1. **Unsupervised Learning: K-means Clustering**

Consider a list of the magnitude of the earthquakes around the Pacific Rim measured in Richter scale (and rounded for ease of manual arithmetic). Group them into three clusters using K-means. Show all the steps. **Start with centers at {2, 8,and 7} and insert all the elements in the list below in a cluster.** If there is tie for distance add it to the higher cluster. Based on the cluster size, what do you think is the most probable strength of the next earth quake in that region? Use Euclidian distance for closeness.

**{2, 8, 7, 9, 8, 2, 3, 1, 4, 5, 1, 8, 2, 6}**

1. **(15 points) Semi-supervised learning: Classification using K-NN**
   1. Plot the data given in the table below that shows the age, loan amount (in thousands) and whether they have **defaulted** or not. (NOTE: “Default” means they did not pay back the loan.)
   2. Using this information and K-NN determine if a client of age 30 and a loan of $90K will default or not. Use (i) K = 3, (ii) K =5.
   3. How does value of K affect the result?

|  |  |  |
| --- | --- | --- |
| Age | Loan (X 1000) | Defaulted |
| 50 | 90 | N |
| 55 | 70 | N |
| 25 | 40 | N |
| 35 | 60 | N |
| 45 | 80 | N |
| 20 | 20 | N |
| 35 | 100 | N |
| 50 | 20 | Y |
| 20 | 95 | Y |
| 40 | 60 | Y |
| 60 | 100 | Y |
| 50 | 120 | Y |
| 30 | 110 | Y |