SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Homework #2

Course: *Machine Learning (CS405)* – Professor: *Qi Hao*
Due date: *11:59pm, October 7th, 2020*

Question 1

(*a*) **[True or False]** If two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian.

(*b*) We consider a partitioning of the components of x into three groups x_a, x_b, and x_c, with a corresponding partitioning of the mean vector μ and of the covariance matrix $\sum$ in the form

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \\ \mu_c \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{pmatrix}.$$

Find an expression for the conditional distribution $p(x_a|x_b)$ in which x_c has been marginalized out.

Question 2

Consider a joint distribution over the variable

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

whose mean and covariance are given by

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \mu \\ \mathbf{A}\mu + \mathbf{b} \end{pmatrix}, \quad \text{cov}[\mathbf{z}] = \begin{pmatrix} \mathbf{\Lambda^{-1}} & \mathbf{\Lambda^{-1}A^{\mathrm{T}}} \\ \mathbf{A\Lambda^{-1}} & \mathbf{L^{-1}} + \mathbf{A\Lambda^{-1}A^{\mathrm{T}}} \end{pmatrix}.$$

(*a*) Show that the marginal distribution $p(\mathbf{x})$ is given by $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Lambda^{-1}})$.

(*b*) Show that the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is given by $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L^{-1}})$.

Question 3

Show that the covariance matrix Σ that maximizes the log likelihood function is given by the sample covariance

$$\Sigma_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{ML}})(\mathbf{x}_n - \mu_{\mathrm{ML}})^{\mathrm{T}}.$$

Is the final result symmetric and positive definite (provided the sample covariance is nonsingular)?

Hints.

(*a*) To find the maximum likelihood solution for the covariance matrix of a multivariate Gaussian, we need to maximize the log likelihood function with respect to Σ. The log likelihood function is given by

$$\ln p(\mathbf{X}|\mu,\Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_n - \mu).$$

(*b*) The derivative of the inverse of a matrix can be expressed as

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}$$

We have the following properties

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{Tr}(\mathbf{A}) = \mathbf{I}, \quad \frac{\partial}{\partial \mathbf{A}}\ln|\mathbf{A}| = (\mathbf{A}^{-\mathbf{1}})^{\mathrm{T}}.$$

Question 4

(*a*) Derive an expression for the sequential estimation of the variance of a univariate Gaussian distribution, by starting with the maximum likelihood expression

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2.$$

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

(*b*) Derive an expression for the sequential estimation of the covariance of a multivariate Gaussian distribution, by starting with the maximum likelihood expression

$$\Sigma_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{\mathrm{ML}})(\mathbf{x}_n - \mu_{\mathrm{ML}})^{\mathrm{T}}.$$

Verify that substituting the expression for a Gaussian distribution into the Robbins-Monro sequential estimation formula gives a result of the same form, and hence obtain an expression for the corresponding coefficients a_N.

Hints.

(*a*) Consider the result $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ for the maximum likelihood estimator of the mean μ_{ML}, which we will denote by $\mu_{\mathrm{ML}}^{(N)}$ when it is based on N observations. If we dissect out the contribution from the final data point $\mathbf{x}_N$, we obtain

$$\mu_{\mathrm{ML}}^{(N)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n = \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\mu_{\mathrm{ML}}^{(N-1)}$$

$$= \mu_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \mu_{\mathrm{ML}}^{(N-1)}).$$

(*b*) Robbins-Monro for maximum likelihood

$$\theta^{(N)} = \theta^{(N-1)} + a_{(N-1)}\frac{\partial}{\partial\theta^{(N-1)}}\ln p(x_N|\theta^{(N-1)}).$$

Question 5

Consider a D-dimensional Gaussian random variable $\mathbf{x}$ with distribution $N(x|\mu,\Sigma)$ in which th4e covariance Σ is known and for which we wish to infer the mean μ from a set of observations $\mathbf{X} = \{x_1, x_2, ..., x_N\}$. Given a prior distribution $p(\mu) = N(\mu|\mu_0,\Sigma_0)$, find the corresponding posterior distribution $p(\mu|\mathbf{X})$.

Program Question

You should download the `HW2_programQuestion.ipynb` **file first.**
In this coding exercise, we will implement the K-nearest Neighbors (KNN) algorithm. You are provided with a Jupyter Notebook in which you will have to fill in the functions as instructed therein. Be sure to read the notebook thoroughly for the instructions and also comment your code appropriately.

This is a classification problem and we will use the Breast Cancer dataset. This dataset is included as part of sklearn. We have loaded the dataset for you in the Jupyter notebook. Please familiarize yourself with the dataset first.

K	Norm	Accuracy (%)
3	L1	—
3	L2	—
3	L-inf	—
5	L1	—
5	L2	—
5	L-inf	—
7	L1	—
7	L2	—
7	L-inf	—

Table1: Accuracy for the KNN classification problem on the validation set

A training data (X train) is provided which has several datapoints, and each datapoint is a p-dimensional vector (i.e., p features). You are also provided with separate validation (X val) and test sets (X test). Your task is to implement the K-nearest neighbors algorithm to determine the ideal combination of the value of K and the metric norm. For this you have to play with different combinations of metric norm and different values of K.

Compute the accuracy for the X val set for every combination of K and metric norm. Once, you have decided the ideal value of K and a relevant metric norm using the validation set, use those values to report the accuracy for the test set X test. You have to use the following values of K = 3, 5 and 7. The different metrics norms to be implemented are: L1, L2 and L-inf. Do not use any library to implement the norms.

(*a*) How could having a larger dataset influence the performance of KNN?

(*b*) Tabulate your results in Table 1 for the **validation set** as shown below and include that in your file.

(*c*) Finally, mention the best K and the norm combination you have settled upon from the above table and report the accuracy on the test set using that combination.

(*d*) The Autograder for your code submission will grade on the correctness of your implementation for the following functions as given in the Jupyter notebook: **distanceFunc, computeDistancesNeighbors, Majority** and **KNN**.

Reference. The dataset and question are from Kaggle and University of Pennsylvania.