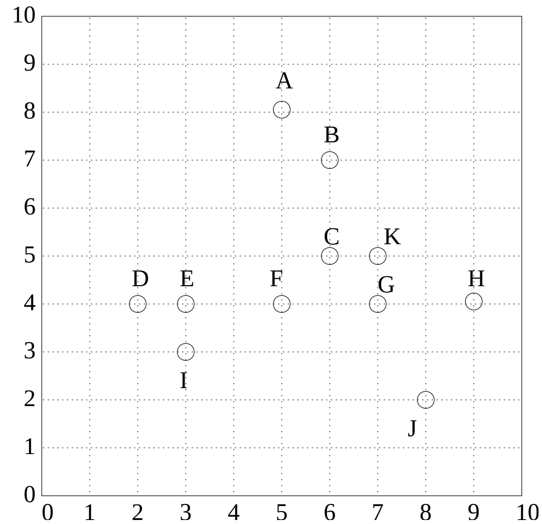
CSCI4390/6390 – Data Mining
Fall 2007, Quiz 11
Total Points: 30

Name:____

1. (15 points) Consider the dataset shown below:



Assuming the following discrete kernel function:

$$K(\mathbf{z}) = \begin{cases} 1 & \text{If } \sum_{j=1}^{d} |z^j| \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

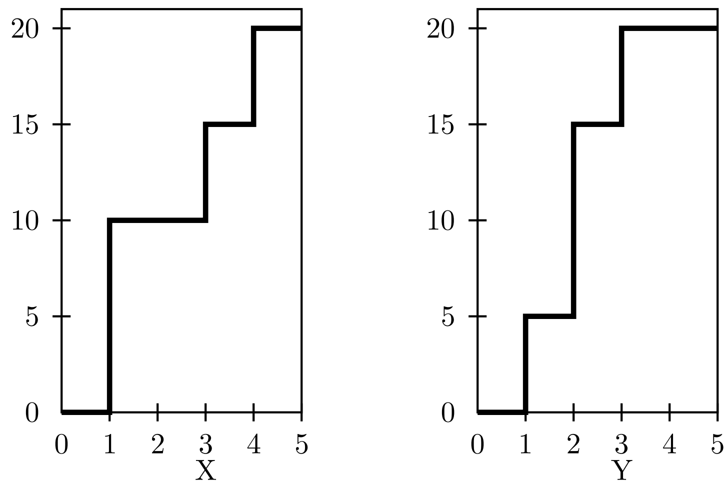
where z^j is the component of $\mathbf{z}$ along the j-th dimension.

Compute the gradient vector at the point F, with $h = 2$. Recall that the gradient vector at a point $\mathbf{x}$ is computed as:

$$\nabla K(\mathbf{x}) = \sum_{\mathbf{x}_i} (\mathbf{x}_i - \mathbf{x}) K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

.

2. (15 points) Let the cumulative probability distributions measured at bin boundaries (at multiples of the unit length) be as follows for dimensions X and Y:



For example, along the X dimension, there are 10 points in the range $[1, 2]$, and none in the range $[2, 3]$, that is why the cumulative distribution does not change between the two consecutive intervals.

Given the cumulative distribution above, find all single dimensional dense clusters, given that minimum density threshold is 10 points. Report only the minimal dense clusters along each dimension, i.e., the smallest closed intervals with enough density. Note that these clusters can overlap.

Finally, what can you say about the minimal dense clusters in the joint 2D space spanned by X and Y?