

6.790 Homework 6

Please hand in your work via Gradescope via the link at <https://gradml.mit.edu/info/homeworks/>. If you were not added to the course automatically, please use Entry Code R7RGGX to add yourself to Gradescope. Make sure to assign the problems to the corresponding pages in your solution when submitting via Gradescope.

1. Latex is not required, but if you are hand-writing your solutions, please write clearly and carefully. You should include enough work to show how you derived your answers, but you don't have to give careful proofs.
2. Homework is due on Thursday November 21 at 11PM.
3. Lateness and extension policies are described at https://gradml.mit.edu/info/class_policy/.

Contents

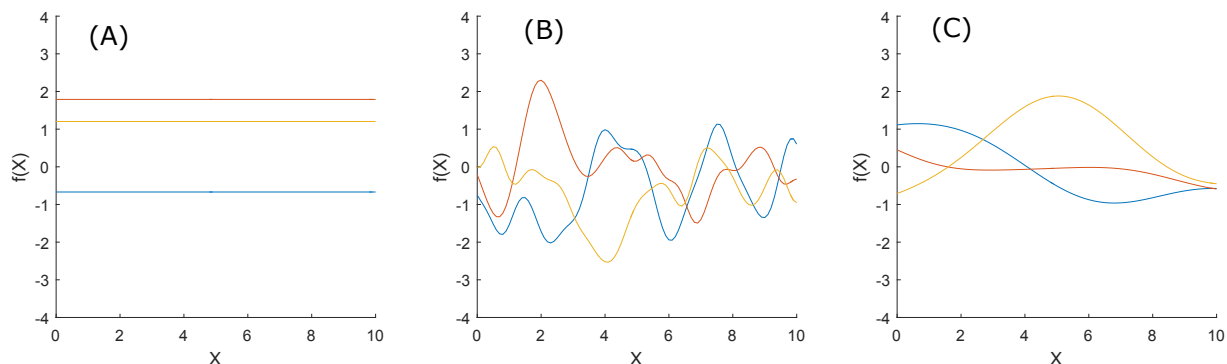
1	Need a smoothie?	2
2	Calculation with squared exponential kernel	3
3	Covariance or not?	3
4	Mixed-up mixture	3
5	More mixture	4
6	Missing data	5
7	Gradient descent for Gaussian mixture	6
8	Principles of principal components	8

1 Need a smoothie?

1. Your friend David just learned a new technique called Gaussian process and he's trying to generate some 1D examples to build some intuition on the different covariance functions for a zero-mean Gaussian process prior. Unfortunately he accidentally spilled the smoothie his girlfriend bought him over his laptop and destroyed his laptop that contains the code he used to generate the plots. Conveniently he has printed out some plots for different covariance functions earlier and roughly remembers what kind of functions he used to generate these plots. As a good friend who excelled in 6.7900, you are trying to comfort him by labeling the plots.

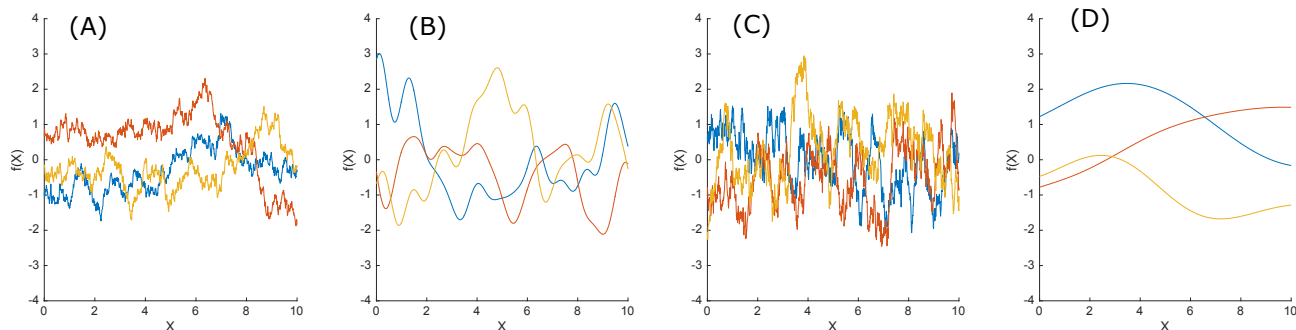
(a) The following plots contain random functions drawn from a covariance function with a squared exponential kernel $k(x, z) = \exp(-\frac{1}{2\tau^2}\|x-z\|^2)$ with different values of τ . Indicate which one of them corresponds to:

- i. $\tau = 0.5$
- ii. $\tau = 3$
- iii. $\tau \rightarrow \infty$



(b) Qualitatively describe what your random function would look like if you draw with a squared exponential kernel when $\tau \rightarrow 0$.

(c) In addition to the squared exponential kernel $k(x, z) = \exp(-\frac{1}{2\tau^2}\|x-z\|^2)$, David has also tried an exponential kernel in the form of $k(x, z) = \exp(-\frac{\|x-z\|}{\tau})$. He has tried two values for τ for both kernels and the values are $\tau = 3$ and $\tau = 0.5$. For each of the plots below, indicate which kernel function it is generated with which τ value.



2 Calculation with squared exponential kernel

Consider a Gaussian process with $\text{Mean}[f(x)] = 0$, and a squared exponential kernel of the form

$$k(x, z) = \sigma_f^2 \exp\left(\frac{-\|x - z\|^2}{2l^2}\right),$$

where l is the characteristic length scale and σ_f is the signal standard deviation. Assume we get observations of y values with noise variance σ_N^2 .

Let $l^2 = 0.5$, $\sigma_f^2 = 5$, $\sigma_N^2 = 0.01$. Assume you have been given observations $((1, 1), (2, -1))$ (these are points in (x, y) space). What is the mean and variance of the prediction at a new query point $x_* = 1.5$? Write down the solution in terms matrices and vectors, you don't need to hand-calculate the final numerical results¹

3 Covariance or not?

Recall that for a Gaussian process model the predictive distribution of the response y_* in a test case with inputs x_* has mean and variance given by

$$\begin{aligned} E[y_* | x_*, \text{training data}] &= k^T C^{-1} y \\ \text{Var}[y_* | x_*, \text{training data}] &= v - k^T C^{-1} k, \end{aligned}$$

where y is the vector of observed responses in training cases, C is the matrix of covariances for the responses in training cases, k is the vector of covariances of the response in the test case with the responses in training cases, and v is the prior (co)variance of the response in the test case.

- Suppose we have just one training case, with $x_1 = 3$ and $y_1 = 4$. Suppose also that the noise-free covariance function is $K(x, x') = 2^{-|x-x'|}$, and the variance of the noise is $1/2$. Find the mean and variance of the predictive distribution for the response in a test case for which the value of the input is 5.
- Repeat the calculations, but using $K(x, x') = 2^{+|x-x'|}$. What can you conclude from the result of this calculation?

4 Mixed-up mixture

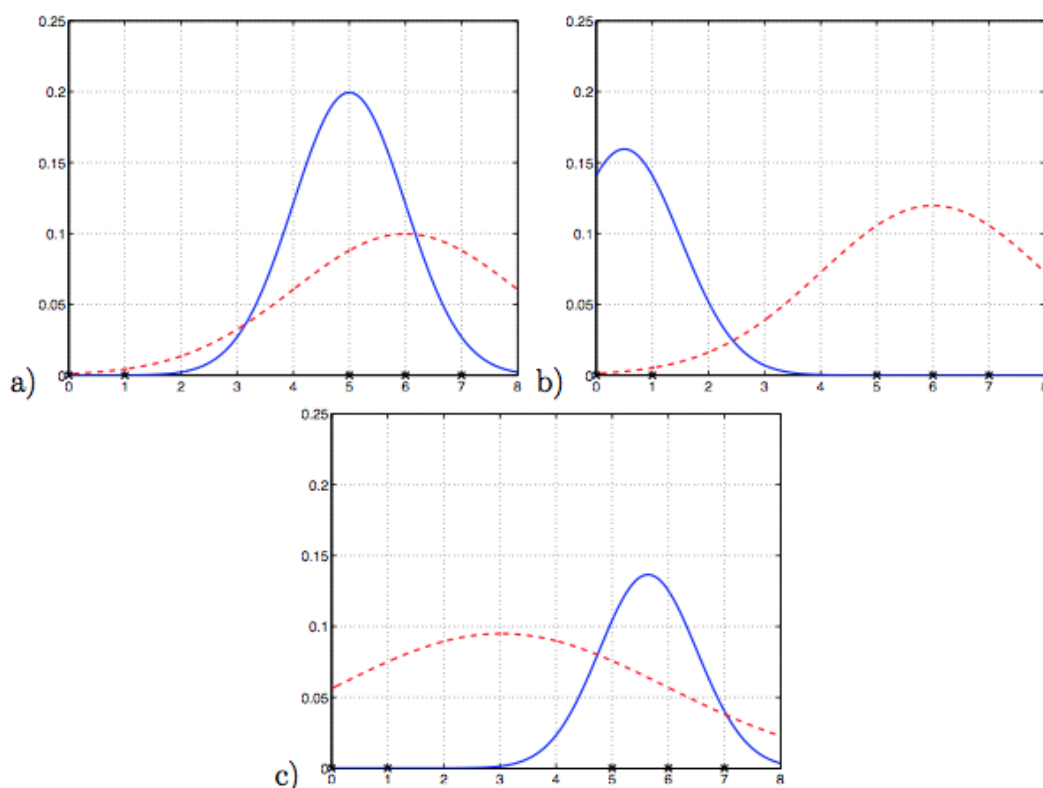
Here we are estimating a mixture of two Gaussians via the EM algorithm. The mixture distribution over x is given by

$$P(x; \theta) = P(1)N(x; \mu_1, \sigma_1^2) + P(2)N(x; \mu_2, \sigma_2^2)$$

¹If you are trying to refer to parameter estimate formula from the textbook "Gaussian Processes for Machine Learning", notice that in eq. (2.21), the predictive distribution is derived with respect to f_* , in order to derive the predictive distribution for y_* , you need to add $\sigma_N^2 I$ to the lower right block of the covariance matrix. This would also impact subsequent equations such as eq. (2.24).

Any student in this class could solve this estimation problem easily. Well, one student, devious as they were, scrambled the order of figures illustrating EM updates. They may have also slipped in a figure that does not belong. Your task is to extract the figures of successive updates and explain why your ordering makes sense from the point of view of how the EM algorithm works. All the figures plot $P(1)N(x; \mu_1, \sigma_1^2)$ as a function of x with a solid line and $P(2)N(x; \mu_2, \sigma_2^2)$ with a dashed line. The sampled data points are given along the x axis in the figures at $x = 0, 1, 2, 5, 6, 7$ and are denoted by the cross marks on the x axis.

- (a) (True/False) In the mixture model, we can identify the most likely T posterior assignment, i.e., j that maximizes $P(j | x)$, by comparing the values of $P(1)N(x; \mu_1, \sigma_1^2)$ and $P(2)N(x; \mu_2, \sigma_2^2)$

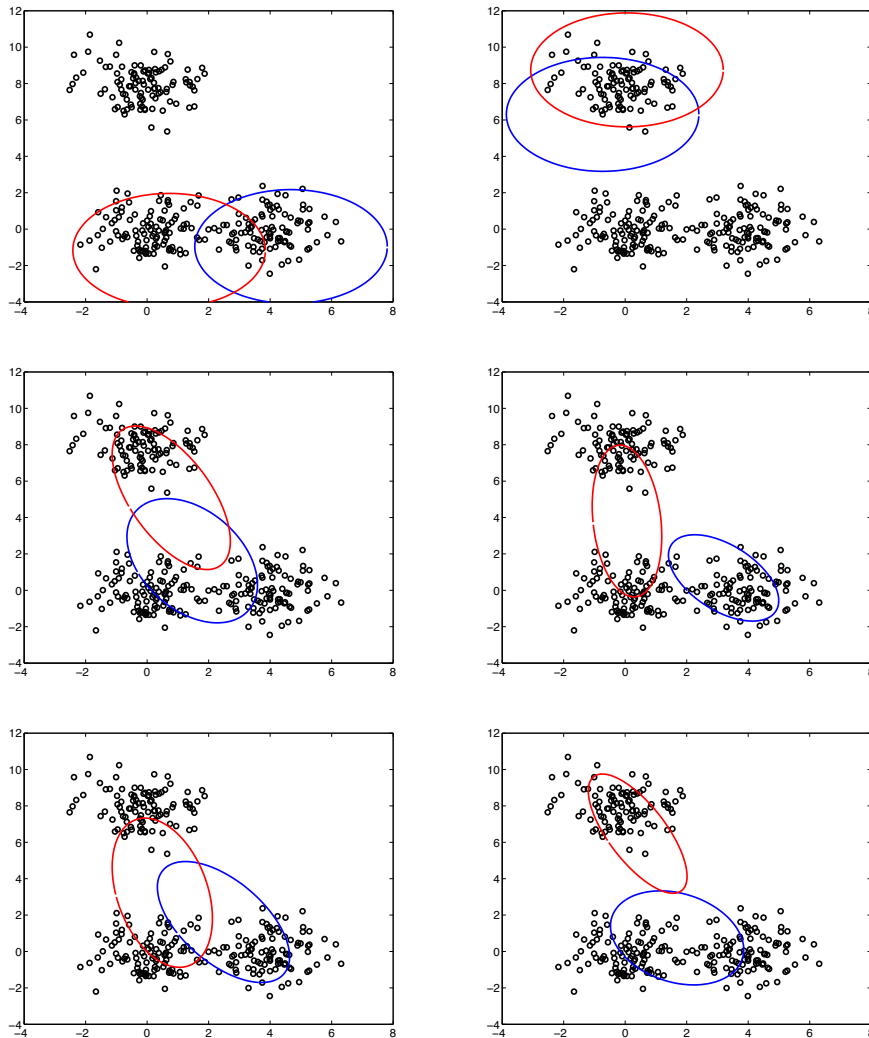


- (b) Assign two figures to the correct steps in the EM algorithm.
- Step 0: () initial mixture distribution
 - Step 1: () after one EM-iteration
- (c) Briefly explain how the mixture you chose for “step 1” follows from the mixture you have in “step 0”.

5 More mixture

We estimated a two Gaussians mixture model based on two-dimensional data shown in the figure below. The mixture was initialized randomly in two different ways and run for three iterations based on each initialization. However, the figures got mixed up (yes, again!). Please

draw an arrow from one figure to another to indicate how they follow from each other (you should draw only four arrows). The ellipses represent the 1 standard deviation equi-probability contours of the Gaussians. The small circles represent the sampled data points.



6 Missing data

We'll start with a very simple problem, in which single attribute of a single data set is missing. There are two attributes, A and B, and this is our data set, \mathcal{D} :

i	A	B
1	1	1
2	1	1
3	0	0
4	0	0
5	0	0
6	0	H ***missing **

7	0	1
8	1	0

Assume the data is *missing completely at random* (MCAR): that is, that the fact that it is missing is independent of its value.

Our goal is to estimate $\Pr(A, B)$ from this data. We'd really like to find the maximum-likelihood parameter values, if we can. The likelihood is

$$\mathcal{L}(\theta) = \log \Pr(\mathcal{D}; \theta) = \log (\Pr(\mathcal{D}, H = 0; \theta) + \Pr(\mathcal{D}, H = 1; \theta)) \quad .$$

- (a) Kim is lazy and decides to ignore $x^{(6)}$ all together, and estimate the parameters:

$$\hat{\theta}^1 = \begin{pmatrix} \Pr(A = 0, B = 0) & \Pr(A = 1, B = 0) \\ \Pr(A = 1, B = 0) & \Pr(A = 1, B = 1) \end{pmatrix} = \begin{pmatrix} 3/7 & 1/7 \\ 1/7 & 2/7 \end{pmatrix} = \begin{pmatrix} .429 & .143 \\ .143 & .285 \end{pmatrix}$$

What is $\mathcal{L}(\hat{\theta}^1)$?

- (b) Jan thinks we should let H be the 'best' value it could have, that is to make the log likelihood as large as possible, and so tries setting $H = 0$ and then $H = 1$ and computes the log likelihood of the complete data in both cases. What value gives the highest complete-data log likelihood? What is the likelihood value?
- (c) Evelyn thinks this is all unprincipled messing around and says we should optimize the thing we want to optimize! That is,

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta) \quad .$$

Evelyn also thinks we can just use the code for gradient descent that we already built in 6.7900 to do this job.

Is Evelyn right?

- (d) Ariel was paying close attention in lecture and thinks this problem is an example of estimation in the presence of a latent variable and that we should use EM.

Let's start with the guess

$$\theta_0 = \begin{pmatrix} .25 & .25 \\ .25 & .25 \end{pmatrix}$$

What is the formula for the E step in this problem? What is the numerical result in this particular case?

- (e) Ariel's roommate Angel joins in the EM game and computes the M step, to get θ_1 . What is the numerical value in this case, and why?
- (f) Will EM always find a solution that maximizes \mathcal{L} ?

7 Gradient descent for Gaussian mixture

- It is typical to fit a Gaussian mixture model to data using EM (expectation maximization) but we can also use gradient descent!

Assume we have a latent discrete variable Z with values in $\{1, \dots, K\}$ and an observable continuous variable X with values in \mathbb{R}^d .

The likelihood of the data, as a function of the parameters, is

$$\mathcal{L}(\pi, \mu, \Sigma) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \log \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k)$$

Assuming we know K , we would like to find π , μ , and Σ to **maximize** this quantity.

- (a) The first problem we face is that our parameters are constrained to be in a limited space: the π have to constitute a probability distribution (be in the range $[0, 1]$) and the σ have to be valid covariance matrices (positive definite). For simplicity, let's assume that

$$\Sigma_j = I\sigma_j^2$$

for $\sigma_j > 0$ (that is, that the covariances are round).

What is a different parameterization for π and the σ_j values so that we can do *unconstrained* gradient descent on them?

- (b) Now, let's look at a very simple version of this problem, with a single data point in 1D, with two components. We get

$$\mathcal{L}(\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1, \sigma_2) = \pi_1 \mathcal{N}(x; \mu_1, \sigma_1) + \pi_2 \mathcal{N}(x; \mu_2, \sigma_2)$$

We find that

$$\frac{\partial \mathcal{L}}{\partial \mu_1} = \pi_1 (x - \mu_1) \frac{\exp(-\frac{(x - \mu_1)^2}{2\sigma_1^2})}{\sqrt{2\pi}\sigma_1^3}$$

and, of course, get a symmetric result for $\partial \mathcal{L} / \partial \mu_2$.

- i. For a training example x , if we do one SGD update, in what directions will μ_1 and μ_2 move?
 - ii. What governs which μ parameter will be changed the most?
- (c) Staying with the simple 1D 2-component version, letting $\pi_1 = \exp(a_1) / (\exp(a_1) + \exp(a_2))$, we find that

$$\frac{\partial \mathcal{L}}{\partial a_1} = \frac{\exp(a_1 + a_2)}{(\exp(a_1) + \exp(a_2))^2} (\mathcal{N}(x; \mu_1, \sigma_1) - \mathcal{N}(x; \mu_2, \sigma_2))$$

and get a symmetric result for $\partial \mathcal{L} / \partial a_2$. Let's assume that given the current parameters, x is much more likely given μ_1, σ_1 than given μ_2, σ_2 . In one SGD update with input x , describe how π_1 and π_2 would be changed.

- (d) Finally, in this same problem, but letting $\sigma_1 = \exp(b_1)$, we find that

$$\frac{\partial \mathcal{L}}{\partial b_1} = \pi_1 \mathcal{N}(x; \mu_1, \exp(b_1)) \left(\frac{(\mu_1 - x)^2}{\exp(2b_1)} - 1 \right)$$

and get a symmetric result for $\partial \mathcal{L} / \partial b_2$. In one SGD update with input x , describe how σ_1 and σ_2 would be changed.

3. 8 Principles of principal components

4. Principal components are related to several other ideas we have come across in class so far. We'll explore this in two dimensions. Imagine we have a data set $\mathcal{D} = \{(x_1^{(i)}, x_2^{(i)})\}_{i=1}^n$.

- (a) In homework 0, we observed that the eigenvectors of the covariance matrix of a multi-dimensional Gaussian corresponded to axes of ellipses describing equi-probability contours.

Show that the eigenvector of the covariance matrix with the largest corresponding eigenvalue is equivalent to the first “principal” component of the data.

- (b) One way to describe the first principal component is that it is the line such that the sum of the perpendicular distances of the points to the line is minimized. This sounds sort of like what's happening in ordinary least squares. Explain why they are different and draw a picture of a small data-set (4 points) in which the solutions are substantially different.