maximizing the expected loglikelihood

$$\sum_{j=1}^{M} \mathrm{E}\left[I(x_0 = j)\right] \log q_j + \sum_{i=1}^{T} \sum_{j=1}^{M} \sum_{k=1}^{M} \mathrm{E}\left[I(x_{i-1} = j)I(x_i = k)\right] \log p_{jk}$$

  In fact, we get

$$\hat{q}_j = \mathrm{E}\left[I(x_0 = j)\right] \text{ and } \hat{p}_{jk} = \frac{\sum_{i=1}^{T} \mathrm{E}\left[I(x_{i-1} = j)I(x_i = k)\right]}{\sum_{k=1}^{M} \sum_{i=1}^{T} \mathrm{E}\left[I(x_{i-1} = j)I(x_i = k)\right]}$$

- Set $\theta^{old} = ((\hat{q}_1, \ldots, \hat{q}_M), (\hat{p}_{11}, \ldots, \hat{p}_{MM}))$ and iterate until convergence.