Expectation Maximum Algorithm

Maximum Likelihood Estimation

Motivation

- We have data points $x_1, x_2 \dots x_n$ drawn from set X
- We have parameter $\theta \in \Omega$ (parameter space)
- We have distribution $P(x|\theta)$ so that

$$\sum_{x \in X} P(x|\theta) = 1 \ and \ P(x|\theta) \geq 0 \ \forall x$$

- Assume that we have $x_1, x_2 \dots x_n$ drawn **independently identically** from distribution $P(x|\theta^*)$ for $\theta^* \in \Omega$
- The likelihood will be

$$Likelihood(\theta) = P(x_1, x_2, \dots x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta) \ (because \ of \ IID)$$

- So with **"given"** dataset $x_1, x_2 \dots x_n$ we want to maximum likelihood function so that the distribution $P(x|\theta)$ is "close" to your "given" dataset

$$\operatorname*{argmax}_{\theta \in \Omega} L(\theta) = \operatorname*{argmax}_{\theta \in \Omega} \prod_{i=1}^{n} P(x_i|\theta) . L \ is \ log - likelihood$$

Example

- Let start with simple example that we flip the coin.
- $X = \{H, T\}$ so data points is sequence of heads or tails. In this example **HHTTHHHTHH**
- θ is single parameter that probability of coin coming up heads. $\Omega = [0,1]$
- $P(x|\theta)$ will be defined as $P(x|\theta) = \begin{cases} \theta \ if \ x = H \\ 1 - \theta \ if \ x = T \end{cases}$
- $L(\theta) = \theta^{count(H)}(1-\theta)^{n-count(H)}$. n is number of your data points
- The problem that we want to find θ that can fit with your example? We try to maximize

$$\begin{aligned}
\operatorname*{argmax}_{\theta \in \Omega} L(\theta) &= \operatorname*{argmax}_{\theta \in \Omega} \log\left(\theta^{count(H)}(1-\theta)^{n-count(H)}\right) \\
&= \operatorname*{argmax}_{\theta \in \Omega} count(H)\log(\theta) + \ (n - count(H)\log(1-\theta) \\
&= \operatorname*{argmax}_{\theta \in \Omega} 7\log(\theta) + 3\log(1-\theta) \\
\text{Derivative to find maximum point} &\Rightarrow \frac{7}{\theta} - \frac{3}{1-\theta} = 0 \\
&\Leftrightarrow \theta = \frac{7}{10} = \frac{count(H)}{n}
\end{aligned}$$

Expectation Maximum

Motivation

- The problem is if your distribution is mixture of K Gaussian

$$P(x|\theta) = \sum_{1}^{K} w_i N(\mu_i, \Sigma_i)$$

- How can we use apply maximum likelihood in Gaussian Mixture Model with numbers parameter can be over 100 parameters? How's about in case our model is formulated in term of "observed" and "unobserved" data. "Unobserved" in this case refer to **quantities.** For example, in given data points you don't know gaussian model that your sample is drawn from. If we can measure them, we can estimate the parameters by maximum likelihood. That's why Expectation Maximum is used to solve this case

Algorithm

- From Gaussian mixture model, we will have likelihood function like:

$$L(x_1, x_2, \dots x_n|\theta) = L(x_1, x_2, \dots x_n|\mu_1, \Sigma_1 \dots \mu_K, \Sigma_K) = \prod_{i=1}^{N} \sum_{j=1}^{k} w_j f(x_i|\mu_j, \Sigma_j)$$

$$z_{ij} = \begin{cases} 1 \ if \ x_i \ drawn \ from \ f_i \\ 0, otherwise \end{cases}$$

- If z_{ij} is known, we can estimate θ. That means you have data points for specific distribution
- If θ is known, we can estimate z_{ij}. For example if $\frac{|x_i-\mu_i|}{\Sigma_i} < \frac{|x_j-\mu_j|}{\Sigma_j}$, so that means x_i is likely drawn from f_i rather than f_j
- That intuitive idea of Expectation Maximum. We **iterate:**
  - Expectation step: we calculate expect value z_{ij} with given parameters
  - Maximum step: calculate "MLE" of parameter given $E(z_{ij})$

Mathematical Understanding

- If x_i is known, θ unknown. Your MLE function will be

$$L(x_1, x_2, \dots x_n|\theta)$$

- If we also know z_{ij}, consider

$$L(x_1, x_2, \dots x_n, z_{11}, z_{12} \dots z_{nk}|\theta)$$

- Now, we don't know z_{ij}. We maximum *expected likelihood* of visible data where expectation is over distribution of hidden data

$$E(L(x_1, x_2, \dots x_n, z_{11}, z_{12} \dots z_{nk}|\theta))$$

E-step: Find $\boldsymbol{E}(z_{ij})$

- Assume θ known (from previous iteration or initial θ)
- Z_{ij} are event that x_i drawn from f_j
- D is observed datum x_i
- Expect value of $z_{ij} = P(Z_{ij}|D)$ for each x_i

$$P(Z_{ij}|D) = \frac{p(\mathrm{D}|\mathrm{Z_{ij}}).p(Z_{ij})}{\sum_Z p(\mathrm{D}|\mathrm{Z})\mathrm{p}(\mathrm{Z})}$$

M-step: Reestimate $\boldsymbol{\theta^{t+1}}$

$$\theta^{t+1} = argmax_{\theta^t} \sum_i \sum_j E(z_{ij}) \log p(x_i, Z_{ij}; \theta^t)$$

Some example for EM

Three coin problem

- We have 3 coins. The problem is given sequence of Head and Tail from tossing coin 1,2 under condition. If coin 0 tossed before is H, we will toss coin 1 and If coin 0 is T, we toss coin 2.
- Define the problem

$$Y_0 = \{h, t\}, X = \{H, T\}, \theta = \{\lambda, p_1, p_2\}$$

$$p(Y_0|\theta) = \begin{cases} \lambda \; if \; Y = h \\ 1-\lambda \; if \; Y = t \end{cases}$$

$$p(x|Y_0, \theta) = \begin{cases} p_1, & y = h \\ p_2, & y = t \end{cases}$$

- Our partially observed data [H,T,H,T,H] with initial parameter $\lambda = 0.3, p_1 = 0.6, p_2 = 0.5$

E-step:

$$P(y=h|X=H) = \frac{P(X=H|y=h)p(y=h)}{p(X=H|y=h)p(y=h) + p(X=H|y=t)p(y=t)} = \frac{p_1\lambda}{p_1\lambda + p_2(1-\lambda)}$$

$$P(y=h|X=T) = \frac{(1-p_1)\lambda}{(1-p_1)\lambda + (1-p_2)(1-\lambda)};$$

$$P(y=t|X=H) = \frac{p_2(1-\lambda)}{p_1 + p_2(1-\lambda)};$$

$$P(y=t|X=T) = \frac{(1-p_2)(1-\lambda)}{(1-p_1)\lambda + (1-p_2)(1-\lambda)};$$

- With defined parameter $\lambda = 0.3, p_1 = 0.6, p_2 = 0.5$

$$P(y=h|X=H) = 0.34; P(y=t|X=H) = 0.66$$
$$P(y=h|X=T) = 0.225; P(y=t|X=T) = 0.775$$

- After filling hidden variables for each sample {H,T,H,T,H} we will have

3 <H>	$(<H>, h)\ P(y = h \mid X = H) = 0.34$
	$(<H>, t)\ P(y = t \mid X = H) = 0.66$
2 <T>	$(<T>, h)\ P(y = h \mid X = T) = 0.225$
	$(<T>, t)\ P(y = t \mid X = T) = 0.775$

M-step:

- New estimate for parameter:

$$\lambda = \frac{count(y = h)}{count(y)} = \frac{0.34 * 3 + 0.225 * 2}{5}$$

$$p_1 = \frac{count(X = H, y = h)}{count(y = h)} = \frac{0.34 * 3}{0.34 * 3 + 0.225 * 2}$$

$$p_2 = \frac{count(X = H, y = t)}{count(y = t)} = \frac{0.66 * 3}{0.66 * 3 + 0.775 * 2}$$

- We continue until it converge