$$\max_{u} \operatorname{var}(z) \qquad \|u\|_2^2 = 1$$

$$S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

$$J(w) = \frac{(\tilde m_1 - \tilde m_2)^2}{\tilde s_1^2 + \tilde s_2^2}$$

$$\max_{w} \frac{(\tilde m_1 - \tilde m_2)^2}{\tilde s_1^2 + \tilde s_2^2}$$

$$\tilde m_1 = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad w^T\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = w^T m_1$$

$$\tilde s_1^2 = \sum^{n} z_i z_i^T \qquad w^T\Big(\underbrace{\sum_{i=1}^{n} x_i x_i^T}_{s_1^2}\Big) w = w^T s_1^2 w$$

$$\frac{(w^T m_1 - w^T m_2)^2}{w^T s_1^2 w + w^T s_2^2 w}$$

$$w^T \overbrace{(m_1 - m_2)(m_1 - m_2)^T}^{S_B} w$$

$$w^T (s_1^2 + s_2^2) w$$

$$\max_{w} (w^T S_B w)$$

$$\text{s.t.} \quad w^T S_W w = K$$

$$w^T S_B w - \lambda (w^T S_W w - k)$$

$$2 S_B w - 2\lambda S_W w = 0$$

$$S_W^{-1} S_B w = \lambda w$$

$$\max \lambda w^T S_W w = \max \lambda k$$

Mercer's theorem:

$$\iint g(x) k(x,y) g(y)\, dx\, dy \geq 0 \rightarrow \boxed{\text{Valid kernel}}$$

$$F_T(x) = \sum_{t}^{T} \alpha_t h_t(x) = F_{T-1}(x) + \alpha_T h_T(x)$$

$$\sum_{i=1} w_i^{(T-1)} e^{-y_i \alpha_T h_T(x_i)} = \sum_{i: y_i \neq h_T(x_i)} w_i^{(T-1)} e^{\alpha_T} + \sum_{i: y_i = h_T(x_i)} w_i^{(T-1)} e^{-\alpha_T}$$

$$\frac{dE}{d\alpha_T} = e^{\alpha_T} \sum_{i\in M} w_i^{(T-1)} - e^{-\alpha_T} \sum_{i \in C} w_i^{(T-1)} = 0$$

$$\alpha_T = \frac{1}{2} \ln \frac{\sum_{i\in C} w_i^{(T-1)}}{\sum_{i \in M} w_i^{(T-1)}} \qquad \alpha_T = \frac{1}{2}\ln\frac{1-\epsilon_T}{\epsilon_T}$$

$$\langle \vec w, x_1 x_2 \rangle = 0 \rightarrow w^T(x_2 - x_1) = w^T x_2 - w^T x_1 = \underbrace{w^T x_2 + b}_{} - \underbrace{w^T x_1 - b}_{0} = 0$$

$$\max_{w,b} \operatorname{margin}(w,b)\, h = \left|\|x_1 x_2\|_2 \cos\alpha\right| = \left|\frac{\|w\|_2}{\|w\|_2} \|x_1 x_2\| \cos\right|$$

$$\frac{1}{\|w\|_2} |w^T(x_1 - x_2)| = \frac{1}{\|w\|_2} |\underbrace{w^T x_1 + b}_{1} - \underbrace{w^T x_2 - b}_{-1}|$$

$$\max_{w,b} \frac{2}{\|w\|_2} \quad \text{s.t. } y_i(w^T x_i + b) \geq 1 \qquad \min_{w,b} \|w\|_2^2 \quad \text{s.t. } y_i(w^T x_i + b) \geq 1$$

$$L(w,b,\lambda) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{n} \lambda_i (y_i(w^T x_i + b) + 1)$$

$$\frac{\partial L}{\partial w} = w + \sum_{i=1}^{n} - \lambda_i y_i x_i = 0 \rightarrow \boxed{w = \sum_{i=1}^{n} \lambda_i y_i x_i}$$

$$-\sum_i \lambda_i y_i b \quad \frac{\partial L}{\partial b} = 0 - \sum_{i=1}^{n} \lambda_i y_i = 0 \Rightarrow \boxed{\sum_{i=1}^{n} \lambda_i y_i = 0}$$

$$g(\lambda) = \frac{1}{2}\Big(\sum_i \lambda_i y_i x_i\Big)^T\Big(\sum_i \lambda_i y_i x_i\Big) - \sum_{i=1} i + \sum_{i=1} \lambda_i y_i \Big(\sum_j \lambda_j y_j\Big)$$

$$= \frac{1}{2}\Big(\sum_i\sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j\Big) - \sum_i\sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i$$

$$g(\lambda) = \frac{-1}{2}\sum_i\sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i$$

$$w^* = \sum_{i=1}^{n} \lambda_i y_i x_i \qquad \lambda_i g_i(x^*) = 0 \qquad \underbrace{\lambda_k^*}_{\neq 0} (y_k(w^T x_k + b)) = 0$$

$$\min_{w,b,e} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} e_i \quad \text{s.t. } y_i(w^T x_i + b) \geq 1 - e_i, \quad e_i \geq 0$$

$$\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} e_i + \sum_{i=1} \lambda_i (1 - y_i(w^T x_i + b)) + \sum_{i=1}^{n} \mu_i (-e_i)$$

$$\nabla_{e_j} L = C + (-\lambda_j) - \mu_j = 0 \Rightarrow \mu_j = C - \lambda_j \geq 0 \rightarrow 0 \leq \lambda_j \leq C \qquad \mu_i, \lambda_i \geq 0$$

$$\max_{\lambda} \frac{-1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_i \lambda_i$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0, \quad 0 \leq \lambda_i \leq C$$

$$p(x|\theta) = (1-\alpha) N(x; \mu_0, \Sigma_0) + \alpha \underbrace{N(x; \mu_1, \Sigma_1)}_{\text{component}}$$

$$P(x|\theta) = (1-\alpha) N_0 + \alpha N_1$$

$$P(x,z|\theta) = [(1-\alpha)N(x;\mu_0,\Sigma_0)]^{1-z} [\alpha N(x;\mu_1,\Sigma_1)]^{z}$$

$$E_z[L(\theta)] = E\Big[\sum_{i=1} \log P(x_i, z_i|\theta)\Big] = \sum_{i=1} E_z[\log P(x_i, z_i|\theta)]$$

$$= \sum_{i=1}^{n} E_z[(1-z_i)\log(1-\alpha) + (1-z_i)\log N(x_i|\mu_0,\Sigma_0) + z_i \log\alpha + z_i \log N(x_i|\mu_1,\Sigma_1)]$$

: for expectation

$$= \sum_{i=1}^{n} \log(1-\alpha)(1 - E[z_i]) + \log N(x_i|\mu_0,\Sigma_0)(1 - E[z_i]) + E[z_i]\log\alpha + E[z_i]\log N(x_i|\mu_1,\Sigma_1))$$

$$E[z_i] = E[z]$$

$$E[z] = E[z|x] = P(z=1|x) = \frac{P(x|z=1)P(z=1)}{P(x)} = \frac{N(x|\mu_1^t,\Sigma_1^t)\alpha^t}{(1-\alpha^t)N(x|\mu_0^t,\Sigma_0^t) + \alpha^t N(x|\mu_1^t,\Sigma_1^t)} = \gamma^t$$

$$Q(\theta) = E_z[L(\theta)] = \sum_{i=1}^{n} \log\alpha(1-\gamma_i^t) + \log N(x_i|\mu_0,\Sigma_0)(1-\gamma_i^t) + \gamma_i^t \log\alpha + \gamma_i^t \log N(x_i|\mu_1,\Sigma_1)$$

: M

$$\frac{dQ}{d\alpha} = 0 \rightarrow \sum_{i=1}^{n}\Big(\frac{-1}{1-\alpha}(1-\gamma_i^t) + \frac{\gamma_i^t}{\alpha}\Big) = 0$$

$$\frac{n - \sum \gamma_i^t}{1-\alpha} = \frac{\sum \gamma_i^t}{\alpha} \rightarrow \boxed{\alpha = \frac{\sum_{i=1}^{n} \gamma^t}{n}} \quad **$$

$$P(x|\theta) = \sum_{k=1}^{K} \alpha_k N(x|\mu_k, \Sigma_k) \qquad \sum_{k=1}^{K} \alpha_k = 1 \quad \alpha_k \geq 0 \qquad P(z,z|\theta) = \prod^{K} (\alpha_k N(x_i|\mu_k,\Sigma_k))^{z_k}$$

$$L(\theta) = \log P(D, H|\theta) = \sum_{i=1}^{n} \log P(x_i, z_i|\theta)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}(\log\alpha_k + \log N(x_i|\mu_k,\Sigma_k))$$

$$E_{z_i|x_i}[z_{ik}] = \Pr(z_{ik}=1|x_i,\theta^t)$$

$$= \frac{P(x_i|z_{ik}=1,\theta^t)\,P(z_{ik}=1|\theta^t)}{P(x_i|\theta^t)}$$

$$\frac{N(x_i|\mu_k^t,\Sigma_k^t)\alpha_k^t}{\sum_{k=1}^{K}\alpha_k N(x_i|\mu_k^t,\Sigma_k^t)} = \gamma_{ik}^t$$

$$Q(\theta) = E_{z_i}[L(\theta)] = \sum_{i=1}^{n}\sum_{k=1}^{K} \underbrace{E[z_{ik}]}_{\text{boolean}}(\log\alpha_k + \log N(x_i|\mu_k,\Sigma_k))$$

$$\frac{\partial Q(\theta)}{\partial\alpha_k} = 0$$

$$\boxed{\hat\alpha_k = \frac{\sum \gamma_{ik}^t}{n}} \qquad \boxed{\hat\mu_k = \frac{\sum_{i=1}^{n}\gamma_{ik}^t x_i}{\sum_{i=1}^{n}\gamma_{ik}^t}}$$

$$Q(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}^t(\log\alpha_k + \log N(x_i|\mu_k,\Sigma_k)) - \lambda\Big(\sum_{k=1}^{K}\alpha_k - 1\Big)$$

$$\frac{\partial Q(\theta)}{\partial\alpha_j} = \sum_{i=1}^{n}\gamma_{ij}^t \frac{1}{\alpha_j} - \lambda = 0 \Rightarrow \boxed{\alpha_j = \frac{\sum_{i=1}^{n}\gamma_{ij}^t}{\lambda}} \quad j = 1,\dots,K$$

$$\alpha_j = \frac{\sum_{i=1}^{n}\gamma_{ij}^t}{\sum^{K}\sum^{n}\gamma_{ij}^t} = \frac{\sum_{i=1}\gamma_{ij}^t}{n}$$

سوال ۷ خوشه‌بندی (۱۵ نمره)

می‌خواهیم مجموعه داده‌های $\mathcal{X}$ را به k خوشه تقسیم کنیم. همانطور که می‌دانید، الگوریتم k-means با انتخاب k مرکز دسته به صورت تصادفی آغاز می‌شود. فرض کنید که راهبرد زیر را برای انتخاب k مرکز دسته اولیه داشته باشیم:

مرکز دسته اول را تصادفی انتخاب کرده و مرکز دسته‌های بعدی را به گونه‌ای انتخاب می‌کنیم که از مرکز دسته‌هایی که قبلا انتخاب شده‌اند بیشترین فاصله را داشته باشد.

به کمک یک مثال نشان دهید که این راهبرد الزاما به دسته‌بندی بهینه منجر نمی‌شود. به عبارت دیگر نشان دهید که ممکن است هزینه راهبرد پیشنهادی از هزینه الگوریتم بهینه خوشه‌بندی k-means که در عبارت زیر نمایش داده شده، بیشتر است:

$$\phi = \sum_{x\in\mathcal{X}} \min_{c\in\mathcal{C}} \|x - c\|^2$$

۷ سوال

فرض کنید k=۲ است و داده‌ها در فضای دوبعدی باشند. حال فرض کنید m نمونه در نقطه‌ی (۰و۱-) و m نمونه در نقطه‌ی (۱و۰) قرار دارند و یک نقطه مثلاً در مکان (۱۰۰و۰) قرار دارد. انتخاب بهینه‌ی مرکز دو خوشه این است که نقطه‌ی (۱و۰) و (۰و۱-) به‌عنوان دو مرکز انتخاب شوند. اما الگوریتم پیشنهادی (اگر $m \to \infty$) حتماً نقطه‌ی (۱۰۰و۰) را به‌عنوان یکی از مراکز انتخاب می‌کند. که هزینه‌ی آن با بزرگ شدن m به سمت بی‌نهایت می‌رود.

بنابراین می‌توان تابع توزیع آمیخته را محاسبه کرد:

$$p(x|\theta) = (1-\alpha)P(x;\lambda_0) + \alpha P(x;\lambda_1)$$

چنانچه تابع احتمال کامل را بنویسیم خواهیم داشت:

$$p(x,z|\theta) = [(1-\alpha)P(x;\lambda_0)]^{1-z}[\alpha P(x;\lambda_1)]^{z}$$

مرحله Expectation)

$$E_z[L(\theta)] = E_z\left[\sum_{i=1}^{n}\log p(x_i,z_i|\theta)\right] = \sum_{i=1}^{n} E_z[\log p(x_i,z_i|\theta)]$$

$$= \sum_{i=1}^{n} E_z[(1-z_i)\log(1-\alpha) + (1-z_i)\log P(x_i;\lambda_0) + z\log\alpha + z\log P(x_i;\lambda_1)]$$

$$= \sum_{i=1}^{n}(1-E[z_i])\log(1-\alpha) + (1-E[z_i])\log P(x_i;\lambda_0) + E[z_i]\log\alpha + E[z_i]\log P(x_i;\lambda_1)$$

از طرفی داریم:

$$E[z] = E[z|x] = p(z=1|x) = \frac{p(x|z=1)p(z=1)}{p(x)} = \frac{P(x;\lambda_1)\alpha^{[t]}}{(1-\alpha^{[t]})P\left(x;\lambda_0^{[t]}\right) + \alpha^{[t]}P(x;\lambda_1^{[t]})} = \delta^{[t]}$$

بدین ترتیب بدست می‌آید:

$$Q(\theta) = E_z[L(\theta)]$$

$$Q(\theta) = \sum_{i=1}^{n}\left(1-\delta_i^{[t]}\right)\log(1-\alpha) + \left(1-\delta_i^{[t]}\right)\log P(x_i;\lambda_0) + \delta_i^{[t]}\log\alpha + \delta_i^{[t]}\log P(x_i;\lambda_1)$$

$$\frac{dQ}{d\alpha} = 0$$

$$\sum_{i=1}^{n} -\frac{1}{1-\alpha}\left(1-\delta_i^{[t]}\right) + \frac{1}{\alpha}\delta_i^{[t]}$$

$$\frac{n-\sum\delta_i^{[t]}}{1-\alpha} = \frac{\sum\delta_i^{[t]}}{\alpha}$$

$$\alpha = \frac{\sum\delta_i^{[t]}}{n}$$

$$\frac{dQ}{d\lambda_0} = 0$$

$$\frac{dQ}{d\lambda_0} = \sum_{i=1}^{n}\left(1-\delta_i^{[t]}\right)\frac{d}{d\lambda_0}\log P(x_i;\lambda_0) = 0$$

$$\frac{dQ}{d\lambda_0} = \sum_{i=1}^{n}\left(1-\delta_i^{[t]}\right)\left(\frac{x_i}{\lambda_0}-1\right) = 0$$

$$\lambda_0 = \frac{\frac{\sum x_i}{n} - \frac{\sum\delta_i^{[t]}x_i}{n}}{1-\frac{\sum\delta_i^{[t]}}{n}}$$

$$\lambda_0 = \frac{mean - \frac{\sum\delta_i^{[t]}x_i}{n}}{1-\alpha^{[t]}}$$

$$\frac{dQ}{d\lambda_1} = 0$$

$$\frac{dQ}{d\lambda_1} = \sum_{i=1}^{n}\delta_i^{[t]}\frac{d}{d\lambda_1}\log P(x_i;\lambda_1) = 0$$

$$\frac{dQ}{d\lambda_1} = \sum_{i=1}^{n}\delta_i^{[t]}\left(\frac{x_i}{\lambda_1}-1\right) = 0$$

$$\lambda_1 = \frac{\frac{\sum\delta_i^{[t]}x_i}{n}}{\frac{\sum\delta_i^{[t]}}{n}}$$

$$\lambda_1 = \frac{\frac{\sum\delta_i^{[t]}x_i}{n}}{\alpha^{[t]}}$$

سوال ۴ یادگیری جمعی (*Ensemble Learning*) (۱۵+۱۰ نمره)

الف) (۱۰ نمره) برای یک مسأله رگرسیون، دو مدل $y=h_1(x)$ و $y=h_2(x)$ را با استفاده از کمینه کردن مجموع مجذور خطا آموزش داده‌ایم. حال می‌خواهیم با ترکیب این دو مدل، یک مدل قوی‌تر بسازیم. فرض کنید مجموعه داده‌ی $D = \{(x_1,y_1),\dots,(x_n,y_n)\}$ را در اختیار داریم، اگر مدل جدید را به صورت $y=\alpha h_1(x)+\beta h_2(x)$ در نظر بگیریم، مقدار بهینه‌ی ضرایب α و β را بدست آورید. توجه داشته باشید که $\alpha+\beta$ لزوما ۱ نیست. بدست آوردن دستگاه دو معادله و دو مجهول برای α و β کفایت می‌کند و نیازی به حل دستگاه نیست.

ب) (**اختیاری ۱۵ نمره**) اگر خطای دو مدل $h_1(x)$ و $h_2(x)$ به ترتیب e_1 و e_2 باشد و از رابطه‌ی $h(x)=\frac{1}{2}h_1(x)+\frac{1}{2}h_2(x)$ برای ترکیب این دو مدل استفاده کنیم. نشان دهید که خطای (مجموع مجذور خطا) مدل $h(x)$ در رابطه‌ی زیر صدق می‌کند:

$$e \le \frac{1}{4}(e_1+e_2) + \frac{1}{2}\sqrt{e_1e_2}$$

الف)

$$\min_{\alpha,\beta}\sum_{i=1}^{n}\left(\alpha h_1(x_i)+\beta h_2(x_i)-y_i\right)^2$$

$$\frac{\partial L}{\partial\alpha} = \sum_{i=1}^{n} 2h_1(x_i)\left(\alpha h_1(x_i)+\beta h_2(x_i)-y_i\right) = 0$$

$$\Rightarrow \alpha\sum_{i=1}^{n}h_1^2(x_i) + \beta\sum_{i=1}^{n}h_1(x_i)h_2(x_i) = \sum_{i=1}^{n}h_1(x_i)y_i$$

به طرق مشابه $\frac{\partial L}{\partial\beta} = 0 \Rightarrow \alpha\sum_{i=1}^{n}h_1(x_i)h_2(x_i) + \beta\sum_{i=1}^{n}h_2^2(x_i) = \sum_{i=1}^{n}h_2(x_i)y_i$

$$e = \sum_{i=1}^{n}\left(\frac{h_1+h_2}{2}-y_i\right)^2 = \frac{1}{4}\sum_{i=1}^{n}(h_1-y_i+h_2-y_i)^2$$

$$= \frac{1}{4}\sum_{i=1}^{n}(h_1-y_i)^2+(h_2-y_i)^2 + 2(h_1-y_i)(h_2-y_i)$$

$$= \frac{1}{4}(e_1+e_2) + \frac{1}{2}\sum_{i=1}^{n}(h_1-y_i)(h_2-y_i)$$

$$\le \frac{1}{4}(e_1+e_2) + \frac{1}{2}\sum_{i=1}^{n}|h_1-y_i|\,|h_2-y_i|$$

$$= \frac{1}{4}(e_1+e_2) + \frac{1}{2}\sqrt{\left(\sum_{i=1}^{n}|h_1-y_i|\,|h_2-y_i|\right)^2}$$

$$= \frac{1}{4}(e_1+e_2) + \frac{1}{2}\sqrt{e_1e_2}$$

$$J = \frac{1}{n_1n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2}(y_i-y_j)^2 = \frac{1}{n_1n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2}y_i^2+y_j^2-2y_iy_j$$

$$= \frac{1}{n_1n_2}\left(n_2\sum_{y_i\in Y_1}y_i^2 + n_1\sum_{y_j\in Y_2}y_j^2 - 2\sum_{y_i\in Y_1}\sum_{y_j\in Y_2}y_iy_j\right)$$

$$= \frac{1}{n_1}\sum_{y_i\in Y_1}y_i^2 + \frac{1}{n_2}\sum_{y_j\in Y_2}y_j^2 - \frac{2}{n_1n_2}\sum_{y_i\in Y_1}\sum_{y_j\in Y_2}y_iy_j$$

$$= \frac{1}{n_1}\sum_{y_i\in Y_1}y_i^2 + \frac{1}{n_2}\sum_{y_j\in Y_2}y_j^2 - 2\sum_{y_i\in Y_1}\frac{1}{n_1}y_i\sum_{y_j\in Y_2}\frac{1}{n_2}y_j$$

$$J = \frac{1}{n_1}\sum_{y_i\in Y_1}y_i^2 + \frac{1}{n_2}\sum_{y_j\in Y_2}y_j^2 - 2m_1m_2$$

$$= \frac{1}{n_1}\sum_{y_i\in Y_1}y_i^2 + \frac{1}{n_2}\sum_{y_j\in Y_2}y_j^2 - 2m_1m_2 + (m_1^2+m_2^2) - (m_1^2+m_2^2)$$

$$= \frac{1}{n_1}\left(\sum_{y_i\in Y_1}y_i^2 - n_1m_1^2\right) + \frac{1}{n_2}\left(\sum_{y_j\in Y_2}y_j^2 - n_2m_2^2\right) + (m_1^2+m_2^2-2m_1m_2)$$

$$= \frac{1}{n_1}\left(\sum_{y_i\in Y_1}(y_i-m_1)^2\right) + \frac{1}{n_2}\left(\sum_{y_j\in Y_2}(y_j-m_2)^2\right) + (m_1-m_2)^2$$

$$J = (m_1-m_2)^2 + \frac{1}{n_1}S_1^2 + \frac{1}{n_2}S_2^2$$

سوال ۶ کاهش بُعد (۱۵ نمره)

مجموعه داده‌ی زیر را در نظر بگیرید. با استفاده از *PCA* می‌خواهیم داده‌ها را به فضای یک بُعدی ببریم. جهت برداری که روش *PCA* برای نگاشت به فضای یک بُعدی بدست می‌آورد را محاسبه کنید.

$$D = \left\{\begin{bmatrix}10\\8\end{bmatrix},\begin{bmatrix}8\\2\end{bmatrix},\begin{bmatrix}2\\0\end{bmatrix},\begin{bmatrix}4\\6\end{bmatrix}\right\}$$

$$X = \begin{bmatrix}10 & 8\\ 8 & 2\\ 2 & 0\\ 4 & 6\end{bmatrix} \qquad \mu = \begin{bmatrix}6\\4\end{bmatrix}$$

با کم کردن میانگین از نمونه‌ها داریم:

$$X = \begin{bmatrix}4 & 4\\ 2 & -2\\ -4 & -4\\ -2 & 2\end{bmatrix} \qquad S = \frac{1}{4}X^TX = \frac{1}{4}\begin{bmatrix}40 & 24\\ 24 & 40\end{bmatrix} = \begin{bmatrix}10 & 6\\ 6 & 10\end{bmatrix}$$

$$\det(S-\lambda I) = 0 \Rightarrow \lambda^2 - 20\lambda + 64 = 0 \Rightarrow \lambda_1 = 16,\ \lambda_2 = 4$$

$$u = \begin{bmatrix}u_1\\u_2\end{bmatrix}$$

$$Su = \lambda_1 u \Rightarrow 10u_1 + 6u_2 = 16u_1 \Rightarrow \boxed{u_1 = u_2}$$

$$\|u\| = 1 \Rightarrow u_1^2+u_2^2 = 1 \Rightarrow 2u_1^2 = 1 \Rightarrow u_1 = \pm\sqrt{2}/2$$

پس جهت اصلی به صورت $u = \begin{bmatrix}\frac{\sqrt{2}}{2}\\ \frac{\sqrt{2}}{2}\end{bmatrix}$ یا معکوس آن یعنی $\begin{bmatrix}-\frac{\sqrt{2}}{2}\\ -\frac{\sqrt{2}}{2}\end{bmatrix}$ خواهد شد.

هرکدام از دو بردار بالا که اگر بدست آمده باشد نمره کامل تعلق می‌گیرد. اگر روند حل درست باشد ولی محاسبات اشتباه ٪75 نمره داده می‌شود.

نشان دهید اگر توزیع $P(\theta|\alpha)$ یک conjugate prior برای پارامتر θ باشد، آن‌گاه توزیع مخلوط $\sum_{d=1}^{D}\lambda_d P(\theta|\alpha_d)$ هم یک conjugate prior برای θ خواهد بود.

$$D = \{x_1,\dots,x_n\}$$

$$P(\theta|D) \propto P(D|\theta)P(\theta) = P(D|\theta)\sum_{d=1}^{D}\lambda_d P(\theta|\alpha_d)$$

$$= \sum_{d=1}^{D}\lambda_d\underbrace{P(D|\theta)P(\theta|\alpha_d)} = \sum_{d=1}^{D}\lambda_d P(\theta|\alpha'_d)$$

طبق فرض مسأله به صورت $P(\theta|\alpha'_d)$ می‌شود

یعنی به صورت یک توزیع مخلوط شد که مشابه توزیع پیشین است.

توزیع مخلوط نمایی[۳] زیر را در نظر بگیرید:

$$p(x)=\alpha\lambda_1 e^{-\lambda_1 x}+(1-\alpha)\lambda_2 e^{-\lambda_2 x}$$

با فرض داشتن مجموعه داده‌ی $\{x_1,\dots,x_n\}$ می‌خواهیم با استفاده از روش EM، پارامترهای مدل بالا را که شامل α، λ_1 و λ_2 است، بدست آوریم (نمونه‌ها i.i.d. هستند).

الف) (۵ نمره) اگر متغیر تصادفی پنهان z_i را به صورت زیر تعریف می‌کنیم:

$$z_i=\begin{cases}0 & x_i\ \text{comes from the first component}\\ 1 & \text{otherwise}\end{cases}$$

تابع *log-complete likelihood* را تشکیل دهید:

$$\log p(x_1,z_1,\dots,x_n,z_n|\lambda_1,\lambda_2,\alpha)=?$$

ب) (۱۰ نمره) **مرحله E**: امید ریاضی تابع *log-complete likelihood* نسبت به متغیرهای پنهان را بدست آورید. برای این منظور، مقدار $\mathbb{E}_{z_i|x_i,\alpha^t,\lambda_1^t,\lambda_2^t}[z_i]$ را نیز باید محاسبه کنید.

ج) (۱۰ نمره) **مرحله M**: روابط به روز رسانی پارامتر α و λ_1 را بدست آورید. برای سادگی در روابط، امید ریاضی $\mathbb{E}_{z_i|x_i,\alpha^t,\lambda_1^t,\lambda_2^t}[z_i]$ را که در قسمت قبل بدست آمده با نماد γ_i^t نشان دهید.

الف)

$$\log P(x_1,z_1,\dots,x_n,z_n|\lambda_1,\lambda_2,\alpha)\overset{i.i.d}{=}\sum_{i=1}^{n}\log P(x_i,z_i|\lambda_1,\lambda_2,\alpha)$$

$$P(x_i,z_i|\lambda_1,\lambda_2,\alpha)=\left(\alpha\lambda_1 e^{-\lambda_1 x_i}\right)^{1-z_i}\left((1-\alpha)\lambda_2 e^{-\lambda_2 x_i}\right)^{z_i}$$

$$\Rightarrow \log P(x_i,z_i|\lambda_1,\lambda_2,\alpha)=(1-z_i)\left(\log\alpha\lambda_1 e^{-\lambda_1 x_i}\right)+z_i\log\left(\lambda_2(1-\alpha)e^{-\lambda_2 x_i}\right)$$

$$\Rightarrow \text{log complete likelihood}=\sum_{i=1}^{n}(1-z_i)(\log\alpha+\log\lambda_1-\lambda_1 x_i)+z_i(\log(1-\alpha)+\log\lambda_2-\lambda_2 x_i)$$

ب)

$$Q(\theta,\theta^t)=E\left[\log P(x_1,z_1,\dots,x_n,z_n|\lambda_1,\lambda_2,\alpha)\right]$$

$$=\sum_{i=1}^{n}(1-E[z_i])(\log\alpha+\log\lambda_1-\lambda_1 x_i)+E[z_i](\log(1-\alpha)+\log\lambda_2-\lambda_2 x_i)$$

$$E[z_i]=P(z_i=1|x_i,\alpha^t,\lambda_1^t,\lambda_2^t)=\frac{P(x_i|z_i=1,\alpha^t,\lambda_1^t,\lambda_2^t)\,P(z_i=1|\alpha^t,\lambda_1^t,\lambda_2^t)}{P(x_i|\alpha^t,\lambda_1^t,\lambda_2^t)}$$

$$=\frac{\lambda_2^t e^{-\lambda_2^t x_i}(1-\alpha^t)}{\alpha^t\lambda_1^t e^{-\lambda_1^t x_i}+(1-\alpha^t)\lambda_2^t e^{-\lambda_2^t x_i}}=\gamma_i^t$$

ج)

$$\frac{\partial Q}{\partial\alpha}=\sum_{i=1}^{n}(1-\gamma_i^t)\frac{1}{\alpha}+\gamma_i^t\frac{-1}{1-\alpha}=0$$

$$\Rightarrow (1-\alpha)\sum_{i=1}^{n}(1-\gamma_i^t)=\alpha\sum_{i=1}^{n}\gamma_i^t$$

$$\Rightarrow \boxed{\alpha=1-\frac{\sum_{i=1}^{n}\gamma_i^t}{n}}$$

$$\frac{\partial Q}{\partial\lambda_1}=\sum_{i=1}^{n}(1-\gamma_i^t)\left(\frac{1}{\lambda_1}-x_i\right)\cancel{+\gamma_i^t}=0$$

$$\Rightarrow \boxed{\lambda_1=\frac{n-\sum_{i=1}^{n}\gamma_i^t}{\sum_{i=1}^{n}(1-\gamma_i^t)x_i}}$$

سوال ۲ خوشه‌بندی مبتنی بر بهینه‌سازی (۲۴ نمره)

می‌خواهیم با استفاده از یک روش مبتنی بر بهینه‌سازی، داده‌ها را خوشه‌بندی کنیم. برای این منظور، می‌خواهیم کوچک‌ترین ابرکره‌ای که کل داده‌ها را درون خود جای می‌دهد، بیابیم. برای یافتن این ابرکره، مسأله بهینه‌سازی زیر را باید حل کنیم:

$$\min_{r,c} r^2$$

$$\text{subject to: } \|x_i-c\|^2\le r^2,\quad i=1,\dots,n$$

که در رابطه‌ی بالا، r و c به ترتیب شعاع و مرکز ابرکره است. x_i نمونه‌ی i-ام و n تعداد کل نمونه‌ها می‌باشد.

الف) (۵ نمره) مسأله‌ی بهینه‌سازی بالا را به گونه‌ای تغییر دهید تا این امکان وجود داشته باشد که بعضی از نمونه‌ها خارج از ابرکره قرار بگیرند. (**راهنمایی**: مشابه soft-margin SVM عمل کنید).

ب) (۵ نمره) تابع لاگرانژین مسأله‌ی بهینه‌سازی قسمت (الف) را تشکیل دهید.

ج) (۱۰ نمره) مسأله دوگان را برای مسأله‌ی قسمت (الف) بدست آورید.

د) (۴ نمره) توضیح دهید که در این مدل، ضرایب لاگرانژ برای کدام نمونه‌ها صفر و برای کدام نمونه‌ها بزرگتر از صفر می‌شود.

سوال ۲ الف)

$$\min_{r,c,e}\ r^2+\alpha\sum_{i=1}^{n}e_i$$

$$\text{s.t.}\quad \|x_i-c\|^2\le r^2+e_i\qquad i=1,\dots,n$$

$$e_i\ge 0\qquad i=1,\dots,n$$

ب)

$$L(r,c,e,\lambda,\mu)=r^2+\alpha\sum_{i=1}^{n}e_i-\sum_{i=1}^{n}\lambda_i\left(r^2+e_i-\|x_i-c\|^2\right)-\sum_{i=1}^{n}\mu_i e_i$$

$$\mu_i\ge 0$$

$$\lambda_i\ge 0$$

ج)

$$g(\lambda,\mu)=\min_{r,c,e}L(r,c,e,\lambda,\mu)$$

$$\frac{\partial L}{\partial r}=2r-\sum_{i=1}^{n}2r\lambda_i=0\Rightarrow\boxed{\sum_{i=1}^{n}\lambda_i=1}\ ①$$

$$\frac{\partial L}{\partial c}=\sum_{i=1}^{n}\lambda_i(2c-2x_i)=0\Rightarrow c\underbrace{\sum_{i=1}^{n}\lambda_i}_{1}=\sum_{i=1}^{n}x_i\Rightarrow\boxed{c=\sum_{i=1}^{n}x_i}\ ②$$

$$\frac{\partial L}{\partial e_i}=\alpha-\mu_i-\lambda_i=0\Rightarrow\boxed{\mu_i=\alpha-\lambda_i}\ ③$$

با جایگذاری ①،②،③ در لاگرانژین

$$\Rightarrow g(\lambda,\mu)=\cancel{r^2}+\cancel{\alpha\sum_{i=1}^{n}e_i}-\cancel{\sum_{i=1}^{n}\lambda_i r^2}-\cancel{\sum_{i=1}^{n}\lambda_i e_i}+\sum_{i=1}^{n}\lambda_i\|x_i-\sum_{j=1}^{n}x_j\|^2-\cancel{\sum_{i=1}^{n}(\alpha-\lambda_i)e_i}$$

$$\max_{\lambda}\ \sum_{i=1}^{n}\lambda_i\|x_i-\sum_{j=1}^{n}x_j\|^2$$

مسأله دوگان:

یا

$$\max_{\lambda}\ \sum_{i=1}^{n}\lambda_i x_i^T x_i-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j x_i^T x_j$$

$$\text{s.t.}\quad 0\le\lambda_i\le\alpha\qquad i=1,\dots,n$$

$$\sum_{i=1}^{n}\lambda_i=1$$

د) نمونه‌هایی که روی سطح ابرکره یا خارج ابرکره قرار می‌گیرند، ضرایب لاگرانژ آن‌ها بزرگتر از صفر است.