**Assignment 5**

Q1[4]. Consider the following dataset.

|  |  |  |
| --- | --- | --- |
| Age | Income | Owns Car (classification attribute) |
| Young | Middle class | Yes |
| Old | High | Yes |
| Young | Low | No |
| Middle age | High | Yes |
| Young | High | No |
| Old | Middle class | No |
| Middle age | Low | Yes |
| Young | Low | Yes |
| Old | Middle class | No |
| Old | High | Yes |

Convert the dataset in numeric data (e.g, young =1, middle age = 2, old = 3, apply the same for income). Apply the kNN approach to classify someone who is middle age and middle class. Use k = 3. Use the Euclidian distance to compute the distance between two samples. Break ties by rolling a dice.

Q2[6]. Consider the following data.

|  |  |  |
| --- | --- | --- |
| ID (not used for clustering) | age | salary (in thousands) |
| 1 | 23 | 22 |
| 2 | 33 | 50 |
| 3 | 40 | 80 |
| 4 | 11 | 5 |
| 5 | 70 | 30 |

Show two iterations of applying k-means clustering to the data. Use k = 3. Pick records 1, 3, and 5 as the initial centroids.

Q3[2]. What are the main weaknesses of k-means clustering?

Q4[2]. Give an example (not discussed in class) that would be a perfect application for neural networks.

Q5[2]. What is the difference between regression and classification?

Q6[2]. What is the difference between supervised and unsupervised learning?

Q7[2]. Why applying a polynomial transformation is sometimes useful before applying linear regression?

Q8[2]. What are the parameters of the model.fit method? Give a whole program that demonstrates how to use the method.

Q9[2]. What does the method make\_pipeline do? Show an example of how to use it.

Q10[2]. Explain what the method model.predict does. Should it be called before or after the method model.fit?

Q11[2]. When building a neural network using the Keras package, explain what do the methods model.add and model.compile do?

Q12[2]. Explain what the method model.evaluate does? What is the input? What is the return value?

Q13[4]. Explain a metric to test the quality of a clustering algorithm. Your metric should return a number between 0 and 1. Zero means that the algorithm did a horrible job in clustering the elements, while 1 means that the algorithm clustered the elements the same way as a human would. The ground truth, that is, the way a human clustered the elements, should be part of the input to your metric.

Q14[2] Explain one way to determine the value of *k* in *k*-means clustering.