FALL17 10-701 Homework 4 Recitation 1

Ellen Vitercik   Matthew Oresky   Jessica Lin

GMM Question 1

Assume you have points that are generated by one of two possible Gaussian distributions. Which of the following are true?

a) We know how to get a globally optimal solution by deriving the maximum likelihood estimate analytically.

b) Using the EM algorithm to solve this problem, we assume that we know from which Gaussian each point originated.

c) Once the EM algorithm has converged, we know for certain from which Gaussian each point originated.

d) The EM algorithm for this problem guarantees that the likelihood of the data never decreases from one iteration to the next.

GMM Question 1

Assume you have points that are generated by one of two possible Gaussian distributions. Which of the following are true?

a) We know how to get a globally optimal solution by deriving the maximum likelihood estimate analytically.

b) Using the EM algorithm to solve this problem, we assume that we know from which Gaussian each point originated.

c) Once the EM algorithm has converged, we know for certain from which Gaussian each point originated.

d) The EM algorithm for this problem guarantees that the likelihood of the data never decreases from one iteration to the next.

Answer

Answer: (d). A - EM doesnt give the globally optimal solution. B - We can start out with one of the Gaussians being more likely for some points, but we dont know for sure. C - After convergence, we only know the probability values of belonging to a particular Gaussian.

GMM Question 2

Which of the following are true about the EM algorithm as applied to a Gaussian Mixture Model?

a) The choice of initial values of parameters of the Gaussian affects the final estimates.

b) The algorithm is guaranteed to converge.

c) The algorithm is guaranteed to converge to a global maxima.

d) The estimate of the parameters obtained at the end is the Maximum Likelihood Estimate.

GMM Question 2

Which of the following are true about the EM algorithm as applied to a Gaussian Mixture Model?

a) The choice of initial values of parameters of the Gaussian affects the final estimates.

b) The algorithm is guaranteed to converge.

c) The algorithm is guaranteed to converge to a global maxima.

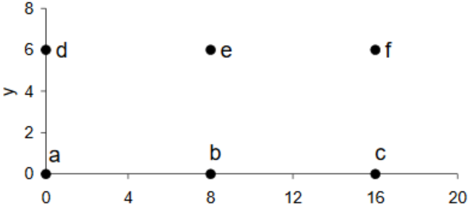
d) The estimate of the parameters obtained at the end is the Maximum Likelihood Estimate.

Answer

A and B are true. C - EM doesnt give the globally optimal solution. D - We cannot solve GMM in closed form to get a clean maximum likelihood expression

K-means Question

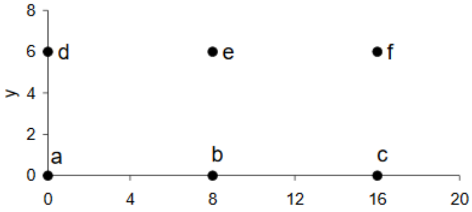
N.B.: Starting cluster centers can be any 3 of the 6 points given.



3-partition	Is it stable?	An example 3-starting configuration that can arrive at the 3-partition after 0 or more iterations of k-means (or write "none" if no such 3-starting configuration)	The number of unique starting configurations that can arrive at the 3-partition
$\{a, b, e\}, \{c, d\}, \{f\}$			
$\{a, b\}, \{d, e\}, \{c, f\}$			
$\{a, d\}, \{b, e\}, \{c, f\}$			
$\{a\}, \{d\}, \{b, c, e, f\}$			
$\{a, b\}, \{d\}, \{c, e, f\}$			
$\{a, b, d\}, \{c\}, \{e, f\}$			

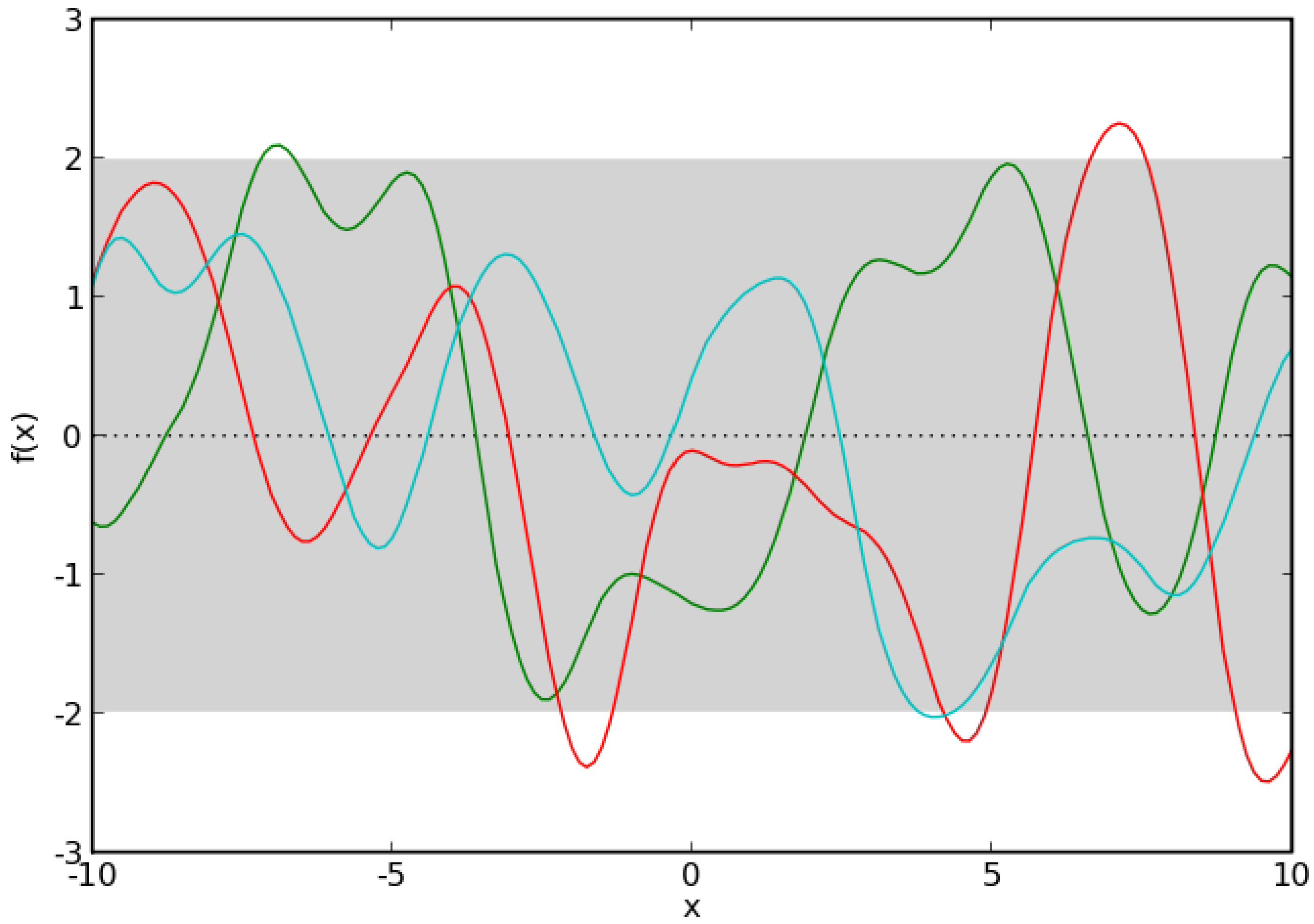
K-means Solution

N.B.: Starting cluster centers can be any 3 of the 6 points given.



3-partition	Stable?	An example 3-starting configuration that can arrive at the 3-partition after 0 or more iterations of k-means (or write "none" if no such 3-starting configuration exists)	# of unique 3-starting configurations that arrive at the 3-partition
$\{a,b,e\},\{c,d\},\{f\}$	N	none	0
$\{a,b\},\{d,e\},\{c,f\}$	Y	{b, c, e}	4
$\{a,d\},\{b,e\},\{c,f\}$	Y	{a, b, c}	8
$\{a\},\{d\},\{b,c,e,f\}$	Y	{a, b, d}	2
$\{a,b\},\{d\},\{c,e,f\}$	Y	none	0
$\{a,b,d\},\{c\},\{e,f\}$	Y	{a, c, f}	1

A *Gaussian process* is a distribution over functions $f\colon \mathbb{R}^d \to \mathbb{R}$.



A *Gaussian process* is a distribution over functions $f: \mathbb{R}^d \to \mathbb{R}$.

It has a mean function:

$$m(\boldsymbol{x}) = \mathbb{E}_{f\sim D}[f(\boldsymbol{x})]$$

A *Gaussian process* is a distribution over functions $f: \mathbb{R}^d \to \mathbb{R}$.

It has a mean function:

$$m(\boldsymbol{x}) = \mathbb{E}_{f\sim D}[f(\boldsymbol{x})]$$

and a covariance function:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{f\sim D}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))]$$

You need to **two things** to define a Gaussian process.

You need to choose the form of:

1. The mean function $m(\boldsymbol{x})$

2. The covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$

How do I choose these functions?

You've chosen the mean and the covariance functions.

You get a training set of labeled points:

$$\{(\boldsymbol{x}_1, f_1), (\boldsymbol{x}_2, f_2)\}$$

You've chosen the mean and the covariance functions.

You get a training set of labeled points:

$$\{(\boldsymbol{x}_1, f_1), (\boldsymbol{x}_2, f_2)\}$$

How do I label a new point $\boldsymbol{x}$?

Let $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

By definition of a Gaussian process,

$$\begin{bmatrix} f_1 \\ f_2 \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x}_1) \\ m(\boldsymbol{x}_2) \\ m(\boldsymbol{x}) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & k(\boldsymbol{x}_1, \boldsymbol{x}_3) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & k(\boldsymbol{x}_2, \boldsymbol{x}_3) \\ k(\boldsymbol{x}_3, \boldsymbol{x}_1) & k(\boldsymbol{x}_3, \boldsymbol{x}_2) & k(\boldsymbol{x}_3, \boldsymbol{x}_3) \end{bmatrix} \right)$$

How do I label a new point $\boldsymbol{x}$?

Question.

You decide $m(\boldsymbol{x}) = \mathbf{0}$ and $k(x, x') = \min\{x, x'\}$.

You see your training set:

$$\boldsymbol{x}_1 = 1, f_1 = 2$$
$$\boldsymbol{x}_2 = 2, f_2 = 5$$

You see a new datapoint $\boldsymbol{x} = 0.5$. Again, let $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

Write down the joint distribution for:

$$\begin{bmatrix} f_1 \\ f_2 \\ f_* \end{bmatrix}$$

Answer.

You decide $m(\boldsymbol{x}) = \mathbf{0}$ and $k(x, x') = \min\{x, x'\}$.

You see your training set:

$$\begin{gathered}\boldsymbol{x}_1 = 1, f_1 = 2 \\ \boldsymbol{x}_2 = 2, f_2 = 5\end{gathered}$$

You see a new datapoint $\boldsymbol{x} = 0.5$. Again, let $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

Write down the joint distribution for:

$$\begin{bmatrix} f_1 \\ f_2 \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0.5 \\ 1 & 2 & 0.5 \\ 0.5 & 0.5 & 0.5 \end{bmatrix}\right)$$

Let $\boldsymbol{x}$ be an unlabled example and $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

Let $X = \begin{bmatrix} - & \boldsymbol{x}_1 & - \\ \vdots & \ddots & \vdots \\ - & \boldsymbol{x}_m & - \end{bmatrix}$ and $f = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix}$.

Then $\begin{bmatrix} \boldsymbol{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_m) \\ m(\boldsymbol{x}) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_m, \boldsymbol{x}_1) & k(\boldsymbol{x}, \boldsymbol{x}_1) \\ \vdots & \ddots & \vdots & \vdots \\ k(\boldsymbol{x}_1, \boldsymbol{x}_m) & \cdots & k(\boldsymbol{x}_m, \boldsymbol{x}_m) & k(\boldsymbol{x}, \boldsymbol{x}_m) \\ k(\boldsymbol{x}_1, \boldsymbol{x}) & \cdots & k(\boldsymbol{x}, \boldsymbol{x}_m) & k(\boldsymbol{x}, \boldsymbol{x}) \end{bmatrix} \right)$

Let $\boldsymbol{x}$ be an unlabled example and $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

Let $k(X,X) = \begin{bmatrix} k(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_m,\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_1,\boldsymbol{x}_m) & \cdots & k(\boldsymbol{x}_m,\boldsymbol{x}_m) \end{bmatrix}$ and $k(X,\boldsymbol{x}) = \begin{bmatrix} k(\boldsymbol{x},\boldsymbol{x}_1) \\ \vdots \\ k(\boldsymbol{x},\boldsymbol{x}_m) \end{bmatrix}$

Then $\begin{bmatrix} \boldsymbol{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_m) \\ m(\boldsymbol{x}) \end{bmatrix}, \begin{bmatrix} k(X,X) & k(X,\boldsymbol{x}) \\ k(X,\boldsymbol{x})^T & k(\boldsymbol{x},\boldsymbol{x}) \end{bmatrix} \right)$

Let $\boldsymbol{x}$ be an unlabled example and $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

Let $k(X, X) = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_m, \boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_1, \boldsymbol{x}_m) & \cdots & k(\boldsymbol{x}_m, \boldsymbol{x}_m) \end{bmatrix}$ and $k(X, \boldsymbol{x}) = \begin{bmatrix} k(\boldsymbol{x}, \boldsymbol{x}_1) \\ \vdots \\ k(\boldsymbol{x}, \boldsymbol{x}_m) \end{bmatrix}$

$$f_* \mid X, \boldsymbol{f}, \boldsymbol{x}$$
$$\sim \mathcal{N}\begin{pmatrix} m(\boldsymbol{x}) + k(X, \boldsymbol{x})^T k(X, X)^{-1}\big(\boldsymbol{f} - m(X)\big), \\ k(\boldsymbol{x}, \boldsymbol{x}) - k(X, \boldsymbol{x})^T k(X, X)^{-1} k(X, \boldsymbol{x}) \end{pmatrix}$$

Question.

You decide $m(\boldsymbol{x}) = \mathbf{0}$ and $k(x, x') = \min\{x, x'\}$.

You see your training set:

$$\boldsymbol{x}_1 = 1, f_1 = 2$$
$$\boldsymbol{x}_2 = 2, f_2 = 5$$

You see a new datapoint $\boldsymbol{x} = 0.5$. Again, let $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

Write down the joint distribution for:

$$f_* \mid X, \boldsymbol{f}, \boldsymbol{x}$$

Hint: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

Answer.

You decide $m(\boldsymbol{x}) = \mathbf{0}$ and $k(x, x') = \min\{x, x'\}$.

You see your training set:

$$\boldsymbol{x}_1 = 1, f_1 = 2$$
$$\boldsymbol{x}_2 = 2, f_2 = 5$$

You see a new datapoint $\boldsymbol{x} = 0.5$. Again, let $f_* = f(\boldsymbol{x})$, where f is sampled from our Gaussian process.

Write down the joint distribution for:

$$f_* \mid X, \boldsymbol{f}, \boldsymbol{x} \sim \mathcal{N}(1, 0.25)$$

You need to **two things** to define a Gaussian process.

You need to choose the form of:

1. The mean function $m(\boldsymbol{x})$

2. The covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$

How do I choose these functions?

Squared exponential kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2}\right)$$

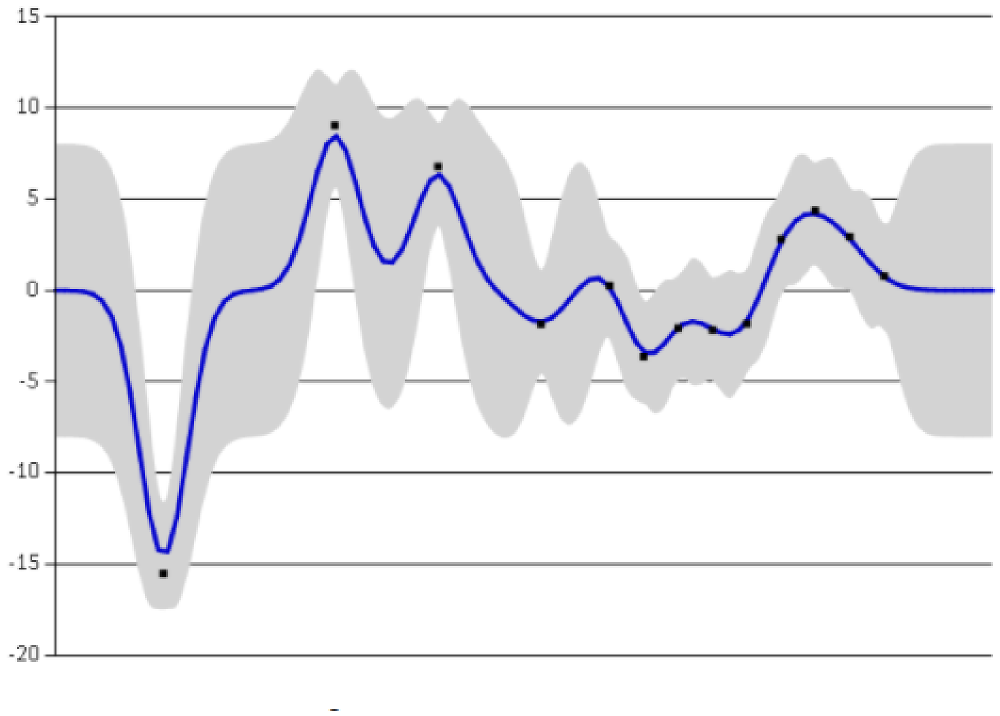
This kernel is **infinitely differentiable**.

It is appropriate for modelling **very smooth** functions.

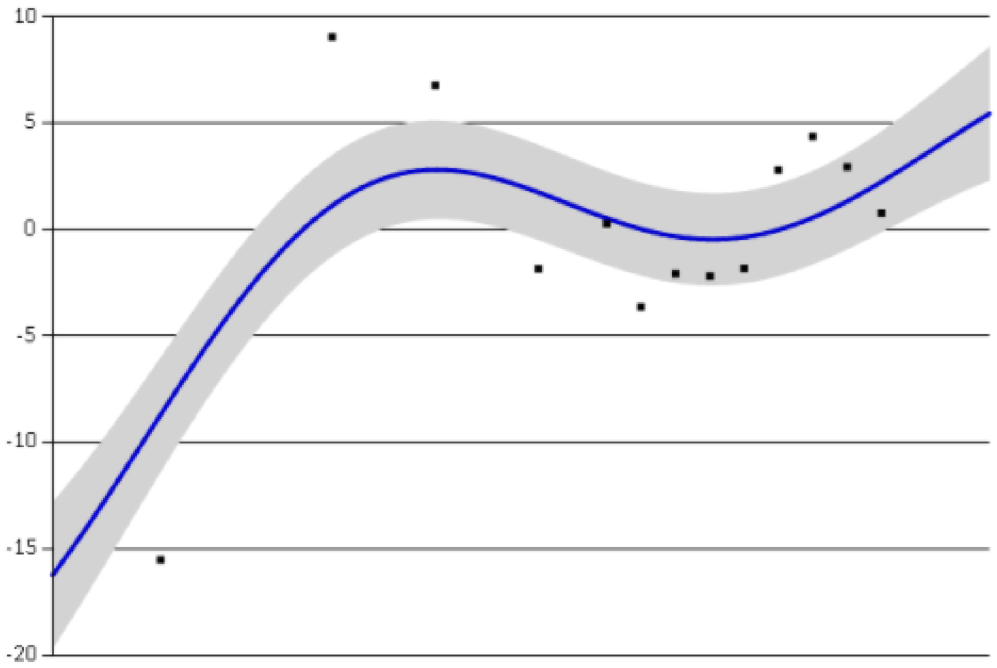
Question.

What happens when ℓ is small? What about when it's large?

As $\ell \to 0, k(\boldsymbol{x}, \boldsymbol{x}') \to 0$. The values look more and more independent.



As $\ell \to \infty, k(\boldsymbol{x}, \boldsymbol{x}') \to 1$. The values look more and more dependent.



Periodic kernel

$$k(x, x') = \exp\left(-\frac{2\sin(\pi|x - x'|/p)^2}{\ell^2}\right)$$