685.621 Algorithms for Data Science
Homework 3
Assigned at the start of Module 6
Due at the end of Module 8 - Tuesday March 23 (Midnight PST, 3AM EST)
Total Points 100/100

Collaboration groups will be assigned on Tuesday march 2nd in Blackboard. Make sure your group starts one thread for the collaborative problems. 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.

1. **Problem 1** ***This is a Collaborative Problem***
   30 Points Total

   In this problem, you will be developing pseudocode and implementing your development in Java, Matlab, Python or R for the Expectation Maximization method.

   (a) The development and implementation should be for a generic number of clusters, features and observations.

   (b) Apply your implementation using the features generated from HW 2 for the numerical data set.

      i. Use the top two ranked features.
      ii. Create 4 clusters using the 4 numerical values that have the best separation.
      iii. Display the 4 numerical values using 4 different colors for a good visual representation.
      iv. Provide an analysis of your results, e.g., what is your observation of the results.

2. **Problem 2 - Chapter 1 and 2 [8]** ***Note this is not Collaborative Problem***
   10 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

3. **Problem 3 - Chapter 2 [8]** ***This is a Collaborative Problem***
30 Points Tota

For your Tic-Tac-Toe implementation from PA1 add an agent to evaluate the board and the next best move.

- Best Move (Provided)
- (Completed from PA1) Implement a method that uses conditional statements to play against a user of your code. You must always check to see if your AI has a WIN and take that move, if not check to ensure your opponent does not have a WIN, if your opponent has a WIN possibility you must block your opponent.
- Develop in pseudocode or code the agent of your choice (this is your algorithm you will need to provide the running time for).
- Implement the agent in your Tic-Tac-Toe game to make the next best move according to your intelligent agent.
- Provide the efficiency (running time) of your algorithm in O-notation.
- Provide the total running time of your algorithm in $T(n)$ as well as showing the cost at each line of code in your algorithm.

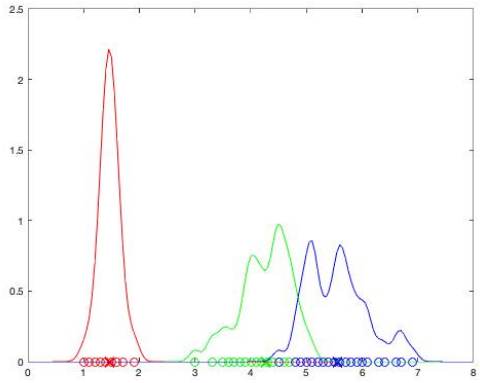
4. **Problem 4**
30 Points Total

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. In this problem the following is to be completed:

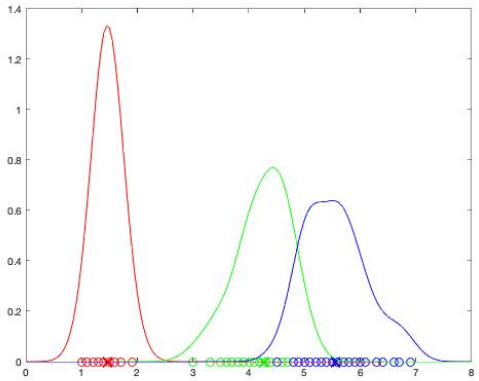
(a) [15 points] ***Note this is not Collaborative Problem***
Using the Gaussian kernel develop psuedo code to create a Parzen windowing system to accomplish the following steps:
   i. Develop the ability to read in data $\mathbf{x}_n$ with n observations and D dimensions (number of features).
   ii. Develop the ability to randomly remove 20% of the observations per class and assign the observations as test data with the remaining 80% of the observations as training data.
   iii. Using the Gaussian kernel in Eq. 24 of the Machine Learning document to develop an algorithm to process an input observations and compare it with the training observations.
   iv. Expand the development to handle multiple classes.

(b) [10 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?

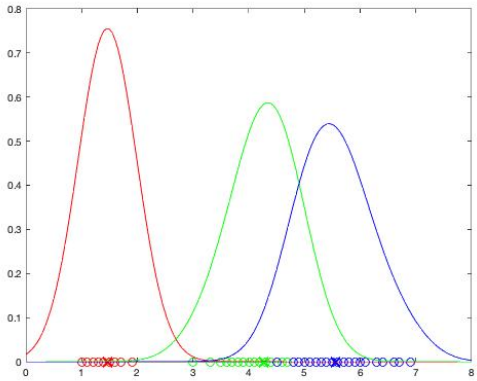
(c) [15 points] ***Note this is not a Collaborative Problem***
   i. Using all observations and the petal length from the Iris data replicate the subfigures in Figure 1.



(a) Gaussian Kernel with h = 0.1

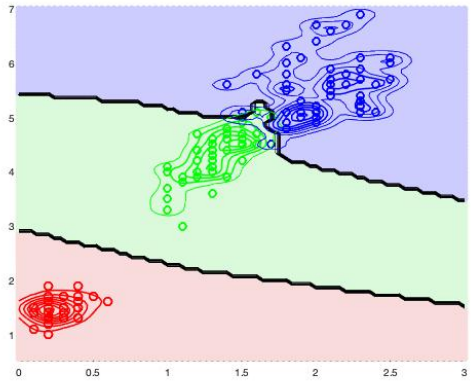


(b) Gaussian Kernel with h = 0.25

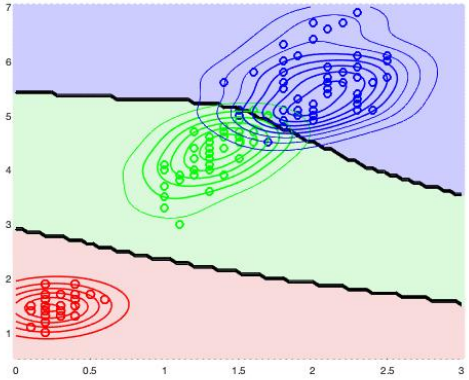


(c) Gaussian Kernel with h = 0.5

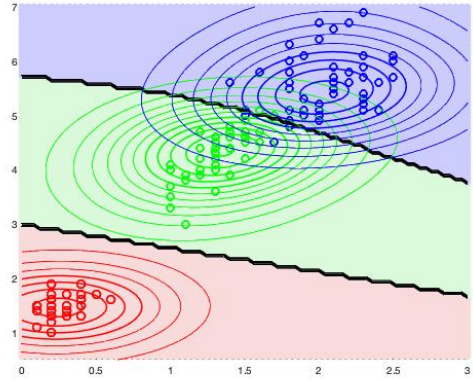
Figure 1: Iris Data - Petal Length with Setosa in Red, Versicolor in Green and Virginica in Blue



(a) Gaussian Kernel with h = 0.1



(b) Gaussian Kernel with h = 0.25



(c) Gaussian Kernel with h = 0.5

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

ii. Using all observations, the petal length and the petal width from the Iris data replicate the subfigures in Figure 2.

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] Machine Learning at Waikato University, *WEKA*, https://www.cs.waikato.ac.nz/ ml/index.html

[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