em.hmm

Lili Mou
lmou@ualberta.ca
lili-mou.github.io
Unsupervised Learning

- Suppose an HMM model is given.

- Training

\[ \mathcal{D} = \left\{ \left( x_1^{(i)}, x_2^{(1)}, \ldots, x_T^{(i)} \right) \right\}_{i=1}^n \]

- Inference

  - Given an unseen sample \( x_1, x_2, \ldots, x_T \)

  - Predict their states \( s_1, s_2, \ldots, s_T \)
General Criteria for Latent Variables

• Training
  – Marginalization
    ▶ Something of $\mathbb{E}$
    ▶ $\mathbb{E}$ of something
    ▶ All sorts of variants

• Inference (depending on applications)
  – Target prediction: Marginalization
  – Latent variable prediction
    ▶ Max a posteriori
    ▶ Sampling
Gaussian Mixture Model

- **Gaussian mixture model:** \( z^{(n)} \rightarrow y^{(n)} \)

\( z^{(n)} \in \{1, \ldots, K\}, y^{(n)} \in \mathbb{R}^d \)

- **Generative process:**
  - Generate \( z^{(n)} \sim \text{cat}(\pi_1, \pi_2, \ldots, \pi_k) \)
  - Given \( z^{(n)} = k \), generate \( y^{(n)} \sim \mathcal{N}(\mu_k, \Sigma_k) \)

---

Expectation Maximization

- Gaussian mixture model: \( z^{(n)} \rightarrow y^{(n)} \)
  \( z^{(n)} \in \{1, \cdots, K\}, y^{(n)} \in \mathbb{R}^d \)

- Expectation maximization

  - **E-step**: Evaluate posterior of each latent category
    \[
    w_k^{(i)} = \frac{\pi_k \mathcal{N}(y^{(n)}; \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(y^{(n)}; \mu_j, \Sigma_j)}
    \]

  - **M-step**: Estimate model parameter
    \[
    \mu_k^{(\text{new})} = \frac{1}{N_k} \sum_{n=1}^N w_k^{(i)} y^{(n)}
    \]
    \[
    \Sigma_k^{(\text{new})} = \frac{1}{N_k} \sum_{n=1}^N w_k^{(i)} (y^{(n)} - \mu_k)(y^{(n)} - \mu_k)^T
    \]
    \[
    \pi_k^{\text{new}} = \frac{N_k}{N} \quad \text{where} \quad N_k = \sum_{i=1}^N w_k^{(i)}
    \]
Likelihood involves marginalization

\[ \log p(Y; \theta) = \log \left( \sum_z p(Y, z; \theta) \right) \]

\[ = \sum_z q(z \mid Y) \log \frac{p(Y, z; \theta)}{q(z \mid Y)} + \sum_z q(z \mid Y) \log \frac{q(z \mid Y)}{p(z \mid y; \theta)} \]

\[ L(q, \theta) \]  
Lower bound

\[ KL(q(Z \mid Y) \parallel p(Z \mid Y)) \]

For those only/over-familiar with VAE:

KL here is different from KL within the lower bound
EM as MLE

- Likelihood involves marginalization

\[
\log p(Y; \theta) = \log \left( \sum_z p(Y, z; \theta) \right)
= \sum_z q(z | Y) \log \frac{p(Y, z; \theta)}{q(z | Y)} + \sum_z q(z | Y) \log \frac{q(z | Y)}{p(z | y; \theta)}
\]

\[
L(q, \theta) \quad \text{KL}(q(Z | Y) \| p(Z | Y))
\]

- **E-step**: Fix \( \theta \), maximize \( L(q, \theta) \) wrt \( q(Z | Y) \)
  - Equivalent to minimize \( \text{KL}(\cdot \| \cdot) \), as \( \log p(Y | \theta) \) is constant
  - \( q(Z | Y) \overset{set}{=} p(Z | Y) \)

- **M-step**: Fix \( q(\cdot | \cdot) \), maximize \( L(q, \theta) \) wrt \( \theta \)
EM as MLE

\[ \ell(\theta_{t+1}) = \sum_i \log p(y_i; \theta_{t+1}) \]

\[ = \sum_i \log \left( \sum_z p(y_i, z; \theta_{t+1}) \right) \]

\[ \geq \sum_i \sum_z q_t(z | y_i) \log \frac{p(y_i, z; \theta_{t+1})}{q_t(z | y_i)} \]

\[ \geq \sum_i \sum_z q_t(z | y_i) \log \frac{p(y_i, z; \theta_t)}{q_t(z | y_i)} \]

\[ = \ell(\theta_t) \]

E-step: make lower bound tight

M-step: \( \theta_{t+1} = \arg \max \{ \cdot \} \)

[Lower bound holds for any \( q_t \)]
\[
\ell(\theta_{t+1}) = \sum_i \log p(y_i; \theta_{t+1}) \\
= \sum_i \log \left( \sum_z p(y_i, z; \theta_{t+1}) \right) \\
\geq \sum_i \sum_z q_t(z | y_i) \log \frac{p(y_i, z; \theta_{t+1})}{q_t(z | y_i)} \\
\geq \sum_i \sum_z q_t(z | y_i) \log \frac{p(y_i, z; \theta_t)}{q_t(z | y_i)} \\
= \ell(\theta_t)
\]

[Lower bound holds for any \( q_t \)]

**M-step:** \( \theta_{t+1} = \arg \max \{ \cdot \} \)

**E-step:** make lower bound tight
Hidden Markov Models

- Observed tokens: $y_1, y_2, \ldots, y_T$
- Latent states: $z_1, \ldots, z_T$
- Generative procedure
  - Choose $z_1$ (omitted here)
  - For every step $t$:
    ▶ Pick $z_t \sim p(z_t | z_{t-1})$
    ▶ Emit $y_t \sim p(y_t | z_t)$
  - Suppose both parametrized by probability tables
- Example
  - $y_1, y_2, \ldots, y_T$: a sequence of words
  - $z_1, z_2, \ldots, z_T$: POS tags
Hidden Markov Models

- **E-step** (expectation for sufficient statistics)
  - Expectation of a state, that is, \( \gamma_t(i) \triangleq \mathbb{E}[z_t = i \mid \cdot] \)
  - Expectation of two consecutive states, that is, \( \xi_t(i, j) \triangleq \mathbb{E}[z_t = i, z_{t+1} = j \mid \cdot] \)
  - Computed by
    \[
    \gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{p(Y)} \quad \xi_t(i, j) = \frac{\alpha_t(i)p_\theta(x_t \mid z_n = i)p_\theta(z_t = j \mid z_{t-1} = i)\beta_t(j)}{p(Y)}
    \]

  where \( \alpha_t(i) \triangleq p(y_{1:t}, z_t = i) \) and \( \beta_t(i) \triangleq p(y_{t+1:T} \mid z_t = i) \)

  are given by dynamic programming
Dynamic Programming

\[ \alpha_t(i) \triangleq p(y_{1:t}, z_t) \]

- Initialization
  \[ \alpha_1(i) \triangleq p(x_1, z_1 = i) = \pi_i \cdot p(x_1 | z_1 = i) \]

- Recursion
  \[ \alpha_t(i) = \sum_j \alpha_{t-1}(i)p(s_t = i | s_{t-1} = j)p(x_t | s_t = j) \]

- Termination
  
  When \( t = T \)
Dynamic Programming

\[ \beta_t(i) \overset{\Delta}{=} p(y_{t+1:T} \mid z_t) \]

- **Initialization**
  \[ \beta_T(i) = 1 \]

- **Recursion**
  \[ \beta_t(i) = \sum_j \beta_{t+1}(j)p(s_{t+1} = j \mid s_t = i)p(x_{t+1} \mid s_{t+1} = j) \]

- **Termination**
  When \( t = 1 \)
Hidden Markov Models

• **E-step** (expectation for sufficient statistics)
  
  - Expectation of a state, that is, $\gamma_t(i) \overset{\Delta}{=} \mathbb{E}[z_t = i \mid \cdot]$  
  - Expectation of two consecutive states, that is,  
    $\xi_t(i, j) \overset{\Delta}{=} \mathbb{E}[z_t = i, z_{t+1} = j \mid \cdot]$  

• **M-step** (MLE by soft counting)

\[
p(z_t = j \mid z_{t-1} = i) = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{i=1}^{T-1} \gamma_t(i)}
\]

\[
p(x \mid z_t = j) = \frac{\sum_{t=1}^{T} \gamma_t(j) \mathbb{I} \{X_t = x\}}{\sum_{t=1}^{T} \gamma_t(j)}
\]
Other Treatments

\[
\log p(Y | \theta) = \log \left( \sum_z p(Y, z | \theta) \right)
\]

- Exact marginalization (enumeration as in GMM, DP as in HMM)
- Choose the single best \( z \)
  - E.g., \( K \)-means clustering
- Choose top-\( N \) latent variables
  - Beam search
- Sampling
- Back propagation
  - If \( Y \) continuous, be careful of the degenerated distribution
  - If \( p(Y | z) \) is by CPT, be aware of the constraint \( \sum_y p(y | z) = 1 \)
Assignment

- Consider a Bayesian network: $X \rightarrow Z \rightarrow Y$
- All variables are discrete, taking $N_x, N_y, N_z$ values, resp.
- Observation: $\{(x_i, y_i)\}_{i=1}^M$
- Goal:
  - Figure out parameters as in conditional probability tables
  - Give an EM algorithm to estimate the parameters. Note that $z$ is unobserved.
Suggested Reading

• CS229
  - Note: http://cs229.stanford.edu/notes/cs229-notes8.pdf
  - Video: https://www.youtube.com/watch?v=ZZGTuAkF-Hw&list=PLEBC422EC5973B4D8&index=12

• Chap 9, Bishop, Pattern Recognition and Machine Learning.

Thank you!

Q&A