Lecture 24: Theoretical foundations of EM

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

Expectation Maximization (EM)

- Mixture of probabilistic models $(M_1, M_2, \ldots, M_k)$ with parameters $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$
- Observation $O = o_1 o_2 \ldots o_N$
- Expectation step
  - Compute $Pr(M_i|o_j)$ for each M_i, o_j
- Maximization step
  - Recompute MLE for each M_i using fraction of O assigned using likelihood
- Repeat until convergence
- How do we justify this procedure?
- What can we say about convergence?

Hidden information

- EM builds a sequence of estimates $\Theta_1, \Theta_2, \ldots, \Theta_n$
- $L(\Theta_j)$ — log-likelihood function, $\ln Pr(O \mid \Theta_j)$
- Want to extend the sequence with Θ_{n+1} such that $L(\Theta_{n+1}) > L(\Theta_n)$
- Refer to Θ_{n+1} as Θ, maximize $L(\Theta) - L(\Theta_n)$
- Likelihood $L(\Theta)$ depends on hidden parameters $Z = \{z_1, z_2, \ldots, z_k\}$
  - $Pr(O \mid \Theta) = \sum_{z \in Z} Pr(O \mid z, \Theta) Pr(z \mid \Theta)$
- Rewrite $L(\Theta) - L(\Theta_n)$ as $\ln \left(\sum_{z \in Z} Pr(O \mid z, \Theta) Pr(z \mid \Theta) \right) - \ln Pr(O \mid \Theta_n)$

Jensen's inequality

- Maximize $\ln \left(\sum_{z \in Z} Pr(O \mid z, \Theta) Pr(z \mid \Theta) \right) - \ln Pr(O \mid \Theta_n)$

Jensen's inequality

Let f be a concave function defined on an interval I. If $x_1, x_2, \ldots, x_n \in I$ and $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$, then

$$f\left(\sum_{i=1}^{n} \lambda_i x_i \right) \geq \sum_{i=1}^{n} \lambda_i f(x_i)$$

- $\ln(\ )$ is a concave function
- To apply Jensen's inequality, need multipliers λ_i that sum to 1

Jensen's inequality

- Maximize $\ln \left(\sum_{z \in Z} Pr(O \mid z, \Theta) Pr(z \mid \Theta) \right) - \ln Pr(O \mid \Theta_n)$
- To apply Jensen's inequality, set λ_i's to be $Pr(z \mid O, \Theta_n)$
  - Probability measure, so $Pr(z \mid O, \Theta_n) \geq 0$ and $\sum_z Pr(z \mid O, \Theta_n) = 1$
- $L(\Theta) - L(\Theta_n)$

  $= \ln \left(\sum_{z \in Z} Pr(O \mid z, \Theta) Pr(z \mid \Theta) \right) - \ln Pr(O \mid \Theta_n)$

  $= \ln \left(\sum_{z \in Z} Pr(O \mid z, \Theta) Pr(z \mid \Theta) \frac{Pr(z \mid O, \Theta_n)}{Pr(z \mid O, \Theta_n)} \right) - \ln Pr(O \mid \Theta_n)$

  $= \ln \left(\sum_{z \in Z} Pr(z \mid O, \Theta_n) \frac{Pr(O \mid z, \Theta) Pr(z \mid \Theta)}{Pr(z \mid O, \Theta_n)} \right) - \ln Pr(O \mid \Theta_n)$

  $\geq \sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln \left(\frac{Pr(O \mid z, \Theta) Pr(z \mid \Theta)}{Pr(z \mid O, \Theta_n)} \right) - \ln Pr(O \mid \Theta_n)$ [Jensen]

$\Delta(\Theta, \Theta_n)$ and $\ell(\Theta \mid \Theta_n)$

- $L(\Theta) - L(\Theta_n)$

  $\geq \sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln \left(\frac{Pr(O \mid z, \Theta) Pr(z \mid \Theta)}{Pr(z \mid O, \Theta_n)} \right) - \ln Pr(O \mid \Theta_n)$ [Jensen]

  $= \sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln \left(\frac{Pr(O \mid z, \Theta) Pr(z \mid \Theta)}{Pr(z \mid O, \Theta_n)} \right) - \underbrace{\sum_{z \in Z} Pr(z \mid O, \Theta_n)}_{1} \ln Pr(O \mid \Theta_n)$

  $= \sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln \left(\frac{Pr(O \mid z, \Theta) Pr(z \mid \Theta)}{Pr(z \mid O, \Theta_n) Pr(O \mid \Theta_n)} \right)$

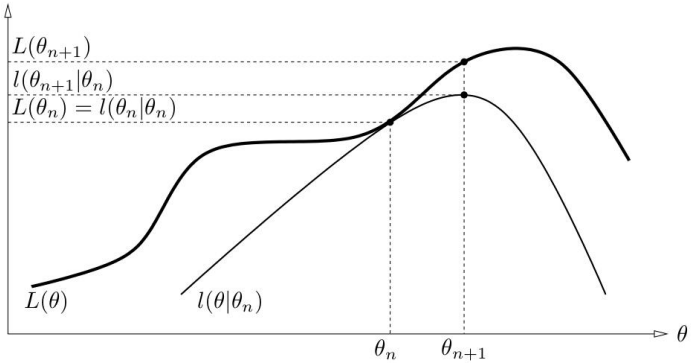
- Define $\Delta(\Theta, \Theta_n)$ to be $\sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln \left(\frac{Pr(O \mid z, \Theta) Pr(z \mid \Theta)}{Pr(z \mid O, \Theta_n) Pr(O \mid \Theta_n)} \right)$
- Hence $L(\Theta) - L(\Theta_n) \geq \Delta(\Theta, \Theta_n)$, or $L(\Theta) \geq L(\Theta_n) + \Delta(\Theta, \Theta_n)$
- Hence $L(\Theta) - L(\Theta_n) \geq \Delta(\Theta, \Theta_n)$, or $L(\Theta) \geq \underbrace{L(\Theta_n) + \Delta(\Theta, \Theta_n)}_{\ell(\Theta \mid \Theta_n)}$

$\ell(\Theta \mid \Theta_n) = L(\Theta_n)$

- $\ell(\Theta \mid \Theta_n) \leq L(\Theta)$
- $$\begin{aligned}
\ell(\Theta_n \mid \Theta_n) &= L(\Theta_n) + \Delta(\Theta_n, \Theta_n) \\
&= L(\Theta_n) + \sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln \left(\frac{Pr(O \mid z, \Theta_n) Pr(z \mid \Theta_n)}{Pr(z \mid O, \Theta_n) Pr(O \mid \Theta_n)} \right) \\
&= L(\Theta_n) + \sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln \left(\frac{Pr(O \mid z, \Theta_n)}{Pr(O \mid z, \Theta_n)} \right) \\
&= L(\Theta_n) + \sum_{z \in Z} Pr(z \mid O, \Theta_n) \ln 1 \\
&= L(\Theta_n)
\end{aligned}$$

Computing Θ_{n+1} from Θ_n

- $\ell(\Theta \mid \Theta_n) \leq L(\Theta)$
- $\ell(\Theta_n \mid \Theta_n) = L(\Theta_n)$
- Increasing $\ell(\Theta \mid \Theta_n)$ also increases $L(\Theta)$
- Choose Θ_{n+1} to maximize $\ell(\Theta \mid \Theta_n)$



Computing Θ_{n+1} from Θ_n

- $$\begin{aligned}\Theta_{n+1} &= \arg\max_{\Theta}\{\ell(\Theta \mid \Theta_n)\} \\ &= \arg\max_{\Theta}\left\{L(\Theta_n) + \sum_z Pr(z \mid O, \Theta_n) \ln \frac{Pr(O \mid z, \Theta) Pr(z \mid \Theta)}{Pr(O \mid \Theta_n) Pr(z \mid O, \Theta_n)}\right\}\end{aligned}$$

  Drop terms constant with respect to Θ

$$\begin{aligned} &= \arg\max_{\Theta}\left\{\sum_z Pr(z \mid O, \Theta_n) \ln Pr(O \mid z, \Theta) Pr(z \mid \Theta)\right\} \\ &= \arg\max_{\Theta}\left\{\sum_z Pr(z \mid O, \Theta_n) \ln \frac{Pr(O, z, \Theta)}{Pr(z, \Theta)} \frac{Pr(z, \Theta)}{Pr(\Theta)}\right\} \\ &= \arg\max_{\Theta}\left\{\sum_z Pr(z \mid O, \Theta_n) \ln Pr(O, z \mid \Theta)\right\} \\ &= \arg\max_{\Theta}\left\{E_{Z|O,\Theta_n}\{\ln Pr(O, z \mid \Theta)\}\right\}\end{aligned}$$

Expectation Maximization (EM), revisited

- $\Theta_{n+1} = \arg\max_{\Theta} \left\{ E_{Z|O,\Theta_n} \{ \ln Pr(O, z \mid \Theta) \} \right\}$
- EM algorithm iterates
  1. E-step: Determine conditional expectation $E_{Z|O,\Theta_n} \{ \ln Pr(O, z \mid \Theta) \}$
  2. M-step: Maximize this expectation with respect to Θ
- What about optimality?
  - Value at convergence need not be a local maximum
  - Could be a local minimum or a saddle point

Summary

- Expectation Maximization is a useful technique to estimate parameters when some information about the generating process is hidden
- We can theoretically justify building up a sequence of estimates through successive expectation + maximization steps
- As in other local search techniques, the stationary value one reaches may turn out to be a saddle point or a local minimum rather than a local maximum