Problem 1

a. K means is only capable of discovering shapes that are convex polygons [1]

Cannot discover 'X' shape because 'X' is not convex. [1]

DBSCAN can discover 'X' shape. [1]

b. K-means is prototype based and it doesn't use actual object in the dataset as centroids. PAM uses actual objects in the data set as representatives [2]; PAM uses an explicit objective function that is minimized whereas K-means minimizes an objective function implicitly without ever call it. [1]

c. Old centroids: (3,3) (4,3) (5,5)

New clusters:
C1 (3,3): (2,2), (0,3)
C2 (4,3): (4,4), (8,0)
C3 (5,5): (4,6), (5,5)

New centroids: (1,5/2) (6,2) (9/2,11/2)

Forgot to compute the new centroids (-1); error in centroid computations (-1); incorrect clustering results (-2 points); more than 2 errors: 0 points.

d. i. Relatively efficient: O(t*k*n*d), where n is # objects, k is # clusters, and t is # iterations, d is the # dimensions.
ii. Storage only O(n)—in contrast to other representative-based algorithms, only computes distances between centroids and objects in the dataset, and not between objects in the dataset; therefore, the distance matrix does not need to be stored.
iii. Easy to use; well studied
iv. Finds local minimum of the SSE fitness function.
v. Implicitly uses a fitness function

Each strength worth 1 point. Maximum of 3 points.

Problem 2

a. 1 cluster {A, C, D, E}

Each core point worth 1 point

b. B & F are outliers, since they are not core or border points [2]

A is border point because it is within the neighbor of a core point, but has only one other point in its ε-neighborhood [1]

c. If Eps= 10, then there will be one cluster containing all the objects.

No partial credit here!

d.

1. Select a random point P
2. If P is a core point, retrieve all points density reachable from P with Epsilon and Midpoints
3. If P is a core point, a cluster is formed
4. Visit the next point P that is not in a cluster yet; if there is no such point terminate!
5. Continue with step 2

Verbal descriptions how clusters are formed are also fine…

e. The two points will be in the same cluster [2]

Problem 3

One possible answer: Ignore SSN as it is not important.

Normalize Rate using Z-score and find distance by L-1 norm

We convert the Oph rating values ‘good, ‘medium’, and ‘poor’: 2:0 using a function ϕ

Find distance by taking L-1 norm and dividing by range i.e. 3

Assign weights 0.4 to Rate, 0.4 to Services and 0.2 to Oph

use the Jaccard distance function for the services: $d_{services}(ser1,ser2)= 1- (|ser1 \cap ser2|)/(|ser1 \cup ser2|)$

Now:

$d(u,v) = 0.4*|(u.Rate)/40 – (v.Rate)/40| + 0.2*|\phi(u.Oph) – \phi(v.Oph)|/2 + 0.4*d_{services}(u.services, v.services)$

2 customers:

c1=(111111111, ‘good’, 120, {A,B,C}) and c2=(222222222, ‘poor’, 80, {C,D,E})!

= 0.4[(120-80)/40] + 0.2*[(2- 0)/2] + 0.4*4/5= 0.3+4+0.2=0.92

If distance functions do not make much sense give 2 points or less

Distance functions are not defined properly [-4-7]

One error [-2 to -3]; two errors [-5 to -6]; more than 2 errors at most 1 point!

Problem 4

a. The box represents IQR/distribution of data [1]

The size of the box is a good estimation for the standard deviation/IQR of the dataset [1]

The line across the box indicates median [0.5]; it has a value of 6 [0.5] for the dataset

The whiskers indicate the minimum and maximum in the data without the outliers [1]

They have the value of 1 [0.5] and 18 [0.5] in the above box plot

The circle above the uppers whisker indicates outlier [0.5]

The location of the box plot between whiskers tells us distribution of data/information about spread of data/ data is skewed [2]

According to the boxplot the following attribute values are outliers: 22 [0.5]

b. The two attributes have no linear relationship. [2]

c. There is overlap between Benign and Malignant classes, Normal is easy to identify.

   Normal has horizontal and vertical spread whereas Benign has a horizontal spread. Function 1 is useful for classifying Normal vs Function 2 is more useful for classifying Benign.

   The distribution is unimodal [Not mentioned this point: -2]

Verbal descriptions how clusters are formed are also fine…

d. Men are taller than women [Not mentioned this point: -2]
   It has no significant gaps
   The number of men is more than number of women
   Female height is skewed towards right
   There are no outliers.

Verbal descriptions how clusters are formed are also fine…

Problem 6

a. Goals:

1) Find representatives for homogeneous groups →**Data Compression**
2) Find "natural" clusters and describe their properties →**"natural" Data Types**
3) Find suitable and useful grouping →**"useful" Data Classes**
4) Find unusual data object →**Outlier Detection**

Classification is supervised learning. Datasets consists of attributes and class labels. Here goal is to predict classes from the attribute values[4]

b.

- Classification: Find a *model* for class attribute as a function of the values of other attributes
- Prediction: Use some variables to predict unknown or future values of other variables
- Association Rule Discovery: Given a set of records each of which contain some number of items from a given collection, produce dependency rules which will predict occurrence of an item based on occurrences of other items
- Sequential Pattern Discovery: Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.
- Regression: Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Deviation Detection / Anomaly Detection: Detect significant deviations from normal behavior

Each task worth 1 point. Maximum of 7.5 points.