CHUNYUAN LI
JIARUI XING

CSE 517A
Machine Learning
THW3

1. (30 points) Parameter Learning for Gaussian Processes (GPs)

(a) Note:

$$\mathbf{f} = (f_1, f_2, \ldots, f_n)^T, \boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$$

Then:

$$p(\mathbf{f}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K_{xx}}),$$
$$\boldsymbol{K_{xx}} = K(\boldsymbol{x}, \boldsymbol{x})$$

Therefore:

$$p\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{f}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{K_{xx}} & \boldsymbol{K_{xx}}\\ \boldsymbol{K_{xx}} & \boldsymbol{K_{xx}}\end{bmatrix}\right)$$

$$\Sigma_{\mathbf{f}|\mathbf{f}} = \boldsymbol{K_{xx}} - \boldsymbol{K_{xx}}\boldsymbol{K_{xx}^{-1}}\boldsymbol{K_{xx}} = 0$$

i.e. $\forall i \in \{1,2,\ldots,n\}, cov_{f_i} = \left[\boldsymbol{\Sigma}_{\mathrm{f|f}}\right]_{ii} = 0$

(b) $log\, p(\mathbf{y}|X, \boldsymbol{\theta})$

$$= log\, \mathcal{N}\left(\mathbf{0}, K_y\right)$$

$$= log\left(\frac{1}{\sqrt{(2\pi)^n|K_y|}} e^{-\frac{1}{2}\boldsymbol{y}^T K_y^{-\mathbf{1}}\boldsymbol{y}}\right)$$

$$= -\frac{1}{2}\boldsymbol{y}^T K_y^{-\mathbf{1}}\boldsymbol{y} - \frac{1}{2}\left[n\, log\, 2\pi + log\, |k_y|\right]$$

(c) $\mathrm{K_y} = \mathrm{LL^T} \Rightarrow \mathrm{K_y^{-1}} = (L^T)^{-1}\mathrm{L^{-1}}$

$$log\, p(\mathbf{y}|X, \boldsymbol{\theta})$$
$$= -\frac{1}{2}\boldsymbol{y}^T K_y^{-\mathbf{1}}\boldsymbol{y} - \frac{1}{2}\left[n\, log\, 2\pi + log\, |K_y|\right]$$
$$= -\frac{1}{2}\boldsymbol{y}^T (L^T)^{-1} L^{-1}\boldsymbol{y} - \frac{1}{2}[n\, log\, 2\pi + log\, |LL^T|]$$

$$= -\frac{1}{2}\boldsymbol{y}^T L^T (L\backslash\boldsymbol{y}) - \frac{1}{2}[n\, log\, 2\pi + log\, |L|^2]$$

$$= -\frac{1}{2}\boldsymbol{y}^T \boldsymbol{\alpha} - \frac{1}{2}[n\, log\, 2\pi + 2\, log\, |L|]$$

(d) $\dfrac{d}{d\boldsymbol{\theta}} log\, p(\boldsymbol{y}|X, \boldsymbol{\theta})$

$$= -\frac{1}{2} y^T \frac{d}{d\theta} K_y^{-1} y - \frac{1}{2} \frac{d}{d\theta} log|K_y|$$

$$= \frac{1}{2} y^T K_y^{-1} \frac{d}{d\theta} K_y K_y^{-1} y - \frac{1}{2} tr\left(K_y^{-1} \frac{d}{d\theta} K_y\right)$$

$$= \frac{1}{2} tr\left(\left(K_y^{-1} y y^T K_y^{-1} - K_y^{-1}\right) \frac{d}{d\theta} K_y\right)$$

$$= \frac{1}{2} tr\left(\left(\alpha\alpha^T - (L^T)^{-1} L^{-1}\right) \frac{d}{d\theta} L L^T\right)$$

2. (25 points) K-means Clustering

(a) The two conditions are equivalent to each other. Proof:

(i)⇒(ii):

If assignment do not change, i.e. $[z_i]_\alpha$ in $\mu_\alpha = \frac{\sum_{i=1}^{n}[z_i]_\alpha \boldsymbol{x}_i}{\sum_{i=1}^{n}[z_i]_\alpha}$ do not change, μ_α will not change, i.e. (ii) holds;

(ii) ⇒(i)

If cluster centers do not change, i.e. μ_α in

$$[z_i]_\alpha = \begin{cases} 1 & \text{if } \alpha = \underset{\alpha}{\operatorname{argmin}} \|x_i - \mu_\alpha\|^2 \\ 0 & \text{otherwise} \end{cases}$$

do not change, $[z_i]_\alpha$ will not change, i.e. (i) holds.

(b) K-means always converge. Proof:

(i) The object function will always decrease after updating assignments (until converge, when cluster centers are fixed):

For any data point x_i, since z_i are assigned with the closet cluster center to it, $\sum_{\alpha=1}^{k}[z_i]_\alpha\|x_i - \mu_\alpha\|^2$ will decrease if z_i changes or will not change if z_i does not change. Therefore, the objective function $\sum_{i=1}^{n}\sum_{\alpha=1}^{k}[z_i]_\alpha\|x_i - \mu_\alpha\|^2$ will always decrease or not change.

(ii) The object function will always decrease after updating cluster centers (until converge, when assignments are fixed):

Here we are going to show $\mu_\alpha = \frac{\sum_{i=1}^{n}[z_i]_\alpha \boldsymbol{x}_i}{\sum_{i=1}^{n}[z_i]_\alpha}$ is the global minima for $\mu_\alpha^* = \underset{\mu_\alpha}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{\alpha=1}^{k}[z_i]_\alpha\|x_i - \mu_\alpha\|^2$.

To find the optimal point, let:

$$\frac{d}{d\mu_\alpha}\sum_{i=1}^{n}\sum_{\alpha=1}^{k}[z_i]_\alpha\|x_i - \mu_\alpha\|^2$$

$$= \sum_{i=1}^{n}\frac{d}{d\mu_\alpha}\|x_i - \mu_\alpha\|^2$$

$$= \sum_{i=1}^{n}\frac{d}{d\mu_\alpha}(x_i - \mu_\alpha)^T(x_i - \mu_\alpha)$$

$$= \sum_{i=1}^{n}\frac{d}{d\mu_\alpha}\left(x_i^T x_i + \mu_\alpha^T\mu_\alpha - 2x_i^{T\mu_\alpha}\right)$$

$$= \sum_{i=1}^{n}\frac{d}{d\mu_\alpha}(x_i^T x_i + \mu_\alpha^T\mu_\alpha - 2x_i^T\mu_\alpha)$$

$$= \sum_{i=1}^{n}(2\mu_\alpha - 2x_i)$$

$$= 2n\mu_\alpha - 2\sum_{i=1}^{n}x_i = 0$$

$$\Rightarrow \mu_\alpha = \frac{1}{n}\sum_{i=1}^{n} x_i$$

And:

$$\frac{d^2}{d\mu_\alpha^2}\sum_{i=1}^{n}\sum_{\alpha=1}^{k}[z_i]_\alpha \|x_i - \mu_\alpha\|^2$$

$$= \frac{d}{d\mu_\alpha}\left[2n\mu_\alpha - 2\sum_{i=1}^{n} x_i\right]$$

$$= 2n > 0$$

Which means the objective function is convex (with fixed assignments) and $\mu_\alpha = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the optimal solution. Therefore, the object function will always decrease after updating cluster centers.

(iii) From (i) and (ii) we know that the objective function will always decease in both two steps of k-means until assignments or cluster centers does not change, the algorithm must converge to some local minima point.

(c) The is possible. A case is when the cluster number user assigned is even greater than the number of data points. …

(d) It is impossible to have non-convex clusters. In figure 2.1, μ_1 and μ_2 are cluster centers of two clusters, C_1 and C_2, that share a common border, and the grey dashed line is the bisector of the two clusters. Then we choose 2 data points x_1 and x_2 from C_1 randomly, and they must lie on the same side of the bisector with μ_1 since they are closer to μ_1 than μ_2. If we take any x_3 between x_1 and x_2, by geometry it must lie on the same side of the bisector with x_1 and x_2, and thus it must be assigned to C_1, i.e. the cluster is convex.

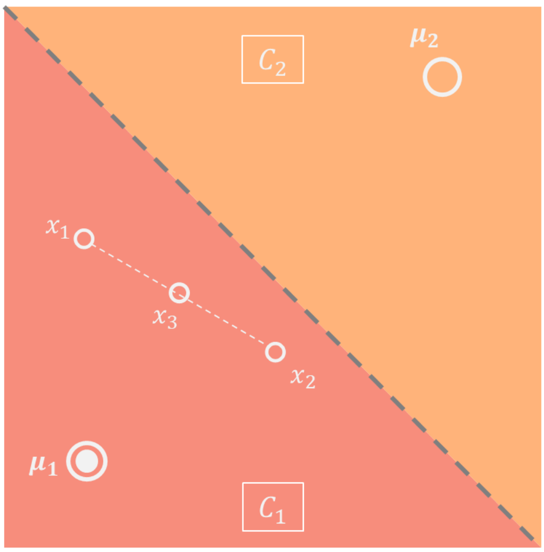

Figure 2.1

3. **(35 points) Expectation-Maximization (EM) for Mixture Model Clustering**

(a) $Pr(x \in j-\text{th})$

$$= [z]_j$$

$$= \frac{p(x|\boldsymbol{\theta_j})\pi_j}{\sum_{i=1}^{k} p(x|\boldsymbol{\theta_i})\pi_i}$$

(b) $\mathcal{L}(X|\Theta)$

$$= \prod_{i=1}^{n} p(x_i|\Theta)$$

$$= \prod_{i=1}^{n} g(x_i|\Theta)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j p(x_i|\boldsymbol{\theta_j})$$

(c)

Algorithm 3.1 General Mixture Model

Initialize $\theta_1, \ldots, \theta_k$

Repeat

for all i & α **do**

$[z_i]_\alpha = \frac{p(x_i|\theta_\alpha)\pi_\alpha}{\sum_{i=1}^{k} p(x_i|\theta_l)\pi_l}$

end for

for $\alpha = 1, \ldots k$ **do**

$\theta_\alpha^* = \underset{\theta_\alpha}{argmax}\, \mathcal{L}(X|\Theta) = \underset{\theta_\alpha}{argmax} \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j p(x_i|\boldsymbol{\theta_j})$

$\pi_\alpha = \frac{1}{n} \sum_{i=1}^{n} [z_i]_\alpha$

end for

$\Theta^* = (\theta_1^*, \ldots, \theta_k^*)$

if $\mathcal{L}(X|\Theta^*) - \mathcal{L}(X|\Theta) > \epsilon$

$\Theta = \Theta^*$

else

exit

(d) Replace $p(x_i|\theta_j)$ in $\mathcal{B}$ with that of Bernoulli distribution:

$$\mathcal{B} = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \log\left(\frac{\pi_j p(x_i|\boldsymbol{\theta_j})}{w_{ij}}\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \log\left(\frac{\pi_j \prod_{m=1}^{d} (\theta_{jm})^{x_{im}} (1-\theta_{jm})^{1-x_{im}}}{w_{ij}}\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \left[log\, \pi_j + \sum_{m=1}^{d} x_{im}\, log\, \theta_{jm} + \sum_{m=1}^{d} (1 - x_{im})\, log\big(1 - \theta_{jm}\big) - log\, w_{ij} \right]$$

To find the MLE estimator, let $\frac{d}{d\theta_{jm}} \mathcal{B} = 0$, i.e.:

$$\frac{d}{d\theta_{jm}} \mathcal{B}$$

$$= \sum_{i=1}^{n} w_{ij} \left[\frac{x_{im}}{\theta_{jm}} - \frac{1 - x_{im}}{1 - \theta_{jm}} \right] = 0$$

$$\Rightarrow \theta_{jm} = \frac{\sum_{i=1}^{n} w_{ij} x_{im}}{\sum_{i=1}^{n} w_{ij}}$$