685.621 Algorithms for Data Science
Homework 3
Assigned at the start of Module 6
Due at the end of Module 8
Total Points 100/100

You are required to participate in the collaborative problem and subproblem separately. Please do not directly post a complete solution, the goal is for the group to develop a solution after everyone has participated. Please ensure you have a write-up with solutions to each problem and subproblems, you are also required to submit code that will be compiled when grading the assignment. In each of the problems you are allowed to use built-in functions.

1. **Problem 1 - Module 4 and 5** ***This is a Collaborative Problem***
   10 Points Total

   In this problem use the developed numerical features from HW2. In this problem the following is to be completed:

2. Use the Fisher's Linear Discriminant Ratio (FDR) from the Data Processing document, specifically Equation 20.

   (a) (5 points) For each combination of number apply the FDR, e.g., 0 vs 1, 0 vs 2, ..., 0 vs 9, ..., 8 vs 6, 8 vs 7, and 8 vs 9.

   (b) (5 points) Place the results in a table.

3. **Problem 2 - Module 6** ***This is a Collaborative Problem***
   35 Points Total

   In this problem, you will be developing pseudocode and implementing your development in Python or R for the Expectation Maximization method. You are allowed to use either the Iris data set or the developed numerical features from HW2. In this problem the following is to be completed:

   (a) (20 points) The development and implementation of the Expectation Maximization method should be for a generic number of clusters, features and observations.

   (b) Apply your implementation using either the Iris data or the features generated from HW 2.

      i. Use the top two ranked features.

      ii. (5 points) Create 3 clusters for the Iris data or 4 clusters using the 4 numerical values that have the best separation.

      iii. (5 points) Display the 3 species or 4 numerical values using different colors for a good visual representation.

      iv. (5 points) Provide an analysis of your results, e.g., what is your observation of the results, how well did the clusters group each class, etc.

4. **Problem 3 - Module 7 - Chapter 1 and 2 [8]** ***Note this is not Collaborative Problem***
5 Points Total

Define in your own words the following terms:

(a) agent
(b) agent function
(c) agent program
(d) artificial intelligence
(e) autonomy
(f) goal-based agent
(g) intelligence
(h) learning agent
(i) logical reasoning
(j) model-based agent
(k) rationality
(l) reflex agent
(m) utility-based agent

5. **Problem 4 - Module 8 - Section 5.3.1 Cross-Validation [7]** ***This is a Collaborative Problem***
15 Points Tota

In this problem you are to develop and implement a k-fold cross validation algorithm. You are allowed to use either the Iris data set or the developed numerical features from HW2 to test your implementation. In this problem the following is to be completed:

(a) (5 points) Develop an algorithm to randomly separate data into groups of testing and training groups based on the number of desired folds/experiments, the term used will be k-fold cross validation. Use the 5-fold cross validation in Figure **??** as a reference.

(b) (5 points) Implement your k-fold cross validation algorithm.

(c) (2.5 points) Test your implementation using either the Iris data set or the numerical features generated from HW2.

(d) (52. points) Perform analysis and provide an explanation of how each k-fold creates subsets of the data, e.g., training and testing sets.

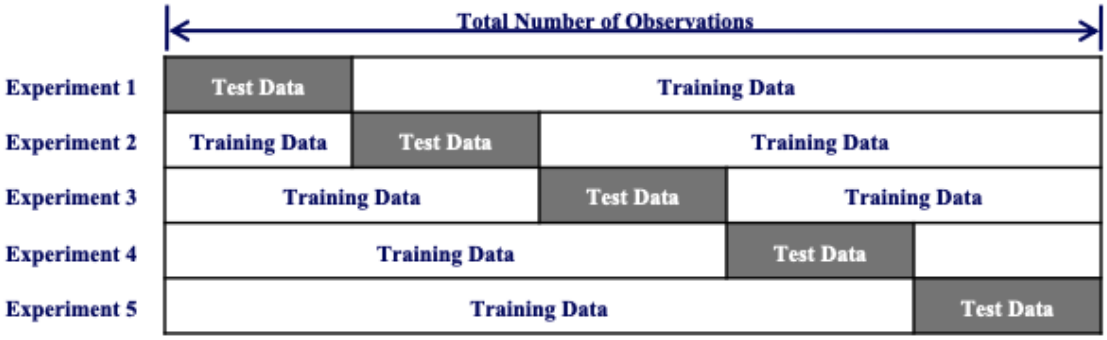


Figure 1: 5-Fold Cross Validation

6. **Problem 5 - Module 8**
35 Points Total

The Parzen window algorithm density model is optimized by maximizing the likelihood of the training data with the use of a Gaussian window surrounding each input data point. You are allowed to use either the Iris data set or the developed numerical features from HW2. In this problem the following is to be completed:

(a) [15 points] ***Note this is not Collaborative Problem***
Using the Gaussian kernel develop psuedo code to create a Parzen windowing system to accomplish the following steps:

i. Develop and implement the ability to read in data $\mathbf{x}_n$ with n observations and D dimensions (number of features).

ii. Using the Gaussian kernel in Eq. 27 of the Machine Learning document to develop and implement an algorithm to process an input observations and compare it with the training observations.

iii. Expand the development and implementation to handle multiple classes.

iv. [5 point bonus] use your solution from Problem 4 for a 5-fold cross validation implementation with the Parzen Window.

(b) [5 points] ***Note this is not a Collaborative Problem***

i. Calculate the running time of the system above in O-notation.

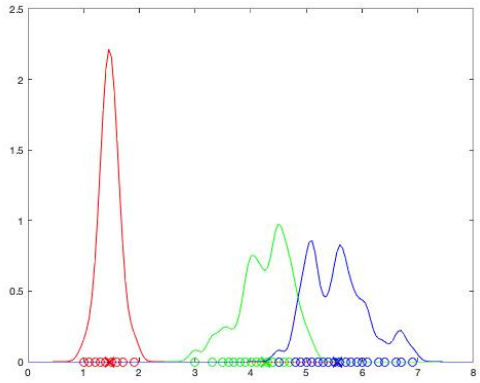
ii. Calculate the total running time of the above system as $T(n)$ with each line of pseudocode or code accounted for.

iii. How does the total running time $T(n)$ compare to the running time in O-notation?

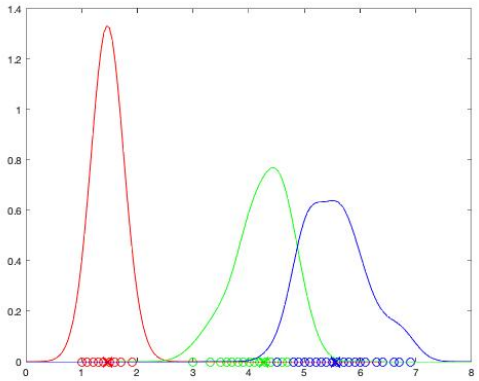
(c) [15 points] ***Note this is not a Collaborative Problem***

i. Using all observations and the petal length from the Iris data replicate the subfigures in Figure **??**.

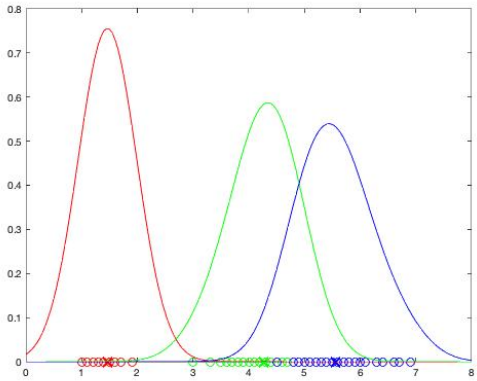
ii. Using all observations, the petal length and the petal width from the Iris data replicate the subfigures in Figure **??**.



(a) Gaussian Kernel with h = 0.1

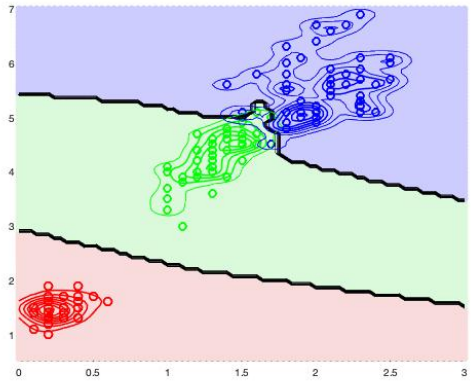


(b) Gaussian Kernel with h = 0.25

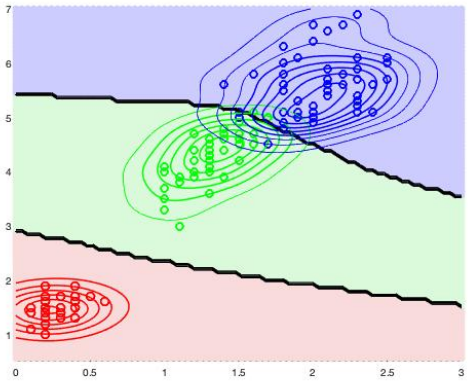


(c) Gaussian Kernel with h = 0.5

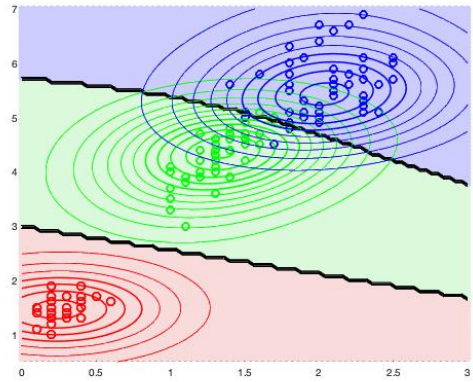
Figure 2: Iris Data - Petal Length with Setosa in Red, Versicolor in Green and Virginica in Blue



(a) Gaussian Kernel with h = 0.1



(b) Gaussian Kernel with h = 0.25



(c) Gaussian Kernel with h = 0.5

Figure 3: Iris Data - Petal Length vs Petal Width with Setosa in Red, Versicolor in Green and Virginica in Blue

References

[1] Bishop, Christopher M., *Neural Networks for pattern Recognition*, Oxford University Press, 1995

[2] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006, https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

[3] Duin, Robert P.W., Tax, David and Pekalska, Elzbieta, *PRTools*, http://prtools.tudelft.nl/

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B., *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society B, Volume 39, Number 1, pp.1–22, 1977

[5] Franc, Vojtech and Hlavac, Vaclav, *Statistical Pattern Recognition Toolbox*, https://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

[6] Fukunaga, Keinosuke, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972

[7] Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron, *Deep Learning*, MIT Press, 2016

[8] Russell, S., and Norvig, P., *Artificial Intelligence A Modern Approach*, 4^{th} Edition, Pearson, 2020

[9] Tomasi, C., *Estimating Gaussian Mixture Densities with EM – A Tutorial*, Duke University Course Notes, 2006, http://www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf, Retrieved Sept 2006