Homework 2

(Total 200 pts)

Due 5:00 pm on September 23, 2022 (Friday)

Canvas submission of your answers (with required plots and source codes attached) in a single PDF file ('hw2.pdf'), and then submit the following source code files:
Q3.m, Q4.py, Q5.m, and Q6.m

1. (20 pts) What is the value of p that maximizes the following log likelihood function $L(p)$? Express your answer in terms of N_1 and N. N is the total number of data samples, out of which there are N_1 samples that belong to class A with class label index being $t_n = 1$. The remaining samples belong to the class B with class label index $t_n = 0$. Show your derivations. Note: you need to determine both the first-order and second-order derivatives to ensure the maximal value is achieved.

   $L(p) = \sum_{n=1}^{N}[t_n \ln p + (1 - t_n)\ln(1 - p)]$, where t_n is the class label index (which is either 0 or 1).

2. (30 pts) Suppose a scalar $f = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{x}$ is a 3×1 column vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, and $\mathbf{A}$ is a 3×3 matrix $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$. Show the detailed derivations to represent the gradient vector $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix}$ in terms of $\mathbf{A}$ and $\mathbf{x}$. Simplify your final expression as much as possible.

3. (40 pts) Performance evaluation metrics in Matlab.
   Binary classification using the simulated dataset:
   http://www.ece.uah.edu/~dwpan/course/ee610/hw/dataset_hw.csv
   (a) Train a Naïve Bayes classifier using the *fitcnb* function in Matlab, and display the confusion matrix using the *resubPredict* and *confusionchart* functions.
   (b) Assume that the class with label value being 1 is the "Positive" class, the other class is the "Negative" class, Fill the table below with the values (rounding to the 4th decimal place) for various metrics.

Accuracy	Sensitivity (Recall)	Specificity	Precision	F1 score

   (c) Attach the confusion matrix chart and the Matlab script.

4. (40 pts) Performance evaluation metrics in sklearn.
   (a) Use the numpy.loadtxt function to load in the file 'dataset_hw.csv' used in Q1.
   (b) Use *GaussianNB* in sklearn to train a Naïve Bayes classifier.

(c) Display the confusion matrix.
(d) Display the classification report.
(e) Compare the results with the confusion matrix and metrics you obtained in Q3.
(e) Attach the confusion matrix plot and the Python code.

5. (40 pts) Non-parametric estimation of probability density estimation.
Read into Matlab the following dataset and estimate the PDF of the x_1 feature (corresponding to the observations in the first column).
http://www.ece.uah.edu/~dwpan/course/ee610/hw/dataset_hw.csv
   (a) Use the *histogram* function (with 'normalization' set to 'pdf') to display the estimated PDF for x_1 for three different bin numbers: 10, 20, and 50. Attach the histograms.
   (b) Use the knnsearch function to estimate the PDF for x_1 using the K nearest-neighbor method. Display the estimated PDF's for three different K values: 10, 20, and 50.
   (c) Attach the Matlab scripts and all the plots.

6. (30 pts) K nearest-neighbor classifiers without and with cross-validations.
   (a) In Matlab, load the dataset:
   >> load fisheriris.
   (b) Train a KNN classifier (using Euclidean distance, and with the 'standardize' option turned on) using K = 2.
   (c) What is the resubstitution loss?
   (d) Using 5-fold cross-validations this time. You can use the *crossval* function to obtain cross-validated classifier models (with a default dataset partitioning being stratified) from the trained model obtained in (b). Initialize the random number generator seed prior to dataset partitioning, by using
   >> rng (1),
   to make sure the results are reproducible.
   (e) Use the *kfoldLoss* function to show the individual losses (by setting the 'Mode' to be 'individual'). Then calculate the average loss for these 5-fold validations.
   (f) Fill in the table below with your answers.

(c) What is the resubstitution loss?	

Five individual classification losses in (e)					Average Loss

   (g) Attach a screenshot running your script, and attach your script.