به نام خدا

پروژه درس آشنایی با ماشین لرنینگ

دکتر سجاد امینی

فاز اول

مبین خطیب

۹۹۱۰۶۱۱۴

محمد علی هاشمی فر

۹۹۱۰۷۶۵۸

Theory Question ۱

1. In your own words, explain how the MM algorithm can deal with nonconvex optimization objective functions by considering simpler convex objective functions.

با به حداقل رساندن مکرر یک تابع جایگزین محدب که مرزهای بالایی تابع هدف غیر محدب را در نقطه فعلی محدود می کند، کار می کند.

ایده این است که تابع هدف غیر محدب را با یک تابع محدب ساده‌تر که به راحتی بهینه می‌شود، تقریب بزنیم. در هر تکرار، الگوریتم MM یک تابع جانشین محدب می سازد که تابع هدف غیر محدب را در نقطه فعلی بزرگ می کند.

سپس این تابع جایگزین برای یافتن نقطه بعدی به حداقل می رسد و همینطور الی آخر تا زمانی که به خواسته مد نظرمان برسیم.تابع بهینه ای که قرار است به تابع هدف نزدیک شود در واقع کوچک یا بزرگ میشود تا زمانی که بدانیم حرکت تابع به بالا یا پایین کافی است .

منظور از بالاو پایین این است که زمانی که قرار است تابع را مینیمایز کنیم هدف حداکثر کردن/حداقل کردن است و زمانی که ماکسیمایز کنیم هدف کوچک کردن/حداکثر کردن است.

Theory Question ۲

$$p(y;\theta) = \sum_{k=1}^{K} p_Z(z_k;\theta) p_{Y|Z}(y|Z = z_k;\theta) \quad (۱)$$

متغیر هایی که بولد شده اند درواقع نشان دهنده یک بردار هستند در اینجا نیز ما برای اینکه به معادله۹ نزدیک تر شویم تلاش میکنیم که y را به $\mathbf{y}$ تبدیل کنیم در واقع میخواهیم آن را از لحاظ برداری نشان دهیم آنگاه خواهیم داشت:

$$p(y_1, y_2, ...y_n;\theta) = \sum_{k=1}^{K} p_Z(z_k;\theta) p_{Y|Z}(y_1, y_2, ...y_n|Z = z_k;\theta) \quad (۲)$$

از آنجا که نمونه های $y_1, y_2, ..., y_n$ مستقل از هم هستند در واقع میتوان اشتراک آنها را به صورت حاصلضربشان نوشت و در نتیجه خواهیم داشت:

$$p(y_1, y_2, ...y_n;\theta) = p(y_1;\theta)p(y_2;\theta)....p(y_n;\theta) = \prod_{i=1}^{N} p(y^{(i)};\theta) \quad (۳)$$

که همانطور که گفتیم این همان وکتور بولد y خواهد شد به این شکل:

$$\prod_{i=1}^{N} p(y^{(i)};\theta) = p_{\mathbf{Y}}(y;\theta) \quad (۴)$$

که حالا اگر از دوطرف لگاریتم طبیعی آن را بگیریم و با توجه به این موضوع که لگاریتم ضرب ها جمع لگاریتم ها میشود:

$$ln p_{\mathbf{Y}}(y;\theta) = ln \prod_{i=1}^{N} p(y^{(i)};\theta) = \sum_{i=1}^{N} \ln p(y^{(i)};\theta) \quad (۵)$$

که در واقع در اینجا به معادله وسطی معادله ۹ رسیدیم و با دخیل کردن متغیر های پنهان که همان z ها هستند و با توجه به قانون بیز میتوان به سمت راست معادله ۹ هم دست پیدا کرد:

$$\begin{aligned}
\ln(p_{\mathbf{Y}}(y;\theta)) = \sum_{i=1}^{N} \ln p(y^{(i)};\theta) &= \ln \prod_{i=1}^{N} \left(\sum_{k=1}^{K} p_{Y,Z}(y^{(i)}, Z^{(i)} = k;\theta) \right) \\
&= \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} p_{Y,Z}(y^{(i)}, Z^{(i)} = k;\theta) \right) \\
&= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} p_Z(z_k;\theta) p_{Y|Z}(y|Z = z_k;\theta)
\end{aligned}$$

از معادله آخر واضح است که به مقصودمان رسیدیم و کار تمام است

در زمینه مدل های مخلوط، توزیع احتمال مشترک داده های مشاهده شده Y و متغیرهای پنهان Z را می توان به صورت زیر بیان کرد:

$$p(Y, Z; \theta) = p(Y|Z; \theta)p(Z; \theta)$$

در جایی که θ پارامترهای مدل را نشان می دهد،$p(Y|Z\theta)$تابع احتمال است، و $p(Z; \theta)$ احتمال قبلی متغیرهای پنهان است.

هدف بهینه‌سازی در میکسچر مدل یافتن مقادیر θ است که احتمال داده‌های مشاهده‌شده Y را با توجه به پارامترهای مدل به حداکثر می‌رساند. این را می توان با محاسبه احتمال نهایی Y بدست آورد:

$$p(Y; \theta) = \int p(Y, Z; \theta), dZ$$

که در آن انتگرال بر روی تمام مقادیر ممکن متغیرهای پنهان Z گرفته می شود.

اغلب بهینه سازی تابع درستنمایی مشترک $p(Y, Z; \theta)$ نسبت به تابع احتمال حاشیه $p(Y; \theta)$ به طور مستقیم آسان تر است. این به این دلیل است که تابع درستنمایی مشترک به حاصلضرب احتمالات شرطی بر روی متغیرهای پنهان تجزیه می شود:

$$p(Y, Z; \theta) = \Pi p(y_n, z_n; \theta)$$

جایی که n نقاط داده را در Y نشان میدهد.

این تجزیه امکان استفاده از الگوریتم (EM) را فراهم می کند، که به طور متناوب بین محاسبه امید شرطی متغیرهای پنهان با توجه به داده های مشاهده شده و برآورد فعلی پارامترهای مدل (گام E) و به حداکثر رساندن امید مورد انتظار expected complete log-likelihood با توجه به پارامترهای مدل (مرحله M) تغییر میکند

با بهینه‌سازی$p(Y, Z; \theta)$ با استفاده از الگوریتم EM، می‌توانیم تخمین‌های حداکثر احتمال پارامترهای θ را به دست آوریم که احتمال نهایی p ؛Y) (□ را به حداکثر می‌رسانند. این رویکرد اغلب کارآمدتر و از نظر عددی پایدارتر از بهینه‌سازی مستقیم $p(\theta, Y)$ به دلیل ساختار تابع احتمال در میکسچر مدل است.

Theory Question ۳

3. Read about variational inference (or variational bayesian methods) and compare it with the procedure we used for the EM algorithm (You might want to check Wikipedia for this!).

خواهیم دید variational inference شامل یافتن مدل و پارامترهای مناسبی است که توزیع مشاهدات را به خوبی نشان می دهد. فرض کنید x مشاهدات و θ پارامترهای مجهول یک مدل ML باشد. در MLE، ما سعی می‌کنیم θ_{ML} را پیدا کنیم که احتمال مشاهدات را با استفاده از مدل ML با پارامترهای زیر به حداکثر می‌رساند:

$$\hat{\theta}_{ML} = argmax_{\theta} p(x;\theta) \qquad (۶)$$

به طور معمول، برای حل مسئله بهینه سازی فوق، مسئله به فرضیات کمی نیاز دارد. یک روش معرفی متغیرهای پنهان z است که مسئله را به مسائل فرعی کوچکتر تقسیم می کند. به عنوان مثال، در Gaussian mixture model می‌توانیم تخصیص عضویت خوشه‌ای را به عنوان متغیرهای تصادفی zi برای هر xi معرفی کنیم که در نتیجه این موضوع خواهیم داشت: $(x_i \mid z_i = k) \sim \mathcal{N}(\mu_k, \sigma_k)$ که مدل را بسیار ساده می کند. این روش را می توان به عنوان بسط الگوریتم انتظار_ بیشینه سازی (EM) از تخمین MAP از محتمل ترین مقدار هر پارامتر تا تخمین کاملاً بیزی که کل توزیع پسین را محاسبه می کند مشاهده کرد. از پارامترها و متغیرهای پنهان همانطور که در EM، مجموعه ای از مقادیر پارامتر بهینه را پیدا می کند، و همان ساختار متناوب EM را دارد، بر اساس مجموعه ای از معادلات به هم قفل شده (وابسته متقابل) که به صورت تحلیلی قابل حل نیستند.

برای بسیاری از کاربردها، varitional Bayes راه حل هایی با دقت قابل مقایسه با نمونه برداری گیبس با سرعت بیشتری تولید می کند. با این حال، استخراج مجموعه معادلات مورد استفاده برای به‌روزرسانی مکرر پارامترها، در مقایسه با استخراج معادلات نمونه‌گیری گیبس، اغلب به مقدار زیادی کار نیاز دارد. این مورد حتی برای بسیاری از مدل‌هایی است که از نظر مفهومی کاملاً ساده هستند، همانطور که در زیر در مورد یک مدل پایه غیر سلسله مراتبی با تنها دو پارامتر و بدون متغیر پنهان نشان داده شده است. با استفاده از خواص امیدریاضی ، عبارت $E_{i\neq j}[lnp(Z,X)]$ می‌توان به تابعی از هایپرپارامترهای ثابت توزیع‌های قبلی بر روی متغیرهای پنهان و امید از متغیر های پنهانی که در پارتیشن فعلی نیستند ساده‌سازی کرد.این وابستگی های دایره ای بین پارامترهای توزیع بر روی متغیرهای یک پارتیشن و امیدمتغیرها در پارتیشن های دیگر ایجاد می کند. این به طور طبیعی یک الگوریتم iterative را پیشنهاد می‌کند، بسیار شبیه به EM که در آن امید متغیرهای پنهان به روشی (شاید به صورت تصادفی) مقداردهی اولیه می‌شوند و سپس پارامترهای هر توزیع به نوبه خود با استفاده از مقادیر فعلی امیدها محاسبه می شود، پس از آن امیدریاضی توزیع جدید محاسبه شده به طور مناسب با توجه به پارامترهای محاسبه شده تنظیم می شود. این الگوریتم تضمین شده است که همگرا شود.به عبارت دیگر، برای هر یک از پارتیشن‌های متغیرها، با ساده‌سازی عبارت توزیع بر روی متغیرهای پارتیشن و بررسی وابستگی عملکردی توزیع به متغیرهای مورد نظر، معمولاً می‌توان خانواده توزیع را تعیین کرد . فرمول پارامترهای توزیع بر حسب هایپرپارامترهای توزیع های قبلی و همچنین بر حسب انتظارات توابع متغیرها در پارتیشن های دیگر بیان می شود. معمولاً این امیدها را می توان به توابع امید خود متغیرها ساده کرد. در بیشتر موارد، توزیع سایر متغیرها از خانواده های شناخته شده خواهد بود و فرمول های امید مربوطه را می توان جستجو کرد. این فرمول ها به پارامترهای آن توزیع ها بستگی دارند، که به نوبه خود به امید در مورد سایر متغیرها بستگی دارد. نتیجه این است که فرمول های پارامترهای توزیع هر متغیر را می توان به صورت مجموعه ای از معادلات با وابستگی های متقابل و غیرخطی بین متغیرها بیان کرد. معمولاً نمی توان مستقیماً این سیستم معادلات را حل کرد. با این حال، همانطور که در بالا توضیح داده شد، وابستگی ها یک الگوریتم تکراری ساده را پیشنهاد می کنند که در بیشتر موارد تضمین شده است که همگرا شوند.

هر دو الگوریتم EM و روش‌های بیزی متغیر برای انجام استنتاج احتمالی در مدل‌های پیچیده استفاده می‌شوند. با این حال، تفاوت هایی بین این دو روش وجود دارد:

۱- الگوریتم EM برای تخمین حداکثر احتمال یا حداکثر تخمین های پسینی (MAP) پارامترهای یک مدل با متغیرهای پنهان استفاده می شود، در حالی که از روش های بیزی متغیر برای تقریب توزیع پسین بر روی متغیرهای پنهان استفاده می شود.

۲- الگوریتم EM یک روش بهینه‌سازی تکراری است که به طور متناوب بین تخمین مقادیر متغیرهای پنهان با استفاده از تخمین‌های فعلی پارامترها و به‌روزرسانی پارامترها با استفاده از مقادیر تخمینی متغیرهای نهفته است. در مقابل، روش‌های بیزی متغیر یک کران پایین‌تر در احتمال حاشیه‌ای داده‌ها را با توجه به پارامترهای تغییرات بهینه می‌کنند.

۳- الگوریتم EM فرض می‌کند که متغیرهای پنهان به طور تصادفی lost شده‌اند و مدل قابل پردازش است. روش‌های بیزی متغیر هیچ فرض خاصی در مورد مکانیسم داده‌های از دست رفته ایجاد نمی‌کنند، اما نیاز دارند که مدل به عنوان یک شبکه بیزی مشخص شود.

۴- الگوریتم EM تضمین شده است که به حداکثر محلی احتمال یا برآورد MAP همگرا می شود، اما ممکن است در local maximum به مشکل بخورد. روش‌های بیزی متغیر چنین تضمین‌هایی ندارند اما می‌توانند تخمین‌های بهتری از توزیع پسین بر روی متغیرهای پنهان ایجاد کنند.

۵- الگوریتم EM را می توان برای مجموعه داده های بزرگ و مدل های پیچیده اعمال کرد، اما ممکن است برای همگرا شدن نیاز به تکرارهای زیادی داشته باشد. روش‌های بیزی متغیر از نظر محاسباتی پیچیده تر هستند، اما می‌توانند برای داده‌های با ابعاد بالا و مدل‌های پیچیده کارآمدتر باشند.

Theory Question ۴

4. Compute estimate of parameters for Gaussian Mixture Models for N observed data $x_{i_{i=1}}^{N} = 1$

1. Determine model parameters and initialize them.

در واقع EM الگوریتمی برای یافتن MLE یا MAP برای مسئله های متغیرهای پنهان است اگر بدانیم هر نقطه به چه خوشه‌ای تعلق دارد (مثلاً متغیرهای z_i)می‌توانیم داده‌ها را تقسیم بندی کنیم و MLE را برای هر خوشه به طور جداگانه پیدا کنیم. هر xi با یک متغیر پنهان همراه است: $z_i = (z_{i۱}, ..., z_{iK})$ در نتیجه complete data برای ما به شکل : $(x, z) = (x_i, z_i), i = ۱, ..., n$ خواهد بود و حالا میتوانیم پارامتر ها را ماکزیمم کردن complete data برای log likelihood انجام دهیم پارامتر متغیر پنهان $z_i k$ نشان دهنده سهم امین-k گاوسی در xi است مشتق log-likelihood برای سه متغیر μ_k, σ_k, π_k که:

$\pi = \pi_۱, ..., \pi_k$

$\mu = (\mu_۱, ..., \mu_k)$

$\Sigma = (\Sigma_۱, ..., \Sigma_k)$

برای سه متغیر مشتق آن ها را روی صفر قرار میدهیم تا معادلات مورد استفاده در الگوریتم EM بدست آوریم برای محاسبه پارامتر های مدل که در بالا ذکر شدند ابتدا باید به این نکته توجه کنیم که:

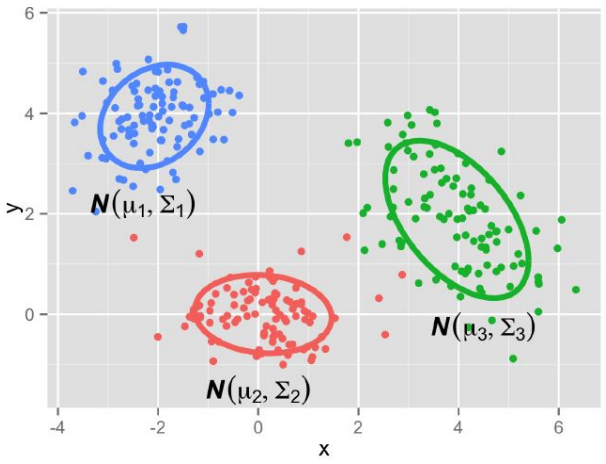
$$-\log Pr(x|\pi, \mu, \Sigma) = -\sum_{i=۱}^{n} \log \sum_{k=۱}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

حالا که متغیر پنهان z نیز ذر این معادله دخیل شود آنگاه شکل معادله به صورت زیر خواهد شد:

$$-\log Pr(x, z|\pi, \mu, \Sigma) = -\sum_{i=۱}^{n} \sum_{k=۱}^{K} \log \pi_k + \log N(x, z|\mu_k, \Sigma_k)$$

در اینجا باید توجه کنیم که π_k و (μ_k, Σ_k) جدا از هم محاسبه خواهند شد و روش محاسبه آنها نیز که ذکر شد با مشتق گیری است که در ادامه آن را پیاده خواهیم کرد. مقدار دهی اولیه را با $\mu_۰, \sigma_۰, I, \pi_۰$ پیگیری میکنیم و محاسبه در مرحله k ام در قسمت۳ انجام خواهد شد

مثال برای سه متغیر سه تایی:



mixture of three Gaussians

2. Compute complete dataset likelihood.

در مرحله E step خواهیم داشت:

$$r_{ik}^t = p(z_{i=k} \mid x_i, \theta^t) = \frac{\pi_k^t p(x_i \mid \theta_k^t)}{\sum_{k'=1}^{K} \pi_{k'}^t p(x_i \mid \theta_{k'}^t)}$$

در مرحله M step خواهیم داشت:

$$\mathrm{LL}^{\mathrm{t}}(\theta) = E\left[\sum_{i=1}^{N} \log p(z_i|\theta) + \log p(x_i|z_i;\theta)\right]$$

$$= E\left[\sum_{i=1}^{N} \log \prod_{k=1}^{K} \pi_k^{(} z_{ik}) + \log \prod_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)^{z_{ik}}\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} E[z_{ik}] \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} E[z_{ik}] \log \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^t \log \pi_k \quad - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^t \log \Sigma_k (\mathbf{x}i - \boldsymbol{\mu}k)^T \Sigma_k^{-1} (\mathbf{x}i - \boldsymbol{\mu}k) + \text{const}$$

در اینجا $z_{ik} = \mathcal{I}(z_i = k)$ یک متغیر one hot از متغیر cathegorical برای z_i میباشد.

3. Find closed-form solution for parameters using EM algorithm.

در اینجا باید توجه کنیم که π_k و (μ_k, Σ_k) جدا از هم محاسبه خواهند شد و روش محاسبه آنها نیز که ذکر شد با مشتق گیری است که در ادامه آن را پیاده خواهیم کرد:

$$۰ = \frac{\partial}{\partial \pi_j}\left[\sum_i \sum_k r_{ik} \log \pi_k + \lambda(۱ - \sum_k \pi_k)\right] \tag{۷}$$

$$\pi_k^{t+۱} = \frac{۱}{n}\sum_i r_{ik}^t = r_k^t/n \tag{۸}$$

$$J(\mu_k, \Sigma_k) = -\frac{۱}{۲}\sum_i r_{ik} \log |\Sigma_k| + (x_i - \mu_k)^T \Sigma_k^{-۱}(x_i - \mu_k) \tag{۹}$$

$$\frac{\partial J(\mu_k, \Sigma_k)}{\partial \mu_k} = \sum_i r_{ik}\Sigma_k^{-۱}(x_i - \mu_k) = ۰ \tag{۱۰}$$

$$\mu_k^{t+۱} = \frac{\sum_i r_{ik}^t x_i}{\sum_i r_{ik}^t} \tag{۱۱}$$

$$\Sigma_k^{t+۱} = \frac{\sum_i r_{ik}^t (x_i - \mu_k^{t+۱})(x_i - \mu_k^{t+۱})^T}{\sum_i r_{ik}^t} = \frac{\sum_i r_{ik}^t x_i x_i^T}{\sum_i r_{ik}^t} - \mu_k^{t+۱}(\mu_k^{t+۱})^T \tag{۱۲}$$

Theory Question ۵

5. Compute estimate of parameters for Categorical Mixture Models for N observed data $x_{i_{i=1}}^{N} = 1$

1. Determine model parameters and initialize them.

در واقع EM الگوریتمی برای یافتن MLE یا MAP برای مسئله های متغیرهای پنهان است اگر بدانیم هر نقطه به چه خوشه‌ای تعلق دارد (مثلاً متغیرهای z_i)می‌توانیم داده‌ها را تقسیم بندی کنیم و MLE را برای هر خوشه به طور جداگانه پیدا کنیم. هر xi با یک متغیر پنهان همراه است: $z_i = (z_{i۱}, ..., z_{iK})$ در نتیجه complete data برای ما به شکل : $(x, z) = (x_i, z_i), i = ۱, ..., n$ خواهد بود و حالا میتوانیم پارامتر ها را ماکزیمم کردن complete data برای log likelihood انجام دهیم پارامتر متغیر پنهان zik نشان دهنده سهم امین-k گاوسی در xi است مشتق log-likelihood برای دو متغیر π_k, θ_k که:

$\pi = \pi_۱, ..., \pi_k$

$\theta = (\theta_۱, ..., \theta_k)$

برای سه متغیر مشتق آن ها را روی صفر قرار میدهیم تا معادلات مورد استفاده در الگوریتم EM بدست آوریم برای محاسبه پارامتر های مدل که در بالا ذکر شدند ابتدا باید به این نکته توجه کنیم که:

$-\log Pr(x|\pi, \theta) = -\sum_{i=۱}^{n} \log \sum_{k=۱}^{K} \pi_k Cat(x|\theta_k)$

حالا که متغیر پنهان z نیز ذر این معادله دخیل شود آنگاه شکل معادله به صورت زیر خواهد شد:

$-\log Pr(x, z|\pi, \theta) = -\sum_{i=۱}^{n} \sum_{k=۱}^{K} \log \pi_k + \log Cat(x|\theta_k)$

2. Compute complete dataset likelihood. 3. Find closed-form solution for parameters using EM algorithm.

در اینجا هر دو پرسش بخش دو و سه را با هم پاسخ خواهیم داد: در ابتدا برای محاسبه E step چون متغیر Z همانطور که در خود جزوه گفته شد و در معادله ۳ جزوه آن را بررسی کردیم متغیر Z دارای توزیع Cat میباشد که پارامتر مربوط آن π میباشد.و مثل سوال قبلی که محاسبه کردیم این بار هم خواهیم داشت:

$$r_{ik}^t = p(z_{i=k} \mid x_i, \theta^t) = \frac{\pi_k^t p(x_i \mid \theta_k^t)}{\sum_{k'=\backslash}^{K} \pi_{k'}^t p(x_i \mid \theta_{k'}^t)}$$

حال برای مرحله M step باید $L^t(\theta))$ را برای توزیع Cat محاسبه و آن را حداکثر کرده و بعد مثل سوال قبل برای یک مرحله بعد پاسخ را محاسبه کنیم: (در معادلات پایین متغیر های θ, π بولد شده در نظر گرفته میشوند منظورمان این است که این متغیر ها بردار هستند)

$$L(t)(\theta) = \sum_i E_{r_i(t)} \left[\ln \left(p_{X,Z}(x_i, z_i; \theta)\right)\right] = \sum_i E_{r_i^{(t)}} \left[\ln \left(p_Z(z_i; \pi) p_{X|Z}(x_i|z_i; \theta)\right)\right] \quad (۱۳)$$

$$= \sum_i E_{r_i^{(t)}} \left[\ln \left(p_Z(z_i; \pi)\right) + \ln \left(p_{X|Z}(x_i|z_i; \theta)\right)\right] \quad (۱۴)$$

$$= \sum_i E_{r_i^{(t)}} \left[\ln \left(\prod_k \pi_k^{I[z_i=k]}\right) + \ln \left(\prod_k Cat(x_n|\theta_k)^{I[z_i=k]}\right)\right] \quad (۱۵)$$

$$= \sum_i E_{r_i^{(t)}} \left[\sum_k I[z_i = k] \ln \pi_k + \sum_k I[z_i = k] \ln \left(Cat(x_i|\theta_k)\right)\right] \quad (۱۶)$$

$$= \sum_i \sum_k E_{r_i^{(t)}} \left[I[z_i = k]\right] \left(\ln \pi_k + \ln \left(Cat(x_i|\theta_k)\right)\right) \quad (۱۷)$$

$$E_{r_i^{(t)}}[\mathcal{I}(z_i = k)] = \sum_{z_i} r_i^{(t)} \mathcal{I}(z_i = k) = r_i^{(t)}(z_i = k) = r_{n,k}^{(t)} \quad (۱۸)$$

با جای گذاری در رابطه اصلی خواهیم داشت:

$$L^{(t)}(\theta) = \sum_i \sum_k E_{r_i^{(t)}} \left[I[z_i = k]\right] \left(\ln \pi_k + \ln \left(Cat(x_i|\theta_k)\right)\right) \quad (۱۹)$$

$$\sum_i \sum_k r_{i,k}^{(t)} \left(\ln \pi_k + \ln \left(Cat(x_i|\theta_k)\right)\right) \quad (۲۰)$$

$$\sum_i \sum_k r_{i,k}^{(t)} \ln \pi_k + \sum_i \sum_k r_{i,k}^{(t)} \ln(Cat(x_i|\theta_k)) \quad (۲۱)$$

با فرض اینکه متغیر Cat ما C کلاسه باشد:(متغیر xi به صوزت وان هات نوشته شده)

$$\theta_k = \begin{bmatrix} \theta_{k,۱} \\ \theta_{k,۲} \\ \cdots \\ \theta_{k,C} \end{bmatrix} \tag{۲۲}$$

$$x_i = \begin{bmatrix} x_{i,۱} \\ x_{i,۲} \\ \cdots \\ x_{i,C} \end{bmatrix} (where x_{i,c} = I[x_i = c]) \tag{۲۳}$$

$$\begin{aligned} L^{(t)}(\theta) &= \sum_i \sum_k r_{i,k}^{(t)} \ln(\pi_k) \\ &+ \sum_i \sum_k r_{i,k}^{(t)} \ln(\mathrm{Cat}(x_i|\theta_k)) \\ = \sum_i \sum_k r_{i,k}^{(t)} \ln(\pi_k) &+ \sum_i \sum_k r_{i,k}^{(t)} \ln \prod_c (\theta_{k,c})^{x_{i,c}} \\ &= \sum_i \sum_k r_{i,k}^{(t)} \ln(\pi_k) \\ &+ \sum_i \sum_k \sum_c r_{i,k}^{(t)} \ln(\theta_{k,c})^{x_{i,c}} \end{aligned}$$

حال برای ماکزیمم کردن عبارت بالا باید نسبت به پارامتر های $\theta_{k,c}, \pi_k$ مشتق بگیریم و برابر صفر قرار دهیم و میدانیم که سیگمای تتا ها روی ۱ تا c برابر یک خواهد شد در نتیجه با یک مسئله بهینه سازی مقید سر و کار داریم و به کمک ضرایب لاگرانژ:

$$L(\theta, \lambda) = L^{(t)}(\theta) + \lambda \left(۱ - \sum_c \theta_{k,c} \right) \tag{۲۴}$$

$$\frac{\partial L(\theta, \lambda)}{\partial \theta_{k,c}} = ۰ \Rightarrow \sum_i r_{i,k}^{(t)} x_{i,c} / \theta_{k,c} = \lambda \Rightarrow \theta_{k,c} = \frac{۱}{\lambda} \sum_i r_{i,k}^{(t)} x_{i,c} \tag{۲۵}$$

$$\frac{\partial L(\theta, \lambda)}{\partial \lambda} = ۰ \Rightarrow \sum_c \theta_{k,c} = ۱ \Rightarrow \frac{۱}{\lambda} \sum_i r_{i,k}^{(t)} \left(\sum_c x_{i,c} \right) \tag{۲۶}$$

$$= \frac{۱}{\lambda} \sum_i r^{(t)} i, k(۱) = ۱ \Rightarrow \lambda = \sum i r_{i,k}^{(t)} = r_k^{(t)} \tag{۲۷}$$

$$r_k^{(t)} = \sum_i r_{i,k}^{(t)} \tag{۲۸}$$

$$\theta_{k,c}^{(t+۱)} = \frac{\sum_i r^{(t)} i, k x_{i,c}}{\sum_i r_{i,k}^{(t)}} = \frac{\sum_i r_{i,k}^{(t)} x_{i,c}}{r_k^{(t)}} \tag{۲۹}$$

$$\theta_k^{(t+۱)} = \frac{۱}{r_k^{(t)}} \begin{bmatrix} \sum_i r_{i,k}^{(t)} x_{i,۱} \\ \sum_i r_{i,k}^{(t)} x_{i,۲} \\ \cdots \\ \sum_i r_{i,k}^{(t)} x_{i,C} \end{bmatrix} \tag{۳۰}$$

$$L(\theta, \lambda) = L^{(t)}(\theta) + \lambda(۱ - \sum_k \pi_k) \Rightarrow \tag{۳۱}$$

$$\frac{\partial L(\theta, \lambda)}{\partial \pi_k} = ۰ \Rightarrow \sum_i \frac{r_{i,k}^{(t)}}{\pi_k} = \lambda \Rightarrow \pi_k = \frac{۱}{\lambda} \sum i r_{i,k}^{(t)} \tag{۳۲}$$

$$\frac{\partial L(\theta, \lambda)}{\partial \lambda} = ۰ \Rightarrow \sum_k \pi_k = ۱ \Rightarrow \frac{۱}{\lambda} \sum_i \sum_k r_{i,k}^{(t)} = ۱ \tag{۳۳}$$

$$\frac{۱}{\lambda} \sum_i \sum_k r_{i,k}^{(t)} = \frac{۱}{\lambda} \sum_i \sum_k \frac{\pi_k^{(t)} p(x_i; \theta_k^{(t)})}{\sum_{k'} \pi_{k'}^{(t)} p(x_i; \theta_{k'}^{(t)})} \tag{۳۴}$$

$$= \frac{N}{\lambda} \tag{۳۵}$$

$$= ۱ \tag{۳۶}$$

$$\pi_k^{(t+۱)} = \frac{۱}{\lambda} \sum_{i=۱} r_{i,k}^{(t)} = \frac{۱}{N} \sum_{i=۱} r_{i,k}^{(t)} = \frac{r_k^{(t)}}{N} \tag{۳۷}$$