No Coding Farmer

Tao Hu
Department of Computer Science
Peking University
No.5 Yiheyuan Road Haidian District, Beijing, P.R.China
taohu@pku.edu.cn

Abstract

Some Miscellaneous Summary.

Author Info: Taohu ,Peking University

Contents

1 Expectation Maximization Introduction

1.1 EM Induction

$L(\theta) = \sum_{i=1}^{M} logp(X;\theta) = \sum_{i=1}^{m} log \sum_z p(X,Z;\theta)$

let θ_i be some distribution over z's ($\sum_z \theta_i(z) = 1, \theta_i(z) \geq 0$)

$\sum_i logp(X^{(i)};\theta)$

$= \sum_i log \sum_{Z^{(i)}} \theta_i(Z^{(i)}) \frac{p(X^{(i)},Z^{(i)};\theta)}{\theta_i(Z^{(i)})}$

$\geq \sum_i \sum_{Z^{(i)}} \theta_i(Z^i) log \frac{P(X^{(i)}),Z^{(i)};\theta)}{\theta_i(Z^{(i)})} (f(x) = logx \quad is \quad concave.)$

let $\frac{P(X^{(i)}),Z^{(i)};\theta)}{\theta_i(Z^{(i)})} = C$

the equality can be only reached when $\frac{P(X^{(i)}),Z^{(i)};\theta)}{\theta_i(Z^{(i)})}$ is a constant.

we can get: $\sum_i \frac{p(X^{(i)},Z^{(i)};\theta)}{C} = 1$ namely: $\sum_i p(X^{(i)},Z^{(i)};\theta) = C$

further induction: $\theta_i(Z^{(i)}) = \frac{p(X^{(i)},Z^{(i)};\theta)}{\sum_i p(X^{(i)},Z^{(i)};\theta)} = p(Z^{(i)}|X^{(i)};\theta)$

so the procedure of EM algorithm is:

Repeat Until Convergence:

- E-step: for each i,get $Q_i(Z^{(i)}) = p(Z^{(i)}|X^{(i)};\theta)$
- M-step: $\theta := argmax_\theta \sum_i \sum_{Z^i} i(Z^{(i)}) log \frac{p(X^{(i)},Z^{(i)}),\theta)}{Q_i(Z^{(i)})}$

1.2 EM convergence proof

let $l(\theta^{(t)}) = \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) log \frac{p(X^{(i)},Z^{(i)},\theta)}{Q_i^{(t)}(Z^{(i)})}$

then,we have the following inequality:
$l(\theta^{(t+1)})$
$\geq \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) log \frac{p(X^{(i)},Z^{(i)},\theta^{(t+1)})}{Q_i^{(t)}(Z^{(i)})}$
$\geq \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) log \frac{p(X^{(i)},Z^{(i)},\theta^{(t)})}{Q_i^{(t)}(Z^{(i)})}$
$\geq l(\theta^{(t)})$

the first inequality is because :$l(\theta) \geq \sum_i \sum_{Z^{(i)}} Q_i^{(t)}(Z^{(i)}) log \frac{p(X^{(i)},Z^{(i)},\theta)}{Q_i^{(t)}(Z^{(i)})} \forall \theta, Q_i$

the second inequality is because of the maximum of the M-step.

Hence, EM causes the likelihood to converge monotonically.

1.3 Different Writing Style of EM Algorithm

There are many writing style of EM algorithm. here I just mention the book <Statistics Learning Method> by LiHang who is very famous in China.

EM algorithm from LiHang(Li-version):

Algorithm 1 EM from LIHang

Require: observation X,hidden variable Z,joint distribution $P(X,Z|\theta)$,conditional distribution $P(Z|Y,\theta)$

while Not convergence **do**

E-Step: let $\theta^{(i)}$ is the i-th estimate of θ,

$Q(\theta,\theta^{(i)}) = E_z[logP(X,Z|\theta)|X,\theta^{(i)}] = \sum_Z logP(X,Z|\theta)P(Z|X,\theta^{(i)})$

M-step: $\theta^{(i+1)} = argmax_\theta Q(\theta,\theta^{(i)})$

end while

output model parameter θ

it seems that Li-version is different from the above version. however, they are the same. because:

- the above version just consider every data, so that it include subscript i. however Li-version only consider one data.
- the above version can be transformed to Li-version.

  $\sum_Z Q(Z)log\frac{P(X,Z:\theta)}{Q(Z)}$
  $= \sum_Z P(Z|X;\theta^{(t)})log\frac{p(X,Z;\theta)}{p(Z|X;\theta^{(t)})}$
  $= \sum_Z P(Z|X;\theta^t)logP(X,Z;\theta) - \sum_Z P(Z|X;\theta^{(t)})logP(Z|X;\theta^{(t)})$

  as the variable is θ,so $\sum_Z P(Z|X;\theta^{(t)})logP(Z|X;\theta^{(t)}$ can be removed.
- $Q(\theta,\theta^{(i)}) = \sum_Z logP(X,Z|\theta)P(Z|X,\theta^{(i)})$ can be also written as $Q(\theta,\theta^{(i)}) = \sum_Z logP(X,Z|\theta)P(Z,X,\theta^{(i)})$,because X is a observation.

2 EM applications

2.1 Gaussian Mix Model

GMM can be solved by EM. **notice here we use the expectation of EM:**

$Q(\theta,\theta^{(i)})$
$= E_\gamma[logP(y,\gamma|\theta)|y,\theta^{(i)}]$
$= E[\sum_{k=1}^K[n_k log\alpha_k + \sum_{j=1}^N \gamma_{jk}[log\frac{1}{\sqrt{2\pi}} - log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2]]]$
$= \sum_{k=1}^K[(E\gamma_{jk})log\alpha_k + \sum_{j=1}^N(E\gamma_{jk})[log\frac{1}{\sqrt{2\pi}} - log\sigma_k - \frac{1}{2\sigma_k^2}(y_j-\mu_k)^2]]$

here $(E\gamma_{jk})$ can be easily calculated.

$\hat{\mu_k}, \hat{\sigma_k^2}$ can be acquired by derivation.

$\hat{\alpha_k}$ can be acquired by the derivation on the Lagrangian($\sum_i^K \alpha_k = 1$).

2.2 Hidden Markov Model

HMM Learning Method is also called Baum-Welch algorithm.the target is learning $\lambda = (A,B,\pi)$.

Q function is:

$Q(\lambda,\bar{\lambda}) = \sum_I logP(O,I|\lambda)P(O,I|\bar{\lambda})$

$P(O,I,\lambda) = \pi_{i1}b_{i1}(o_1)a_{i1i2}b_{i2}(o_2)...a_{i_{T-1}i_T}b_{i_T}(o_T)$

so the Q function can also be written as:

$$Q(\lambda,\bar{\lambda}) = \sum_I log\pi_{i1}P(O,I|\bar{\lambda}) + \sum_I(\sum_{t=1}^{T-1} loga_{i,i+1})P(O,I|\bar{\lambda}) + \sum_I(\sum_{t=1}^{T} logb_{it}(o_t))P(O,I|\bar{\lambda})$$

note here: I is not only one state. it includes state length from 1 to T,which all start from i_1

so we can solve the maximum of Q function by derivation on the Lagrangian polynomial (because exists these limitations: $\sum_{i=1}^{N}\pi_i = 1, \sum_{j=1}^{N} a_{ij} = 1, \sum_{i=1}^{M} b_i = 1$)

2.3 Naive Bayesian

2.4 other papers

We can use softmax to model transition probability, normal distribution to model emission probability.

it's a good example in Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models, the AIO-HMM can be more complicated,which can be enriched by the graphic model by M.I Jordon.

3 VAE

here is a complete VAE tutorial [1]

$$max \quad logP(x)$$
lhs = $log\int P(x,z)dz$
= $log\int P(x/z)p(z)dz$
= log $\int \frac{P(x/z)}{q(z/x)}q(z/x)p(z)dz$
= $logE_{q(z/x)}[\frac{p(x/z)}{q(z/x)}p(z)]$
jenson's inequality,we can know: $\geq E_{q(z/x)}[log\frac{p(x/z)}{q(z/x)}p(z)]$
=$E_{q(z/x)}[logp(x/z)] + E_{q(z/x)}[log\frac{p(z)}{q(z/x)}]$
= $E_{q(z/x)}[logp(x/z)] - E_{q(z/x)}[log\frac{q(z/x)}{p(z)}]$
= $E_{q(z/x)}[logp(x/z)] - KL(q(z/x)||p(z))$

4 ADMM

HOG feature dimension: 648

mid layer 1 num: 100

mid layer 2 num: 50

output layer: 1

Algorithm 2 ADMM_NN

Inputs:

data number:n=10000,
data dimension: m=648,
hidden layer 1 unit number: a=100
hidden layer 2 unit number: b=50
output layer unit number: 2
initial feature:a_0 m-n dimension,
W_1 : a-m dimension
z_1 : a-n dimension
a_1 : a-n dimension
W_2 : b-a dimension
z_2 : b-n dimension
a_2 : b-n dimension
W_3 : 1-b dimension
z_3 : 1-n dimension
labels:y 1-n dimension
λ : 1-n dimension
activation function h is ReLu.

Initialize:

allocate $\{a_l\}_{l=1}^{L}$,$\{z_l\}_{l=1}^{L}$ with i.i.d Gaussian Distribution,and λ
Cache: $a_0^{\dagger}$
Warm Start:
for i=1,..,100 **do**
  for l=1,2,...,L-1 **do**
$W_l \leftarrow z_l a_{l-1}^{\dagger}$
$a_l \leftarrow (\beta_{l+1} W_{l+1}^T W_{l+1} + \gamma_l I)^{-1}(\beta_{l+1} W_{l+1}^T z_{l+1} + \gamma_l h_l(z_l))$
$z_l \leftarrow argmin_z \gamma_l ||a_l - h_l(z)||^2 + \beta_l ||z - W_l a_{l-1}||^2$

  end for
  $W_L \leftarrow z_L a_{L-1}^{\dagger}$
  $z_L \leftarrow argmin_z l(z, y) + < z, \lambda > + \beta_L ||z - W_L a_{l-1}||^2$
end for
Start ADMM:
while not converge **do**
  for l=1,2,...,L-1
do $W_l \leftarrow z_l a_{l-1}^{\dagger}$
$a_l \leftarrow (\beta_{l+1} W_{l+1}^T W_{l+1} + \gamma_l I)^{-1}(\beta_{l+1} W_{l+1}^T z_{l+1} + \gamma_l h_l(z_l))$
$z_l \leftarrow argmin_z \gamma_l ||a_l - h_l(z)||^2 + \beta_l ||z - W_l a_{l-1}||^2$
  end for
  $W_L \leftarrow z_L a_{L-1}^{\dagger}$
  $z_L \leftarrow argmin_z l(z, y) + < z, \lambda > + \beta_L ||z - W_L a_{l-1}||^2$
  $\lambda \leftarrow \lambda + \beta_L (z_L - W_L a_{L-1})$
end while

---

z_l argmin procedure:

$$z_l = \begin{cases} max(\frac{a_l \gamma_l + W_l a_{l-1} \beta_l}{\gamma_l + \beta_l}, 0) & \text{z} \geq 0 \\ min(W_l a_{l-1}, 0) & \text{z} \leq 0 \end{cases}$$

choose one minimizer z from two choices.

z_L argmin procedure:

when $y_i = 0$:
$f(z) = \beta z^2 - (2\beta w\_a - \lambda) z + max(z, 0)$
$z^* = max(\frac{2\beta w\_a - \lambda - 1}{2\beta}, 0)$ or

$z^* = min(\frac{2\beta w\_a - \lambda}{2\beta}, 0)$
choose one which make f(z) smaller.

when $y_i = 1$:
$f(z) = \beta z^2 - (2\beta w\_a - \lambda)z + max(1 - z, 0)$
$z^* = max(\frac{2\beta w\_a - \lambda}{2\beta}, 1)$ or
$z^* = min(\frac{2\beta w\_a - \lambda + 1}{2\beta}, 1)$
choose one which make f(z) smaller.

5 Key steps you must know when building a DL Framework

5.1 Convolution

Acknowledgments

References

[1] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.