Clustering (contd.) EM Algorithm

October 6, 2001
Instructor: Larry Ruzzo
Notes: Tushar Bhangale

Probability Review

Sample Space: The set of all possible outcomes is *sample space* (Ω) $P(\Omega) = 1$.
And probability of any event A: $P(A) \leq P(\Omega)$

Conditional Probability: The probability of an event given that another event has occurred is called a conditional probability. The conditional probability of *A* given *B* is denoted by P(*A*|*B*) ans is computed as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

P(A|B) is also called as the *posterior* probability of A i.e. probability of A after observing that event B has occurred. In this case P(A) is also called as *prior* probability.

Bayes' Rule:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

It is often easier to compute P(B|A) than P(A|B). Bayes' rules makes it possible to evaluate P(A|B).
Coin problem: Consider 2 biased coins, one (H_{biased}) has P(Head) = 0.99 and the other (T_{biased}) has P(Tail) = 0.99. One of them is drawn randomly ($P_{Hbiased} = P_{Tbiased} = 0.5$) and tossed. Thus the prior probability of $P_{Hbiased} = 0.5$. What is the posterior probability of H_{biased} given the fact that a Head occurred $P(H_{biased}|H)$?

$$P(H_{biased} \mid H) = \frac{P(H \mid H_{biased})P(H_{biased})}{P(H)} = \frac{P(H \mid H_{biased})P(H_{biased})}{P(H_{biased}) \times 0.99 + P(T_{biased}) \times 0.01} = \frac{0.99 \times 0.5}{0.5 \times 0.99 + 0.5 \times 0.01} = 0.99$$

Thus the posterior probability $P(H_{biased}|H)$= 0.99 where the prior probability of $P(H_{biased})$ was 0.5.

Notations used:

Z_{ij} {0,1} is a binary variable such that Z_{ij}=1 if $X_i \in$ Gaussian with μ_j and Z_{ij}= 0 otherwise.
Event A = sample X_i is drawn from $N(\mu_1, \sigma_1)$, $P(A) = \tau_1$
Event B = sample X_i is drawn from $N(\mu_2, \sigma_2)$, $P(B) = \tau_2$
Event D = $X_i \in [X, X + dx]$

Calculating E(Z_{ij}):

P(D|A) can be calculated using: $P(D \mid A) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu_j)^2}{2\sigma^2}} dx$

And P(A|D) can be calculated using P(D|A) and applying Bayes' rule as:

$$P(A \mid D) = \frac{P(D \mid A)P(A)}{P(D)}$$

where, P(D) = P(D|A)P(A) + P(D|B)P(B) if A and B are mutually exclusive and exhaustive.

$$P(D) = \sum_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma^2}} \tau_j$$

$$P(A \mid D) = \frac{\sum_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma^2}} \tau_j}{\sum_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma^2}}}$$

D is the observed data and A is the model. P(A|D) is the posterior probability after seeing the data D that it came from model A.

And $E(Z_{ij}) = P(A|D)$.

Clustering can also be classified into hard clustering and soft clustering. Hard clustering is where every data point is assumed to belong to only one cluster. Soft clustering involves assigning a certain probability for the data point belonging to each cluster.

If τ_js are unknown but Zs are known, μs and τs can be calculated by using maximum likelihood estimation. If Zs are unknown, bayesian estimation has to be used to calculate Z_i.

EM Algorithms

EM stands for estimation-maximization. There are two types of EM algorithms.

Classification Em Algorithms: (Hard clustering)

Steps:

1. Given μs and τs, estimate Z_i
2. Assign each x_i to the best cluster
3. Re-estimate μs and τs
4. Reiterate

(General) EM Algorithm: (soft clustering)

Steps:

1. Random initialization of μs and τs
2. Using these values of μs and τs, estimate Zs
3. Given distribution of Zs, re-estimate μs and τs
4. Reiterate

Consider that the data points belong to a mixture of two Gaussians with means μ_1 and μ_2 and variance σ^2. Assuming equal likelihood of the data point belonging to each cluster i.e. $\tau_1 = \tau_2$, for any data point, the posterior probability (given the μs) of it belonging to any cluster, is given by,

$$P(X_i, Z_{i1}, Z_{i2} \mid \mu_1, \mu_2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\pi\sigma^2}\sum_{j=1}^{2} Z_{ij}(x_i-\mu_j)^2}$$

The joint probability for all the points is:

$$P((X_1, Z_{11}, Z_{12}), (X_2, Z_{21}, Z_{22}) \ldots \mid \mu_1, \mu_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\pi\sigma^2}\sum_{j=1}^{2} Z_{ij}(x_i-\mu_j)^2}$$

The goal is to maximize this probability, which is equivalent to maximizing the log of the function.

$$\max \log(P((X_1, Z_{11}, Z_{12}), (X_2, Z_{21}, Z_{22}) \ldots \mid \mu_1, \mu_2)) = \max \sum_{i=1}^{n} \left\{ \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{2} Z_{ij}(x_i - \mu_j)^2 \right\}$$

now, maximizing expected value of log P i.e. max E(log P), treating Z_i as a random variable drawn from distributions defined by μ_1^t, μ_2^t

$$\max E(\log(P((X_1, Z_{11}, Z_{12}), (X_2, Z_{21}, Z_{22}) \ldots \mid \mu_1, \mu_2))) = \max \sum_{i=1}^{n} \left\{ \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{j=1}^{2} E(Z_{ij})(x_i - \mu_j)^2 \right\}$$

Finding μ_1 and μ_2 that maximize E(log P) is equivalent to finding μ_1 and μ_2 that minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{2} E(Z_{ij})(x_i - \mu_j)^2$$

$$\frac{\partial}{\partial \mu_1} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{2} E(Z_{ij})(x_i - \mu_j)^2 \right\} = -2 \sum_{i=1}^{n} E(Z_{ij})(x_i - \mu_1) = 0$$

$$\mu_1 = \frac{\sum_{i=1}^{n} E(Z_{i1}) x_i}{\sum_{i=1}^{n} E(Z_{i1})} \ \textit{and}\ \mu_2 = \frac{\sum_{i=1}^{n} E(Z_{i2}) x_i}{\sum_{i=1}^{n} E(Z_{i2})}$$

$$\textit{similarly for k clusters},\ \mu_k = \frac{\sum_{i=1}^{n} E(Z_{ik}) x_i}{\sum_{i=1}^{n} E(Z_{ik})}$$

Same technique can be used to estimate unknown τs and σs if they are not the same for each cluster.

<u>EM Algorithm (proof of convergence):</u>

Let X be the visible data
Y the hidden data
θ, θ^t the parameters where θ^t is the value of the parameters at time t

$$P(Y \mid X, \theta) = \frac{P(Y \cap X \mid \theta)}{P(X \mid \theta)}$$

$$\forall \textit{ fixed } y, \log P(X \mid \theta) = \log(P(X, Y \mid \theta)) - \log P(Y \mid X, \theta)$$

$$\log P(X \mid \theta) = \sum_{y} P(Y \mid X, \theta^t) \log(P(X, Y \mid \theta)) - P(y \mid X, \theta^t) \log P(Y \mid X, \theta) \ldots\ldots\ldots\ldots (1)$$

$$\textit{Let } Q(\theta \mid \theta^t) = \sum_{y} P(Y \mid X, \theta^t) \log(P(X, Y \mid \theta))$$

$\textit{Subtracting } \log P(X \mid \theta^t) \textit{ from } (1),$

$$\log P(X \mid \theta) - \log P(X \mid \theta^t) = Q(\theta \mid \theta^t) - Q(\theta^t \mid \theta^t) + \sum_y P(y \mid X, \theta^t) \log \frac{P(Y \mid X, \theta^t)}{P(Y \mid X, \theta)}$$

$H\left(P(Y \mid X, \theta) \parallel P(Y \mid X, \theta^t)\right) \geq 0$ *is the relative entropy*

$\theta^{t+1} = \theta$ *that* max*imizes* $Q(\theta \mid \theta^t)$