King Mongkut's University of Technology **Name:** ______________________

Machine Learning **I.D. Number:** ________________

Suthep Madarasmi, Ph.D.

Take Home Quiz 2 due Sat Feb 11, 2023 (New) **Score:** ______ **/ 65**

1. 1.5 hr. Min Distance classifier on 3 Gaussian Classes. Modify your KNN program from quiz 1 with 3 classes to create 50 random points for each class: red, blue, and yellow from a *2D Gaussian distribution* (see Gaussian Data.ipynb) with means: (20, 30), (40, 40), (50, 40) and (s_x, s_y) of (3, 10), (10, 10), (15, 15), for red, blue, and yellow, respectively. Use 70% of the dataset as training data and 30% as testing data.
    - *5 pts.* Plot the 3 classes using the training data.
    - *5 pts.* Using KNN with K = 5, report the total accuracy of the testing data.
    - Using the minimum distance classifier:
        - *5 pts.* Report the training data cluster mean for each class of red, blue, and yellow.
        - *5 pts.* Report the total accuracy of the testing data using this classifier.

2. 4 hrs. **Code Naive Bayes from Scratch**. Write a program to read the Iris dataset, split into 2 parts: training and testing just like it was done in the example. Then write your own code to:

a. *5 pts.* Find the mean and standard deviation for each of the 4 features of each of the 3 classes from the training data. μ_ik and σ_ik for i = 1..4, k = 1..3. You will get pdfs P(x_i | c_k) for each class using the Gaussian distribution equation with μ_ik and σ_ik for i = 1..4, k = 1..3. This gets you pdfs: P(x_1, x_2, x_3, x_4 | c_1), P(x_1, x_2, x_3, x_4 | c_2), P(x_1, x_2, x_3, x_4 | c_3).

b. *5 pts.* Find the P(c_k) by counting the percent frequency of each class in your training data. Now we have P(c_k | x_1, x_2, x_3, x_4) ∝ P(x_1, x_2, x_3, x_4 | c_k) * P(c_k).

c. *5 pts.* Then for each (x_1, x_2, x_3, x_4) in your test data: find the class k of 1, 2, or 3 for which P(c_k | x_1, x_2, x_3, x_4) is maximum, put that k into array my_predicted_labels

d. *5 pts.* Calculate and print the accuracy score from your implementation of Naive Bayes from scratch

e. *5 pts.* Use sklearn's GaussianNB classifier to report the accuracy score. Compare your result to sklearn's.

3. 1 hr. Try the *digits* datasets. Change the "Naive Bayes and KNN Iris and Cancer.ipynb" program to allow the user to also select the digits dataset by entering "digits", in addition to "iris" and "cancer". Present the output results for both Naive Bayes and KNN classifiers using Sklearn. What is the best value of K in KNN?

4. *10 pts.* 1.5 hr. Normalize data option. Add an option to "Naive Bayes and KNN Iris and Cancer.ipynb" to ask the user whether to normalize the dataset by converting each feature into a Z-distribution by making mean = 0, and standard deviation = 1. For this problem, compare the accuracy results for the breast cancer dataset on the sklearn's KNN classifier using normalized vs. unnormalized data.

Submission. All your work should be put into 1 pdf file and uploaded to KMUTT LMS before the due date.

Late Quiz Submission Policy. 7 days late allowed, after that 2 point deducted per day late.

Honor Code Agreement. Work copied from others including the internet is cheating, resulting in an F grade.