COMS 4771 HW 3 (Spring 2020)

Due: Apr 10, 2020 at 11:59pm

This homework is to be done **alone**. No late homeworks are allowed. To receive credit, a typesetted copy of the homework pdf must be uploaded to Gradescope by the due date. You must show your work to receive full credit. Discussing possible solutions for homework questions is encouraged on piazza and with your peers, but you must write their own individual solutions. You should cite all resources (including online material, books, articles, help taken from specific individuals, etc.) you used to complete your work.

1 Non-parametric Regression via Bayesian Modelling

Here we will study a generative modelling technique via Gaussians for non-parametric regression.

Before getting into regression we need to derive some facts about multivariate Gaussian distributions. Let $x \in \mathbb{R}^d$ be distributed normally as

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) = \mathcal{N}(\mu, \Sigma)$$

(i) Derive the marginal distribution of x_1?

(ii) (*warning! tedious calculations*) Let $\Sigma^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix}$. Using the facts that[1]

- $\Sigma^{11} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathsf{T}})^{-1} = \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}$
- $\Sigma^{22} = (\Sigma_{22} - \Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}\Sigma_{12})^{-1} = \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^{\mathsf{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathsf{T}})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$
- $\Sigma^{12} = -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}\Sigma_{12})^{-1} = (\Sigma^{21})^{\mathsf{T}}$

Show that the joint distribution on x can be written as

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}\Big((x_1 - \mu_1)^{\mathsf{T}}\Sigma_{11}^{-1}(x_1 - \mu_1) + (x_2 - b)^{\mathsf{T}}A^{-1}(x_2 - b) \Big) \Big\},$$

where $b = \mu_2 + \Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}(x_1 - \mu_1)$, and $A = \Sigma_{22} - \Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}\Sigma_{12}$.

(iii) Now using the fact that $|\Sigma| = |\Sigma_{11}||\Sigma_{22} - \Sigma_{12}^{\mathsf{T}}\Sigma_{11}^{-1}\Sigma_{12}|$, show that the joint on x can be decomposed as product

$$\mathcal{N}(x_1; \mu_1, \Sigma_{11})\mathcal{N}(x_2; b, A),$$

where b and A are as defined in previous part.

[1] Feel free to prove these facts for yourself, if you are bored :).

(iv) What is the conditional distribution of x_2 given x_1?

Now we are ready to do some regression via generative modelling. Like in any generative model, we first need a prior over our objects of interest (in this case, the regression functions), then given some data (also known as evidence), we shall compute the posterior (in this case, those regression functions that agree with the given data).

A prior over the regression functions. A function $f : \mathbb{R} \to \mathbb{R}$ can be thought of as an infinite length vector. Suppose we want to know the value of this function at positions $x_1, \ldots, x_n \in \mathbb{R}$, we can write it down the result as a vector $\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$. A simple way to *generate* random functions then is to simply model it as the Gaussian distribution, specifically $\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}(\mu_n, \Sigma_{n\times n})$.

Throughout our discussion below, we will use 500 equal spaced points between the range of -10 and 10 as our locations $x_1, \ldots, x_n$, that is $x_1 = -10, x_2 = -9.959, \ldots, x_{500} = 10$.

(v) For $\mu_n = \vec{0}$ and $\Sigma_{n\times n} = I$, draw 4 random functions and show their plots[2]. What can you say about the smoothness of these functions? What happens if Σ is set to all ones matrix? Play with various values of μ and Σ, what effect does it have on the distribution of the random functions? Explain why these effects are occurring.

(vi) Usually $\mu_n = \vec{0}$ and $\Sigma_{n\times n} = K$, where $K_{ij} = k(x_i, x_j)$ for some kernel function k. A popular choice is $k : (x_i, x_j) \mapsto \exp\{-(x_i - x_j)^2/h\}$, for some fixed parameter h. Draw 4 random functions from this setting of μ and Σ ($h = 5$). and show their plots. What can you say about the smoothness of these functions?

(vii) If one is interested in random periodic functions, qualitatively explain what setting of μ and Σ would be appropriate? Pick a μ and Σ which can generate periodic functions of periodicity 3 units. Draw 4 random functions from that setting and plot them to verify.

The posterior over the regression functions. Of course, in the problem of regression, one is not interested in drawing random functions, but instead, understanding/predicting the trend in data given some observations. Suppose we are given a training data $(\bar{x}_1, \bar{y}_1), \ldots, (\bar{x}_m, \bar{y}_m) = (\overline{\mathbf{X}}, \overline{\mathbf{Y}})$. Then using the suggested model we can model the joint distribution as

[2] use x-axis range -10 to 10, y-axis range -3 to 3 for all your plots.

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \\ f(\bar{x}_1) = \bar{y}_i \\ \vdots \\ f(\bar{x}_m) = \bar{y}_m \end{bmatrix} = \begin{bmatrix} f(\mathbf{X}) \\ f(\overline{\mathbf{X}}) = \overline{\mathbf{Y}} \end{bmatrix} \sim \mathcal{N}\big(\mu_{n+m}, \Sigma_{(n+m)\times(n+m)}\big) = \mathcal{N}\left(\begin{bmatrix} \mu_n \\ \mu_m \end{bmatrix}, \begin{bmatrix} \Sigma_{nn} & \Sigma_{nm} \\ \Sigma_{mn} & \Sigma_{mm} \end{bmatrix} \right)$$

$$= \mathcal{N}\left(\begin{bmatrix} \mu_n \\ \mu_m \end{bmatrix}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) & K(\mathbf{X},\overline{\mathbf{X}}) \\ K(\overline{\mathbf{X}},\mathbf{X}) & K(\overline{\mathbf{X}},\overline{\mathbf{X}}) \end{bmatrix} \right).$$

Let $y_1, \ldots, y_n = \mathbf{Y}$ be the regression values we are interested in knowing for a set of (test) locations $x_1, \ldots, x_n = \mathbf{X}$, then we can write the above joint model more compactly as

$$\begin{bmatrix} \mathbf{Y} \\ \overline{\mathbf{Y}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_n \\ \mu_m \end{bmatrix}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) & K(\mathbf{X},\overline{\mathbf{X}}) \\ K(\overline{\mathbf{X}},\mathbf{X}) & K(\overline{\mathbf{X}},\overline{\mathbf{X}}) \end{bmatrix} \right).$$

Hence, given the training data $(\overline{\mathbf{X}}, \overline{\mathbf{Y}})$ we are interested in knowing the posterior $\mathbf{Y}\big|\overline{\mathbf{Y}}$

(viii) Using the result from part (iv), what is the posterior $\mathbf{Y}\big|\overline{\mathbf{Y}}$?

(ix) For training data $\{(-6,3),(0,-2),(7,2)\}$ and K induced by kernel function $k : (x_i, x_j) \mapsto \exp\{-(x_i - x_j)^2/5\}$, draw 4 random functions from the posterior and plot the resulting functions. Make sure to depict the three training datapoints on the same plot. What do you notice?

(x) For the training data in part (ix) and the periodic Σ used in part (vii), draw 4 random functions from the posterior and plot the resulting functions (along with the training data). What do you notice in this case?

Notice that this Bayesian modelling technique provides a (posterior) *distribution* over regression values for the test locations. One can use the mean value as the final prediction over the test locations.

(xi) What is the mean of the posterior $\mathbf{Y}\big|\overline{\mathbf{Y}}$?

(xii) Plot the mean "function" for parts (ix) and (x) as well.