[CSC 5825 Fall 2017]

Due. Before the Class of Nov, 6, 2017

Full credit: 100 points with extra 10 points bonus

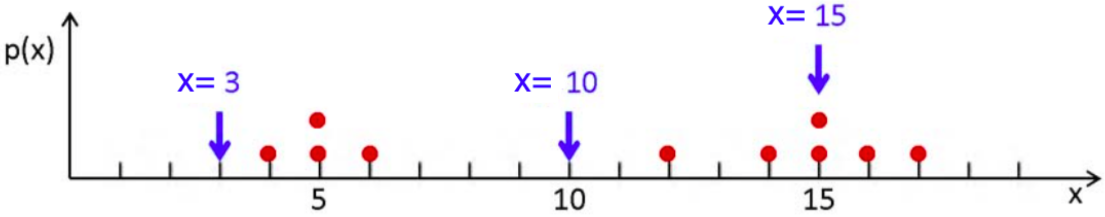
October 23, 2017

Question 1. Nonparametric Methods (20 points)
Recall the kernel estimator you learned in class:

$$\hat{p}(x) = \frac{1}{Nh}\sum_{t=1}^{N} w(\frac{x - x^t}{h})$$

$$w(\mu) = \begin{cases} \frac{1}{2} & if|\mu| < 1 \\ 0 & otherwise \end{cases}$$

Estimate the density $\hat{p}(x)$ at $x = 3, 10$ and 15 with $h = 4$, given the training data $x = \{4, 5, 5, 6, 12, 14, 15, 15, 16, 17\}$.



Question 2. Clustering Methods (80 points) 1. Iimplement the k-means clustering algorithm. Your function should allow users to give a data set, k and initial centers as the input parameters. Your function should output cluster label for each state and the converged centers. (30 points)

2. Implement an agglomerative hierarchical clustering algorithm. Your function should allow users to give data set and linkage method (way to calculate distances between clusters) as input arguments. Your function must produce the following outputs: the merging process and the height of the dendrogram corresponds to each merging. (30 points)

Hints: you can implement a function that outputs clustering memberships at any stage of merging. Also see below for examples:

merge an $n - 1$ by 2 matrix. n is the number of states. Row i of merge describes the merging of clusters at step i of the clustering. If an element j in the row is negative, then observation $-j$ was merged at this stage. If j is positive then the merge was with the cluster formed at the (earlier) stage j of the algorithm. Thus negative entries in merge

indicate agglomerations of singletons, and positive entries indicate agglomerations of non-singletons.
height a set of $n-1$ real values. The clustering height: that is, the value of the criterion associated with the clustering method for the particular agglomeration.

Use your functions to cluster cancer patients from 21 cancer types and validate your results using RAND index and adjusted RAND index. You may use the existing implementations of RAND index without penalty. (20 points) Each cancer data is stored in a separate csv file. Each row corresponds to a patient. The first column means survival time and the second column is an indicator of death versus live. You should use ALL the other columns to do clustering patients with different cancer types, which means you will need to combine all the files into one. **BONUS:** You may use 21 as a reference number but please find an optimal number of clusters (k) to partition your data and justify your choice. (10 points)

Submission Instructions

To earn the full credit, your must type your solutions with sufficient details either using LaTeX or Microsoft Word **embedded** with source code and plots. Please submit it through the blackboard website. You will get points off for not following these requirements.

Late homework (w/o acceptable documents) will be accepted with penalty. 20% off penalty if late for 24 hours or less. 40% off penalty if late between 24 hours and 48 hours. 60% off penalty if late between 48 hours and 72 hours. 80% off penalty if late between 72 hours and 96 hours. No homework late more than 96 hours will be accepted.