Assignment III (2015.2 - T01)

Submission Deadline: November, 3rd, 2015.

Instructions:

- Submission deadline is November, 3rd, 2015.
- This assignment must be delivered as a report (with Introduction, Methodology, Results, Conclusion and References).
- Source codes must be delivered as attachments.
- When answering the questions, books and papers citations are allowed as long as they are listed in "references" section of the report.
- When submitting this assignment, please inform if you are registered or not in SIGAA.
- If you are registered in SIGAA, provide your university ID number when submitting the assignment.

Exercise 01) This exercises explores the *Conditional Gaussians*. With a simulator or a programming language, allocate a mean vector $\boldsymbol{\mu} \in \mathbb{R}^3$ in a way that

$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$$

and initialize a matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ in a way that

$$\mathbf{K} = \begin{pmatrix} 0.5 & 0.2 & 0.1 \\ 0.6 & 0.18 & 0.4 \\ 0.1 & 0.6 & 0.2 \end{pmatrix}.$$

Now, taking $\mathbf{K}$, generate 3 covariance matrices. Each one follow the following procedures.

I) *Isotropic Covarince Matrix* given by

$$\boldsymbol{\Sigma}_I = k\mathbf{I},$$

where $\mathbf{I}$ is an identity matrix and k is a strictly positive number. For this exercise, define $k = 4$.

II) *Full Covariance Matrix* given by

$$\boldsymbol{\Sigma}_{II} = \mathbf{K}\mathbf{K}^T, \tag{1}$$

III) *Diagonal Covariance Matrix* is a matrix filled with zeros, but the diagonal is the same from the diagonal of the full covariance matrix, so

$$\boldsymbol{\Sigma}_{III} = \operatorname{diag}\left(\mathbf{K}\mathbf{K}^T\right).$$

After defining $\boldsymbol{\Sigma}$, partition it in a way that:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix},$$

where, in this example, $\boldsymbol{\Sigma}_{aa} \in \mathbb{R}^{2\times2}$, $\boldsymbol{\Sigma}_{ab} \in \mathbb{R}^{2\times1}$, $\boldsymbol{\Sigma}_{ba} \in \mathbb{R}^{1\times2}$ and $\boldsymbol{\Sigma}_{bb} \in \mathbb{R}^{1\times1}$. Notice also that $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^T$.

Partition $\boldsymbol{\mu}$ into two vectors $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ in way that

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}.$$

Now, after executing these instructions, using the generated means vector and each of the 3 generated covariance matrices, answer the following questions (for each covariance matrix: $\boldsymbol{\Sigma}_I$, $\boldsymbol{\Sigma}_{II}$ and $\boldsymbol{\Sigma}_{III}$ has an answer. So, each of the following questions has 3 answers. One answer for each generated covariance matrix).

1.1) What is the final value of $\boldsymbol{\Sigma}$?

1.2) Inform the values of: $\boldsymbol{\Sigma}_{aa}$, $\boldsymbol{\Sigma}_{ab}$, $\boldsymbol{\Sigma}_{ba}$ and $\boldsymbol{\Sigma}_{bb}$?

1.3) Assume $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. Inform the values of $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}_{aa}$, $\boldsymbol{\Lambda}_{ab}$, $\boldsymbol{\Lambda}_{ba}$ and $\boldsymbol{\Lambda}_{bb}$.

1.4) What is the value of $\boldsymbol{\Sigma}_{a|b}$?

1.5) Inform the values of: $\boldsymbol{\mu}$, $\boldsymbol{\mu}_a$, $\boldsymbol{\mu}_b$ and $\boldsymbol{\mu}_{a|b}$?

1.6) Using

$$\mathbf{x}_a = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix} \text{ and}$$

$$\mathbf{x}_b = (1),$$

compute the conditional gaussian $p(\mathbf{x}_a|\mathbf{x}_b)$. Remember to compute it for each covariance matrix.

Exercise 02) Using the same procedure and data from **Exercise 01**, calculate the marginal gaussian:

$$p(\mathbf{x}_b) = \int p(\mathbf{x}_b, \mathbf{x}_a) \mathrm{d}\mathbf{x}_a$$

and inform the values of $\mathbb{E}[\mathbf{x}_b]$ and $\mathrm{cov}[\mathbf{x}_b]$.

Exercise 03) Using Student's t-distribution

$$\mathrm{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2},$$

plot $\mathrm{St}(x|\mu, \lambda, \nu)$ for x inside the range $[-10; 10]$, $\mu = 1$ and $\lambda = 2$ for ν equals:

3.1) $\nu = 0.01$
3.2) $\nu = 0.1$
3.3) $\nu = 1$
3.4) $\nu = 10$
3.5) $\nu = 10^2$.
3.6) What happens if $\nu \to +\infty$? Why?

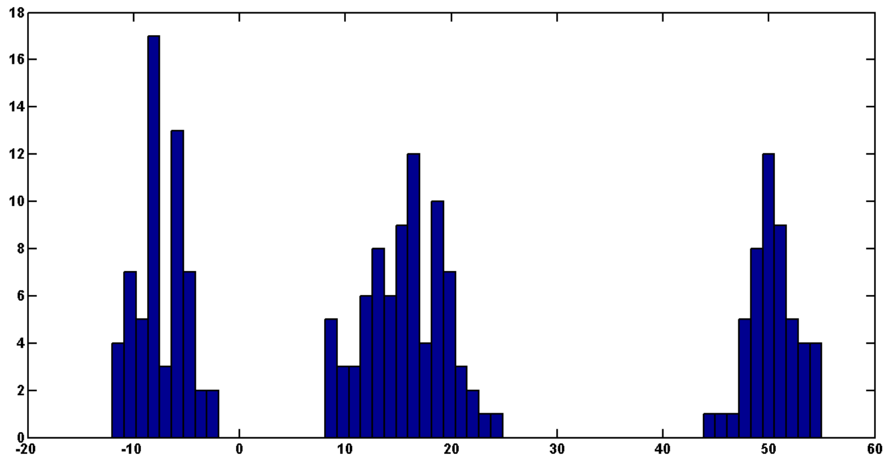

Figure 1: Histogram of the data available in data-B.txt

Exercise 04) A dataset called `data-B.txt`, available in the website of this course has $N = 190$ samples and it has only one dimension. Figure 1 show its histogram. Using *mixture of gaussians*,

$$p(x) = \sum_{1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k),$$

and setting $\sigma_k = 1$ try to fit its histogram using K number of gaussians through trial and error/visual inspection. After that, inform the number of K gaussians used for fitting as well as the means each gaussian. Inform the number of gaussians used (K). After that, determine the value of each π_k.
HINT: Take a look at the peaks of the histogram.

Exercise 05) Using the same dataset used in **Exercise 04** and knowing that `data-B.txt` has only one dimension ($D = 1$), estimate the probability density using *Kernel Density Estimators* (KDE) approach. Firstly, use the *Parzen Window* kernel

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right),$$

and $k(\cdot)$ is a function where

$$k(\mathbf{x}) = \begin{cases} 1, \text{ if } \ |x_i| \leq \frac{1}{2} & \text{for } i = 1, \ldots, D. \\ 0, \text{otherwise} \end{cases}.$$

Then, use the *Gaussian kernel*:

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h\sqrt{2\pi}}\exp\left\{-\frac{||\mathbf{x}-\mathbf{x}_n||^2}{2h^2}\right\}.$$

Test these two kernels and plot the results with the following values of h:

5.1) $h = 0.1$;
5.2) $h = 0.5$;
5.3) $h = 1$;
5.4) $h = 2$.

Exercise 06) This exercise will explore the *Maximum likelihood/Least Squares* for *regression problem*. It will make use of 3 pairs of datasets. Each pair corresponds to the input and output data of a model

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

where $t \in \mathbb{R}$ is the output, ϵ is a noise and $\mathbf{x} \in \mathbb{R}^2$ is the input. The main goal of this problem is to find the best $\mathbf{w} \in \mathbb{R}^M$ that

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$$

is minimal. Recall that $\phi(.)$ is a basis function. Use the 3 pairs of datasets (details about the datasets will be described shortly) to find an optimal value for $\mathbf{w}$. Plot the original output and the estimated output. Calculate the residual sum-of-squares error function at the end of testing phase. Execute Least-Squares regression with the following basis functions:

6.1) $\phi(x)_j = x$;

6.2) $\phi(x)_j = \exp\left\{-\frac{x-\mu_j}{4}\right\}$. For each center basis, use a different value of μ_j.

Execute the algorithm using the following number of basis functions: 2, 4, 8, 16, 32, 64, 128 and 256. Plot the mean squared error (MSE) for each number of basis (plot: MSE $\times$ number of basis) .
Find the optimal number of basis functions (based in MSE).
Recall, that for finding the maximum likelihood $\mathbf{w}$

$$\mathbf{w}_{\text{ML}} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{t},$$

where $\mathbf{t} = (t_1, t_2, \ldots, t_N)^T$ and

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \ldots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \ldots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \ldots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

The first dataset pair, `data-C-learning-input.txt` and `data-C-learning-output.txt` stands for $\mathbf{x}$ and t respectively. Both have $N = 2000$ samples and they must be used in the training step.
The second dataset pair, `data-C-validation-input.txt` and `data-C-validation-output.txt` stands for $\mathbf{x}$ and t respectively. Both have $N = 1000$ samples and they must be used in the validation step. This step is used to determine each number of basis functions is optimal.
The last dataset pair, `data-C-test-input.txt` and `data-C-test-output.txt` stands for $\mathbf{x}$ and t respectively. Both have $N = 1000$ samples and they must be used to test the results. This dataset is supposed to be used after the validation step (after discovering the optimal number of basis functions).

All results (such as sum-of-squares error, and plots) must be generated through this last dataset pair.
HINT: Code everything as a function putting everything inside the function.

Exercise 07) Taking the optimal number of basis functions found in **Exercise 06**, use the Regularized least squares:

$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t},$$

with the following values of λ:

7.1) $\lambda = 0.01$;
7.2) $\lambda = 0.1$;
7.3) $\lambda = 1$;
7.4) $\lambda = 10$;
7.5) $\lambda = 100$.
7.6) Find the optimal value of λ. Compare the regularized least squares with optimal λ with the results of the previous exercise. Use the mean-squared error as a criterion for comparing the results.

HINT: use the function implemented in the previous exercise.

Exercise 08) This exercise simulates the experiment available in section 3.3 of Bishop's book, It explores Bayesian Linear Regression and uses two datasets. The datasets called `data-D-input.txt` and `data-D-output.txt` corresponds to the input and the output of the model:

$$y(x, \mathbf{w}) = (w_0 + w_1 x) + \epsilon.$$

Assuming that the signal is corrupted by a additive gaussian noise (ϵ) with zero-mean and standard deviation 0.2. For sake of simplicity, assume $\phi(\mathbf{x}) = \mathbf{x}$. The noise variance is known. Hence, set the precision parameter $\beta = (1/0.2)^2 = 25$. Also, consider a zero-mean isotropic gaussian governed by a single precision parameter $\alpha = 2$ so that the prior is

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}, \alpha^{-1}\mathbf{I})$$

and the corresponding posterior distribution over $\mathbf{w}$ is given by

$$p(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \text{ with}$$

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^T \mathbf{t} \text{ and} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \mathbf{\Phi}^T \mathbf{\Phi}. \end{aligned}$$

Use Bayesian Linear Regression to determine the values of $\mathbf{w} = (w_0 \ , \ w_1)^T$.

Good luck!