Gaussian Mixture Model and Hidden Markov Model

Adway Mitra
MLFA AI42001 Center for Artificial Intelligence
Indian Institute of Technology Kharagpur

November 1, 2019

Generative Model for Clustering

- Generative model is a story about how the data was created
- We imagine that each of K clusters has a prototype
- Every data point is a "noisy version" of one prototype
- For any datapoint i,
  - first the cluster index Z_i is decided ($Z_i \in \{1, 2, \ldots, K\}$)
  - then the feature X_i is created, as a noisy version of the selected cluster's prototype

Generative Model for Clustering

- Assumption: Each cluster represented by prototypes: $\{\theta_1, \theta_2, \ldots, \theta_K\}$
- for each datapoint i
  - Draw cluster index $Z_i \sim g$ (g: distribution on the clusters)
  - Draw feature vector $X_i \sim f(\theta_{Z_i})$ (f: distribution on the observation space)
- We choose f and g according to application (eg. f can be Gaussian if our observations are real-valued)

Inference and Estimation Problems

- Observed variables: X (our observed datapoints)
- Unknown variables: cluster assignments Z, cluster parameters θ
- Finding Z: Inference problem, $prob(Z|X, \theta)$
- Finding θ: Estimation problem, $\theta = argmaxprob(Z, X, \theta)$
- Challenge: The two problems are linked together!
- Cannot estimate θ directly because of Z

Gaussian Mixture Model

- Each cluster represented by a Gaussian distribution: $\mathcal{N}(\mu_j, \sigma_j)$ ($j \in \{1, \ldots, K\}$)
- Each cluster has a probability π_j, $\pi = [\pi_1, \ldots, \pi_K]$, $\pi_j \geq 0$, $\sum_{j=1}^{K} \pi_j = 1$
- Model parameters: $\{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$
- for each datapoint i
  - Draw cluster index $Z_i \sim Categorical(\pi)$
  - Draw feature vector $X_i \sim \mathcal{N}(\mu_{Z_i}, \sigma_{Z_i})$

Gaussian Mixture Model

- X_i depends only on Z_i, Z_i depends on nothing!
- Joint distribution:
  $prob(Z_1, \ldots, Z_N, X_1, \ldots, X_N) = \prod_{i=1}^{N} prob(Z_i) prob(X_i|Z_i)$
- $prob(Z_i) = \prod_{j=1}^{K} \pi_j^{I(Z_i=j)}$ (I: indicator function)
- $prob(X_i|Z_i) = \prod_{j=1}^{K} (\frac{1}{\sigma_j} exp(-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}))^{I(Z_i=j)}$
- Likelihood function
  $\mathcal{L}(\mu, \sigma, \pi) = \prod_{i=1}^{N} \prod_{j=1}^{K} (\frac{\pi_j}{\sigma_j} exp(-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}))^{I(Z_i=j)}$
- Log-likelihood $=$
  $\sum_{i=1}^{N} \sum_{j=1}^{K} I(Z_i = j)(log \pi_j - log \sigma_j - \frac{(x_i-\mu_j)^2}{2\sigma_j^2}))$

Gaussian Mixture Model

- $\{\mu_{MLE}, \sigma_{MLE}, \pi_{MLE}\} = argmax_{\mu,\sigma,\pi} \mathcal{L}(\mu, \sigma, \pi)$
- Solve $\frac{\partial \mathcal{L}}{\partial \mu_|} = 0$, $\frac{\partial \mathcal{L}}{\partial \sigma_|} = 0$
- $\mu_j = \frac{\sum_{i=1}^{N} I(Z_i = j) x_i}{\sum_{i=1}^{N} I(Z_i = j)}$, i.e. mean of the points in cluster j
- $\sigma_j = \frac{\sum_{i=1}^{N} I(Z_i = j)(x_i - \mu_j)^2}{\sum_{i=1}^{N} I(Z_i = j)}$ i.e. variance of the points in cluster j
- $\pi_j = \frac{\sum_{i=1}^{N} I(Z_i = j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} I(Z_i = j)}$, i.e. relative frequency of the points in cluster j
- Unfortunately we cannot compute these, as we do not know Z!

Expectation Maximization

- As we do not know $I(Z_i = j)$, we consider it as a random variable, with distribution $p(Z_i|X)$
- We replace $I(Z_i = j)$ by its expected value, $\gamma_{ij} = E(I(Z_i = j))$
- As I is binary, $E(I(Z_i = j)) = p(Z_i = j|X)$
- $p(Z_i = j|X) = p(Z_i = j|X_i) = \frac{p(X_i|Z_i=j)p(Z_i=j)}{p(X_i)} = \frac{p(X_i|Z_i=j)p(Z_i=j)}{\sum_{l=1}^{K} p(X_i|Z_i=l)p(Z_i=l)}$
- So, $\gamma_{ij} = \frac{\pi_j \mathcal{N}(X_i;\mu_j,\sigma_j)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(x_i;\mu_l,\sigma_l)}$
- $\mu_j = \frac{\sum_{i=1}^{N} \gamma_{ij} x_i}{\sum_{i=1}^{N} \gamma_{ij}}$, $\sigma_j = \frac{\sum_{i=1}^{N} \gamma_{ij}(x_i - \mu_j)^2}{\sum_{i=1}^{N} \gamma_{ij}}$, $\pi_j = \frac{\sum_{i=1}^{N} \gamma_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_{ij}}$

Expectation Maximization

- We use an iterative algorithm
  1. Initialize μ^0, σ^0, π^0
  2. Repeat
     - 2.1 E-step: Calculate $\gamma_{ij} = \frac{\pi_j^0 \mathcal{N}(X_i; \mu_j^0, \sigma_j^0)}{\sum_{l=1}^{K} \pi_l^0 \mathcal{N}(X_i; \mu_l^0, \sigma_l^0)}$
     - 2.2 M-step: Re-estimate the parameters
     - 2.3 $\mu_j^1 = \frac{\sum_{i=1}^{N} \gamma_{ij} x_i}{\sum_{i=1}^{N} \gamma_{ij}}$, $\sigma_j^1 = \frac{\sum_{i=1}^{N} \gamma_{ij}(x_i - \mu_j)^2}{\sum_{i=1}^{N} \gamma_{ij}}$, $\pi_j^1 = \frac{\sum_{i=1}^{N} \gamma_{ij}}{N}$
  3. If $(\mu^0, \sigma^0, \pi^0) \approx (\mu^1, \sigma^1, \pi^1)$, STOP
  4. Else set $(\mu^0 = \mu^1, \sigma^0 = \sigma^1, \pi^0 = \pi^1)$ and GOTO 2

Expectation Maximization

- When E-M algorithm converges, we get optimal values of the parameters $(\mu^{EM}, \sigma^{EM}, \pi^{EM})$
- Compute posterior distribution $p(Z_i|X_i) = \frac{\pi_j^{EM}\mathcal{N}(X_i;\mu_j^{EM},\sigma_j^{EM})}{\sum_{l=1}^{K}\pi_l\mathcal{N}(X_i;\mu_l^{EM},\sigma_l^{EM})}$
- Soft-clustering instead of hard-clustering as in K-means
- Mode of distribution may be used as cluster assignment

Model Likelihood

- The likelihood of a model: $\mathcal{L}(P) = prob(X)$ the joint distribution of the data according to the model
- If model contains *latent variables* like Z, marginalize over them

$$
\begin{aligned}
\mathcal{L}(\mu, \sigma, \pi) = prob(X) &= \prod_{i=1}^{N} prob(X_i) = \prod_{i=1}^{N} \sum_{k=1}^{K} prob(X_i, Z_i = k) \\
&= \prod_{i=1}^{N} \sum_{k=1}^{K} prob(X_i | Z_i = k) prob(Z_i = k) \\
&= \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \frac{1}{2\pi\sigma_k} exp(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2})
\end{aligned}
$$

Comparing models

- Two different GMMs - with different sets of parameters (μ_a, σ_a, π_a) and (μ_b, σ_b, π_b)
- They can be compared by their likelihoods
- $\mathcal{L}(\mu_a, \sigma_a, \pi_a) > \mathcal{L}(\mu_b, \sigma_b, \pi_b)$ implies that first model *fits* the data better than the second
- Choosing K may be done by this approach

Hidden Markov Model

- Consider sequential observations $x_1, x_2, \ldots, x_T$
- Key assumption in GMM: all the data-points are independent
- For sequential applications, this may not be true any longer!
- eg. a long audio stream with many speakers
- The observation x_t is likely to belong to same speaker as x_{t-1}
- There may be a transition pattern from one speaker to another!

Hidden Markov Model

- Different values of Z indicate *state* of the system (eg. which speaker is talking)
- System may have K states (decided by user)
- Current state Z_t depends on previous states $Z_1, \ldots, Z_{t-1}$
- Instead of $prob(Z_t)$, we need $prob(Z_t|Z_{t-1}, \ldots, Z_1)$
- *Markov Assumption*: Future indepedent of past, given the present!
- Markov model: $prob(Z_t|Z_{t-1}, \ldots, Z_1) = prob(Z_t|Z_{t-1})$
- New parameter instead of π: $A_{ij} = prob(Z_t = j|Z_{t-1} = i)$

Hidden Markov Model

- Each state represented by parameters: p_j ($j \in \{1, \ldots, K\}$) of *emission distribution f*
- *Transition distribution* from state i to state j: $A_{ij} = prob(Z_t = j | Z_{t-1} = i)$ (KxK matrix)
- Each row of matrix A: categorical probability distribution
- An *initial state distribution* π (similar to GMM)
- $Z_1 \sim Categorical(\pi); X_1 \sim f(p_{Z_1})$
- for each datapoint t
  - Draw cluster index $Z_t \sim Categorical(A_{Z_t - 1})$
  - Draw feature vector $X_t \sim f(p_{Z_t})$

Hidden Markov Model

- Common emission distributions: Categorical (discrete observations) or Gaussian (real observations)
- X_t depends on Z_t only, Z_t depends on Z_{t-1} only
- Joint distribution $prob(X, Z) = prob(Z_1)prob(X_1|Z_1)\prod_{t=2}^{T} prob(Z_t|Z_{t-1})prob(X_t|Z_t)$
- Rearranging, $prob(X, Z) = prob(Z_1) \times \prod_{t=2}^{T} prob(Z_t|Z_{t-1}) \times \prod_{t=1}^{T} prob(Z_t|X_t)$
- $prob(Z_1) : \pi$ (initial state distribution), $prob(Z_t|Z_{t-1}) : A$ (transition distribution), $prob(Z_t|X_t) : f(p)$ (emission distribution)

Forward-Backward Algorithm

Inference problem: Given (π, A, p), find posterior distribution $prob(Z_t|X_1,\dots,X_T)$

$$
\begin{aligned}
prob(Z_t|X_1,\dots,X_T) &\propto prob(Z_t, X_1,\dots,X_T) \\
&= prob(Z_t, X_1,\dots,X_t) \\
&x \quad prob(X_{t+1},\dots,X_T|Z_t, X_1,\dots,X_t) \\
&= prob(Z_t, X_1,\dots,X_t) prob(X_{t+1},\dots,X_T|Z_t) \\
&= \alpha_t(Z_t)\beta_t(Z_t)
\end{aligned}
$$

Forward Algorithm

$$
\begin{aligned}
\alpha_t(Z_t) &= prob(Z_t, X_1, \dots, X_t) \\
&= \sum_{Z_{t-1}} prob(Z_t, Z_{t-1}, X_1, \dots, X_t) \\
&= \sum_{Z_{t-1}} prob(Z_{t-1}, X_1, \dots, X_{t-1}) prob(Z_t, X_t | Z_{t-1}, X_1, \dots, X_{t-1}) \\
&= \sum_{Z_{t-1}} \alpha_{t-1}(Z_{t-1}) prob(Z_t, X_t | Z_{t-1}) \\
&= \sum_{Z_{t-1}} \alpha_{t-1}(Z_{t-1}) prob(Z_t | Z_{t-1}) prob(X_t | Z_t)
\end{aligned}
$$

$$
\alpha_1(Z_1) = \prod_{k=1}^{K} \pi_k^{I(Z_1=k)} f(X_1, p_k), prob(Z_t | Z_{t-1}) = \prod_{k,l=1}^{K,K} A_{kl}^{I(Z_{t-1}=k, Z_t=l)}
$$

Backward Algorithm

$$
\begin{aligned}
\beta_t(Z_t) &= prob(X_{t+1}, \ldots, X_T | Z_t) \\
&= \sum_{Z_{t+1}} prob(Z_{t+1}, X_{t+1}, \ldots, X_T | Z_t) \\
&= \sum_{Z_{t+1}} prob(Z_{t+1}, X_{t+1} | Z_t) prob(X_{t+2}, \ldots, X_T | Z_{t+1}, X_{t+1}, Z_t) \\
&= \sum_{Z_{t+1}} prob(Z_{t+1} | Z_t) prob(X_{t+1} | Z_{t+1}) prob(X_{t+2}, \ldots, X_T | Z_{t+1}) \\
&= \sum_{Z_{t+1}} prob(Z_{t+1} | Z_t) prob(X_{t+1} | Z_{t+1}) \beta_{t+1}(Z_{t+1})
\end{aligned}
$$

$$
\begin{aligned}
\beta_{T-1}(Z_{T-1}) &= \sum_{Z_T} prob(X_T, Z_T | Z_{T-1}) \\
&= \sum_{Z_T} prob(Z_T | Z_{T-1}) prob(X_T | Z_T)
\end{aligned}
$$

Parameter Estimation in HMM

Estimation problem: Estimate the parameters (π, A, p), though we don't know Z

$$prob(X, Z) = prob(Z_1)prob(X_1|Z_1)\prod_{t=2}^{T} prob(Z_t|Z_{t-1})prob(X_t|Z_t)$$

$$\mathcal{L}(\pi, A, p) = \prod_{k=1}^{K} \pi_k^{I(Z_1=k)} \times \prod_{t=2}^{T}\prod_{k,l=1}^{K,K} A_{kl}^{I(Z_{t-1}=k, Z_t=l)} \times \prod_{t=1}^{T} f(X_t, p_{Z_t})$$

Replace $I(Z_1 = k)$ by $\gamma_1(k) = E(I(Z_1 = k))$, $I(Z_{t-1} = k, Z_t = l)$ by $\xi_t(kl) = E(I(Z_{t-1} = k, Z_t = l))$

Baum-Welch Algorithm

Input: sequence $\{X_1, \ldots, X_T\}$, emission parameters p

1. Make initial estimates of parameters π^0, A^0
2. Repeat
   2.1 $\pi_k^1 = \gamma_k = \frac{\pi_k^0 f(X_1, p_k)}{\sum_{l=1}^K \pi_l^0 f(X_1, p_l)}$
   2.2 $A_{kl}^1 = \frac{\sum_{t=1}^{T-1} \xi_t(kl)}{\sum_{t=1}^{T-1} \gamma_t(k)}$
   2.3 If $(\pi^0, A^0,) \approx (\pi^1, A^1)$, STOP
   2.4 Else set $(\pi^0 = \pi^1, A^0 = A^1)$ and GOTO 2

$$\gamma_t(k) = \frac{\alpha_t(k)\beta_t(k)}{\sum_{i=1}^K \alpha_t(i)\beta_t(i)}, \xi_t(kl) = \frac{\alpha_{t-1}(k)\beta_t(l)A_{kl}^0 f(X_t, p_l)}{\sum_{i,j=1}^{K,K} \alpha_{t-1}(i)\beta_t(j)A_{ij}^0 f(X_t, p_j)}$$