CSCI4390/6390 – Data Mining
Fall 2009, Exam III
Total Points: 100 + 10 (bonus)

1. (20 points) For the distance matrix below, use the group average method for cluster proximity to generate the hierarchical cluster dendogram. Show the updated distance matrix at each step. Whenever there is a tie, choose the cluster containing the smallest labeled item to merge first.

	A	B	C	D	E
A	0	1	3	2	4
B		0	3	2	3
C			0	1	3
D				0	5
E					0

First we merge A+B. The updated distance matrices will be:

	C	D	E
AB	3	2	3.5
C		1	3
D			5

Next we merge C+D, updated matrix:

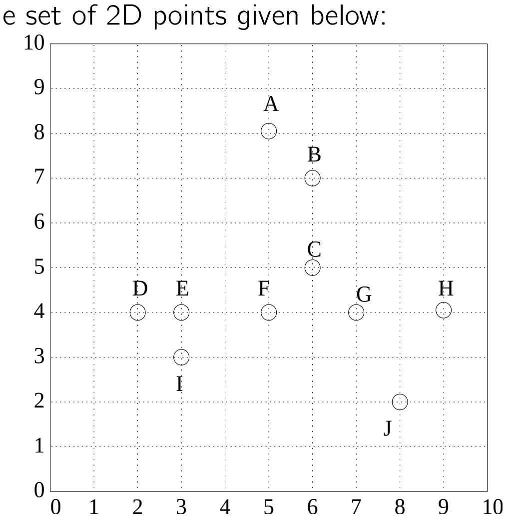
	CD	E
AB	2.5	3.5
CD		4

Next we merge AB+CD, updated matrix:

	E
ABCD	3.75

Finally we merge ACBD+E

2. (20 points) Consider the set of 2D points given below:



Assume $\epsilon = 2$, $minpts = 3$. For any point $\mathbf{x}$ define the ball of radius ϵ around $\mathbf{x}$ as follows:

$$B_\epsilon(\mathbf{x}) = \{\mathbf{y} \; : \; L_{\frac{1}{2}}(\mathbf{x}, \mathbf{y}) \le \epsilon\}$$

where the $L_{\frac{1}{2}}$ is the *fractional norm*, given as:

$$L_{\frac{1}{2}}(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{d} \sqrt{|x_i - y_i|} \right)^2$$

(a) Draw the shape of the ball of radius $\epsilon = 2$ around some point $\mathbf{x}$.

Assuming the center at $(0, 0)$, we can see that that points $(0, \pm 2)$, and $(\pm 2, 0)$ are all within the ball. Also the points $(\pm\frac{1}{2}, \pm\frac{1}{2})$ are within the ball. It is easy to draw the (star-like) shape of the ball from these point coordinates.

(b) Using the DBSCAN approach, idenfity all core, border and outlier points

The core points are: E, F, G

The border points are: D,H,I

The outliers are: A,B,C,J

(c) Report the final density-based clusters (based on DBSCAN).

There is only one cluster: $\{D, E, F, G, H, I\}$

3. (20 points) Using the same dataset as the one in question 2 above, and assuming that $h = 4$, answer the following questions:

(a) What is the probability density at E using the discrete kernel?

The density at E is $p(E) = \frac{1}{11 \cdot 4^2} \cdot 4 = \frac{1}{10 \cdot 4} = \frac{1}{40} = 0.025$.

(b) What is the gradient at E using the Gaussian kernel, but using only the 3 nearest neighbors of E (not including E)?

The gradient is

$$\begin{aligned}
\nabla p(E) &= \frac{1}{10 \cdot 4^4} \frac{1}{0.159} \left[e^{\frac{-1}{2 \cdot 4^2}} \left(\begin{pmatrix} 2 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 4 \end{pmatrix} \right) + e^{\frac{-1}{2 \cdot 4^2}} \left(\begin{pmatrix} 3 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 4 \end{pmatrix} \right) + e^{\frac{-4}{2 \cdot 4^2}} \left(\begin{pmatrix} 5 \\ 4 \end{pmatrix} - \begin{pmatrix} 3 \\ 4 \end{pmatrix} \right) \right] \\
&= \frac{1}{2560} \left[0.154 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + 0.154 \begin{pmatrix} 0 \\ -1 \end{pmatrix} + 0.1405 \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right] \\
&= \frac{1}{2560} \begin{pmatrix} 0.127 \\ -0.154 \end{pmatrix} \\
&= \begin{pmatrix} 4.96 \times 10^{-5} \\ -6.02 \times 10^{-5} \end{pmatrix}
\end{aligned}$$

4. (20 points) Given the points shown in the table below, find all axis-parallel subspace clusters using the level-wise CLIQUE approach. Assume that each dimension has range $[0, 5]$, and assume 5 bins of unit length along each dimension, of the form $[0, 1)$, $[1, 2)$, and so on. Density of a cell is defined as the number of points in that cell. Use a minimum density threshold of 3 points to find the clusters. Merge any clusters that share a face.

	X	Y	Z
p_1	0.5	4.5	2.5
p_2	2.2	1.5	0.1
p_3	3.9	3.5	1.1
p_4	2.1	1.9	4.9
p_5	0.5	3.2	1.2
p_6	0.8	4.3	2.6
p_7	2.7	1.1	3.1
p_8	2.5	3.5	2.8
p_9	2.8	3.9	1.5
p_{10}	0.1	4.1	2.9

We find the following dense intervals in 1D:
$X : [0, 1)$ with points 1,5,6,10
$X : [2, 3)$ with points 2,4,7,8,9

$Y : [1, 2)$ with points 2,4,7
$Y : [3, 4)$ with points 3,5,8,9, and $Y : [4, 5)$ with points 1,6,10, which be merged into the cluster:
$Y : [3, 5)$, with points 1,3,5,6,8,9,10

$Z : [1, 2)$ with points 3,5,9, and $Z : [2, 3)$ with points 1,6,8,10, which will be combined into one cluster: $Z : [1, 3)$ with points 1,3,5,6,8,9,10

For 2D cells we have: $X : [0, 1), Y : [4, 5)$ with points 1,6,10
$X : [2, 3), Y : [1, 2)$ with points 2,4,7
$X : [0, 1), Z : [2, 3)$ with points 1,6,10
$Y : [3, 4), Z : [1, 2)$ with points 3,5,9
$Y : [4, 5), Z : [2, 3)$ with points 1,6,10

Finally we have one 3D cell: $X : [0, 1), Y : [4, 5), Z : [2, 3)$ with points 1,6,10

5. (20 points) Given the two points $\mathbf{x}_1 = (1, 2)$, and $\mathbf{x}_2 = (2, 1)$, use the kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^2$$

to find the kernel principal component.

(a) Compute the kernel matrix $\mathbf{K}$ and center it in feature space.

(b) Find the first principal component, and the corresponding eigenvalue of the centered kernel matrix.

The kernel matrix is

$$\mathbf{K} = \begin{pmatrix} 25 & 16 \\ 16 & 25 \end{pmatrix}$$

We can center it in feature space as follows:

$$\hat{\mathbf{K}} = \mathbf{K} - \mathbf{1}_n\mathbf{K} - K\mathbf{1}_n + \mathbf{1}_n\mathbf{K}\mathbf{1}_n$$

We have: Note that

$$\mathbf{1}_2\mathbf{K} = \begin{pmatrix} 25 & 16 \\ 16 & 25 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 20.5 & 20.5 \\ 20.5 & 20.5 \end{pmatrix}$$

Also

$$\mathbf{K}\mathbf{1}_2 = \begin{pmatrix} 20.5 & 20.5 \\ 20.5 & 20.5 \end{pmatrix}$$

And

$$\mathbf{1}_2\mathbf{K}\mathbf{1}_2 = \begin{pmatrix} 20.5 & 20.5 \\ 20.5 & 20.5 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 20.5 & 20.5 \\ 20.5 & 20.5 \end{pmatrix}$$

Therefore

$$\hat{\mathbf{K}} = \begin{pmatrix} 25 & 16 \\ 16 & 25 \end{pmatrix} - \begin{pmatrix} 20.5 & 20.5 \\ 20.5 & 20.5 \end{pmatrix} = \begin{pmatrix} 4.5 & -4.5 \\ -4.5 & 4.5 \end{pmatrix}$$

We can compute the dominant eigenvector and eigenvalue of $\hat{\mathbf{K}}$ as follows:

$$\begin{pmatrix} 4.5 & -4.5 \\ -4.5 & 4.5 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4.5 \\ -4.5 \end{pmatrix} \to \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\begin{pmatrix} 4.5 & -4.5 \\ -4.5 & 4.5 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 9 \\ -9 \end{pmatrix} \to \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

This implies that the eigenvector is $\mathbf{a} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and the eigenvalue is $\eta_1 = 9$.

We can now extract the actual eigenvalue λ_1 as follows:

$$\lambda_1 = \eta_1/2 = 9/2 = 4.5$$

Also we need to scale $\mathbf{a}$ so that $\|\mathbf{a}\|^2 = \frac{1}{9}$. The right scaling constant is 1/3, so the normalized $\mathbf{a}$ vector should be: $\frac{1}{3\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

6. (**Bonus:** 10 points) The normalized symmetric Laplacian matrix is given as:

$$\mathbf{L}_s = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$$

Answer any **one** of the following questions:

(a) Prove that $\mathbf{L}_s$ has the smallest eigenvalue $\lambda_n = 0$

(b) Prove that $\mathbf{L}_s$ is positive semi-definite.