EM algorithm

Victor Kitov

Latent variables ML

Suppose objects have observed features x and unobserved (latent) features z[1].

- $[x, z] \sim p(x, z, \theta)$, $x \sim p(x, \theta)$
- denote $X = [x_1, x_2, ...x_N]$, $Z = [z_1, z_2, ...z_N]$.

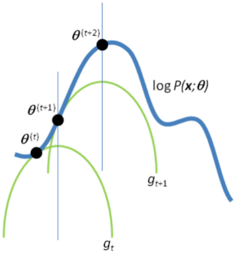
To find $\widehat{\theta}$ we need to solve

$$L(\theta) = \ln p(X|\theta) = \ln \sum_{Z} p(X, Z|\theta) \to \max_{\theta}$$

- This is intractable for unknown Z.
- We need to fallback to iterative optimization, such as SGD.
- Alternatively, we may use EM algorithm, which "averages" over different fixed variants of Z.

[1] They are considered discrete here. Everything holds true for continious latent variables if everywhere you replace summation over Z with integration

General idea of EM algorithm



- Initialize $\widehat{\theta}_0$ randomly, $t = 0$
- Repeat until convergence:
  1. $g_t(\theta)$ is estimated as lower bound for $\ln p(X|\theta)$, tight for $\widehat{\theta}_t$
  2. $\widehat{\theta}_{t+1} = \arg\max_\theta g_t(\theta)$
  3. $t = t + 1$

Distribution of latent variables

Let's introduce $q(Z)$ - some distribution over latent variables Z, $q(Z) \geq 0$, $\sum_Z q(Z) = 1$. Then

$$
\begin{aligned}
L(\theta) = \ln p(X|\theta) &= \ln \sum_Z p(X, Z|\theta) \\
&= \ln \sum_Z q(Z) \frac{p(X, Z|\theta)}{q(Z)} && (1) \\
&\geq \sum_Z q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)} = g(\theta) && (2)
\end{aligned}
$$

On the last step we used Jensen's inequality $\ln\left(\mathbb{E}u_n\right) \geq \mathbb{E}\left(\ln u_n\right)$ applied to

1. $\ln x$ which is strictly concave, because $(\ln x)'' = -\frac{1}{x^2} < 0$
2. for r.v. $U \in \mathbb{R}$ with distribution $p\left(U = \frac{p(X,Z,\theta)}{q(Z)}\right) = q(Z)$ for different Z.

Making lower bound tight

We can select $q(Z)$ so that at fixed θ $L(\theta) = g(\theta)$:

- Since $\ln x$ is strictly concave, equality in inequality (1)-(2) is achieved <=> $U = \mathbb{E}U$ with probability 1.
- This happens when $\frac{p(X,Z|\theta)}{q(Z)} = c$ for some constant c $\forall Z$.
- Using property $\sum_Z q(Z) = 1$ we have

$$c \sum_Z q(Z) = c = \sum_Z p(X, Z|\theta) = p(X|\theta)$$

- So for lower bound $g(\theta)$ to be tight at θ, we need to take

$$q(Z) = \frac{p(X, Z|\theta)}{p(X|\theta)} = p(Z|X, \theta)$$

Equivalent M-step

M-step can be equivalently represented as

$$
\begin{aligned}
\hat{\theta}_{t+1} &= \arg\max_{\theta}\{\sum_{Z} q(Z)\ln\frac{p(X,Z|\theta)}{q(Z)}\} \\
&= \arg\max_{\theta}\{\sum_{Z} q(Z)\ln p(X,Z|\theta) - \overbrace{\sum_{Z} q(Z)\ln q(Z)}^{const(\theta)}\} \\
&= \arg\max_{\theta}\{\sum_{Z} q(Z)\ln p(X,Z|\theta)\} \\
&= \arg\max_{\theta}\{\sum_{Z} p(Z|X,\widehat{\theta}_t)\ln p(X,Z|\theta)\} \\
&= \arg\max_{\theta}\{\mathbb{E}_Z\{\ln p(X,Z|\theta)\}, \quad Z \sim q(Z) = p(Z|X,\widehat{\theta}_t)
\end{aligned}
$$

EM algorithm

INPUT:

training set $X=[x_1,...x_N]$, convergence criteria

ALGORITHM:

$t=0$, θ_0 - init randomly

repeat until convergence:

E-step: set distribution over latent variables:

$$q(Z)=p(Z|X,\hat{\theta}_t)$$

M-step: improve estimate of θ:

$$\hat{\theta}_{t+1}=\arg\max_\theta\{\sum_Z q(Z)\ln p(X,Z|\theta)\}$$

$$t=t+1$$

OUTPUT:

ML estimate $\hat{\theta}_{t+1}$ **for** the training set.

Comments

Theorem 1

EM estimates of θ on each iteration $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \ldots$ lead to non-decreasing sequence of likelihoods $L(\widehat{\theta}_1) \geq L(\widehat{\theta}_2) \geq L(\widehat{\theta}_3) \geq \ldots$

Proof.

1. Suppose that at iteration t we have $L(\widehat{\theta}_t)$.
2. At the E-step among all lower bounds $g(\theta) \leq L(\theta)\,\forall\theta$ we select such lower bound $g_t(\cdot)$, that $L(\widehat{\theta}_t) = g_t(\widehat{\theta}_t)$ (by selecting $q_n(Z)$).
3. On M-step we find $\widehat{\theta}_{t+1} = \arg\max_\theta g_t(\theta)$, so $g_t(\widehat{\theta}_{t+1}) \geq g_t(\widehat{\theta}_t)$
4. Since $g_t(\cdot)$ is lower bound, we have $L(\widehat{\theta}_{t+1}) \geq g_t(\widehat{\theta}_{t+1}) \geq g_t(\widehat{\theta}_t) = L(\widehat{\theta}_t)$

$\square$

Since $L(\widehat{\theta}_t)$ is non-decreasing and is bounded from above ($L(\theta) \leq \sum_{n=1}^N \ln 1 = 0$) it converges.

Comments on EM algorithm

- On M-step $q(Z)$ does not depend on θ, since this parameter was taken fixed from E-step.
- Possible convergence criteria:
  - $\left\|\widehat{\theta}_{t+1}-\widehat{\theta}_t\right\|<\varepsilon$
  - $L(\widehat{\theta}_{t+1})-L(\widehat{\theta}_t)<\varepsilon$
  - maximum number of iterations reached
- EM converges to local optimum
  - to improve quality it is good to
    - re-run algorithm from different initial conditions
    - select estimate that gives the greatest likelihood
- To guarantee convergence it is not required to solve $\widehat{\theta}_{t+1}=\arg\max_\theta g_t(\theta)$ precisely.
  - we can make very coarse (e.g. single step) optimization here
  - this is called GEM algorithm (generalized EM)

Comments on EM algorithm

- EM can also be applied for MAP optimization
- Define $J(Q,\theta) = \sum_Z q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}$.
- We know that $L(\theta) \geq J(Q,\theta)$ for all $Q = Q(Z)$.
- EM algorithm can be viewed as coordinate ascent:
    - E-step maximizes $J(Q,\theta)$ w.r.t. Q[2]
    - M-step maximizes $J(Q,\theta)$ w.r.t. θ

[2] We know that, because we chose such Q that ensure equality in Jensen's inequality.

Table of Contents

Independent observations

- Consider special case, when (x_n, z_n) are i.i.d.[3]
  - Examples:.
    - z_n is unknown mixture component, generating x_n
    - z_n are missing variables in i.i.d. x_n
- **E-step** becomes:

$$q(Z) = p(Z|X, \theta) = p(z_1|x_1, \theta)...p(z_N|x_N, \theta) = q_1(z_1)...q_N(z_N)$$

for

$$q_n(z_n) = p(z_n|x_n, \theta)$$

[3] i.i.d.=independent and identically distributed.

Independent observations

- **M-step** becomes:

$$
\begin{aligned}
\hat{\theta} &= \arg\max_{\theta}\{\sum_{Z} q(Z) \ln p(X, Z|\theta)\} \\
&= \arg\max_{\theta}\{\sum_{Z} q(Z) \sum_{n=1}^{N} \ln p(x_n, z_n|\theta)\} \\
&= \arg\max_{\theta}\{\sum_{n=1}^{N} \sum_{z_1,...z_N} q(z_1, ...z_N) \ln p(x_n, z_n|\theta)\} \\
&= \arg\max_{\theta}\{\sum_{n=1}^{N} \sum_{z_1,...z_N} q_1(z_1)...q_N(z_N) \ln p(x_n, z_n|\theta)\} \\
&= \arg\max_{\theta}\{\sum_{n=1}^{N} \sum_{z_n} q_n(z_n) \ln p(x_n, z_n|\theta)\}
\end{aligned}
$$

Table of Contents

Distribution of latent variables

Suppose we add regularization $R(\theta)$ to log-likelihood.

$$L(\theta) = \ln p(X|\theta) + R(\theta) = \ln \sum_Z p(X, Z|\theta) + \lambda R(\theta)$$

$$= \ln \sum_Z q(Z) \frac{p(X, Z|\theta)}{q(Z)} + \lambda R(\theta) \quad \text{(Yensen's inequality)} \quad (3)$$

$$\geq \sum_Z q(Z) \ln \frac{p(X, Z|\theta)}{q(Z)} + \lambda R(\theta) \quad (4)$$

$$= \sum_Z q(Z) \ln p(X, Z|\theta) + \lambda R(\theta) - \sum_Z q(Z) \ln q(Z) = g(\theta) \quad (5)$$

Since $\sum_Z q(Z) \ln q(Z) = const(\theta)$,

$$\widehat{\theta} = \arg\max_\theta g(\theta) = \arg\max_\theta \left\{ \sum_Z q(Z) \ln p(X, Z|\theta) + \lambda R(\theta) \right\}$$

Making lower bound tight

We can select $q(Z)$ so that at fixed θ $L(\theta) = g(\theta)$:

- Since $\ln x$ is strictly concave, equality in inequality (3)-(4) is achieved <=> $U = \mathbb{E}U$ with probability 1.
- This happens when $\frac{p(X,Z|\theta)}{q(Z)} = c$ for some constant c $\forall Z$.
- Using property $\sum_Z q(Z) = 1$ we have

$$c \sum_Z q(Z) = c = \sum_Z p(X, Z|\theta) = p(X|\theta)$$

- So for lower bound $g(\theta)$ to be tight at θ, we need to take

$$q(Z) = \frac{p(X, Z|\theta)}{p(X|\theta)} = p(Z|X, \theta)$$

EM algorithm

INPUT:

training set $X = [x_1, ...x_N]$, convergence criteria, λ, $R(\theta)$

ALGORITHM:

$t = 0$, $\hat{\theta}_0$ - init randomly

repeat until convergence:

E-step: set distribution over latent variables:

$$q(Z) = p(Z|X, \hat{\theta}_t)$$

M-step: improve estimate of θ

$$\hat{\theta}_{t+1} = \arg\max_\theta \{\sum_Z q(Z) \ln p(X, Z|\theta) + \lambda R(\theta)\}$$
$$t = t + 1$$

OUTPUT:

ML estimates $\hat{\theta}_{t+1}$ **for** the training set.

EM algorithm for MAP estimate

- MAP (maximum a posteriori) estimate:
  - θ is a random variable with prior $p(\theta)$
  - $\widehat{\theta} = \arg\max_\theta \ln p(X, \theta) = \arg\max_\theta \ln p(X|\theta) + \ln p(\theta)$
  - this is equivalent to adding regularization $\lambda R(\theta) = \ln p(\theta)$

INPUT:
training set $X = [x_1, ... x_N]$, convergence criteria, prior $p(\theta)$

ALGORITHM:
$t = 0$, $\hat{\theta}_0$ - init randomly

repeat until convergence:
E-step: set distribution over latent variables:
$$q(Z) = p(Z|X, \hat{\theta}_t)$$
M-step: improve estimate of θ
$$\hat{\theta}_{t+1} = \arg\max_\theta \{\textstyle\sum_Z q(Z) \ln p(X, Z|\theta) + \ln p(\theta)\}$$
$$t = t + 1$$