CHUNYUAN LI
JIARUI XING

CSE 517A

Machine Learning

THW3

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

(a) Note:

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

Then:

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

Therefore:

$$p\left(\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{f}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{K_{xx}} & \boldsymbol{K_{xx}}\\ \boldsymbol{K_{xx}} & \boldsymbol{K_{xx}}\end{bmatrix}\right)$$
$$\Sigma_{\boldsymbol{f}|\boldsymbol{f}} = \boldsymbol{K_{xx}} - \boldsymbol{K_{xx}}\boldsymbol{K_{xx}^{-1}}\boldsymbol{K_{xx}} = 0$$

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

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

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

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

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

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

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

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

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

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

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

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

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, z_i changes only if updated z_i can decrease $\sum_{\alpha=1}^k [z_i]_\alpha \|x_i - \mu_\alpha\|^2$. Otherwise, z_i does not change and so as $\sum_{\alpha=1}^k [z_i]_\alpha \|x_i - \mu_\alpha\|^2$. Therefore, the objective function $\sum_{i=1}^n \sum_{\alpha=1}^k [z_i]_\alpha \|x_i - \mu_\alpha\|^2$ will either decrease or keep unchanged.

(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:

$$\begin{aligned}
&\frac{d}{d\mu_\alpha} \sum_{i=1}^n \sum_{\alpha=1}^k [z_i]_\alpha \|x_i - \mu_\alpha\|^2 \\
&= \sum_{i=1}^n \frac{d}{d\mu_\alpha} \|x_i - \mu_\alpha\|^2 \\
&= \sum_{i=1}^n \frac{d}{d\mu_\alpha} (x_i - \mu_\alpha)^T (x_i - \mu_\alpha) \\
&= \sum_{i=1}^n \frac{d}{d\mu_\alpha} \left(x_i^T x_i + \mu_\alpha^T \mu_\alpha - 2x_i^{T\mu_\alpha}\right) \\
&= \sum_{i=1}^n \frac{d}{d\mu_\alpha} (x_i^T x_i + \mu_\alpha^T \mu_\alpha - 2x_i^T \mu_\alpha) \\
&= \sum_{i=1}^n (2\mu_\alpha - 2x_i)
\end{aligned}$$

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

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

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) It is possible to generate empty cluster. Eg. Empty cluster exists when the user defined cluster number greater than the number of data points.

If the number of clusters is wrong, it might generate empty cluster as well. For an example, in the following figure, we set the number of clusters to be 3. The shapes of the points are meant to show that there are actually 2 natural clusters. But we do not know the natural number of clusters before running a K-means algorithm, so currently assume we have chosen k to be 3.

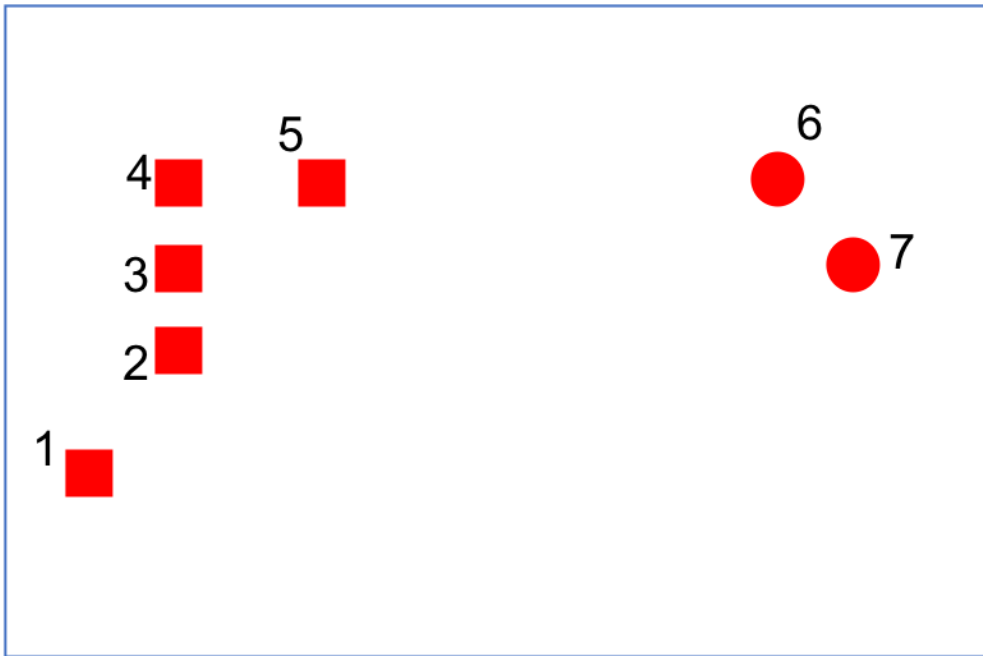

Figure 1 example - initialization

Next assume that points 1, 6, and 7 as initial cluster centers.

At the end of first iteration points 1,2,3 and 4 will be in one cluster. 5 and 6 will be in another cluster. And 7 will be in the last cluster. Note here that the distance between 1 and 5 is larger than the distance between 5 and 6 and so 5 is assigned to the cluster represented by 6 (blue cross). Update the cluster centers and the following figure 2 shows the centers (approximately) and the clusters at the end of first step.

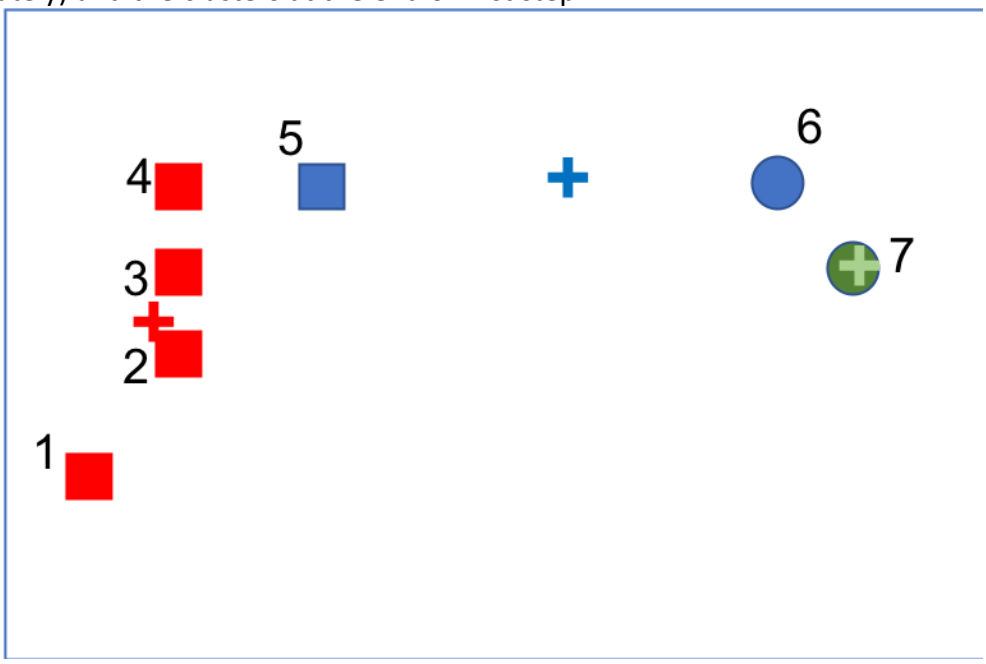

Figure 2 example - after one iteration

Now, the cluster center for the red cluster moved closer to point 5 due to 2, 3 and 4. And the cluster center for the blue cluster moved away from 6 due to point 5. In the next

iteration point 5 will decide that it is closer to the red cluster and point 6 will decide that it is closer to the green cluster. This will cause blue cluster to be empty as shown in figure 3.

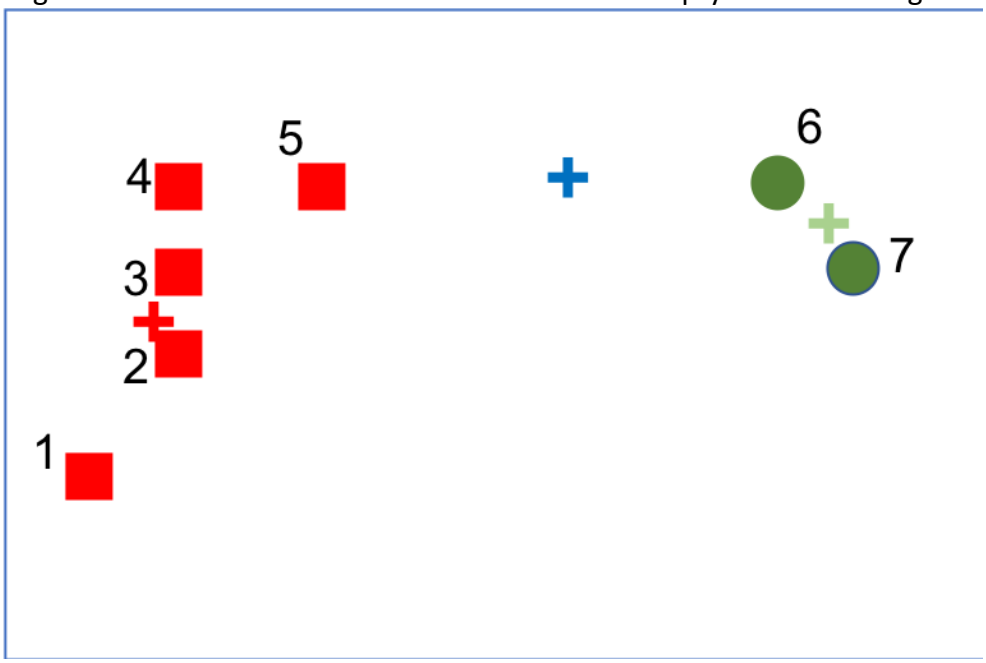

Figure 3 example - after 2nd iteration

(d) It is impossible to have non-convex clusters. In figure 4, μ_1 and μ_2 are cluster centers of two clusters, C_1 and C_2 are two clusters that share a common border, and the black dashed line is the bisector of the two clusters. We choose 2 data points x_1 and x_2 from C_1 randomly. 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.

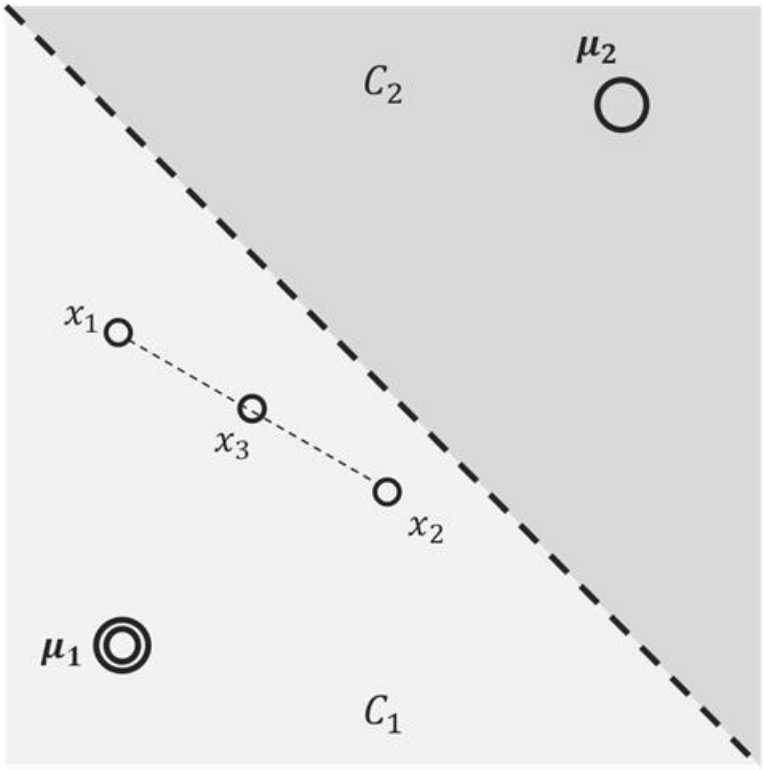

Figure 4

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

(a) $Pr(x \in j^{th}\ distribution)$

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

Where $p(x|\mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^n \Sigma_j}} \exp(-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j))$

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

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

Log likelihood

$$log\mathcal{L}(X|\Theta) = \sum_{i=1}^{n} \log[\sum_{j=1}^{k} \pi_j p(x_i|\theta_j)]$$

(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 $\mathrm{p}(\mathrm{x_i}|\theta_\mathrm{j})$ in $\mathcal{B}$ with that of Bernoulli distribution:

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

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

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

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

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

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

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