CS 557 STATISTICAL PATTERN RECOGNITION & LEARNING
MIDTERM EXAM 2 SOLUTIONS
Fall 2016

QUESTION 1 **(Marks 5)**

Suppose we are designing a system for regression on two input variables, x_1, x_2. Derive the normal equation for the weight w_0 if we use mean square error as the optimization criterion and the following expression for making predictions o for the output variable:

$$o = x_2 x_1 w_0 + x_1 w_1 + x_2^2 w_2$$

SOLUTIONS

(do the detailed working yourself)

If we take the mean square error function as:

$$E = \frac{1}{2}\sum_i \left(o^i - y^i\right)^2$$

$$\frac{\partial E}{\partial w_0} = \sum_i \left(o^i - y^i\right) x_2^i x_1^i$$

Setting the above equal to zero we get

$$\sum_i x_2^i x_1^i y^i = w_0 \sum_i (x_2^i x_1^i)^2 + w_1 \sum_i x_2^i (x_1^i)^2 + w_2 \sum_i x_1^i (x_2^i)^3$$

QUESTION 2 **(Marks: 5)**

Suppose we have four points **w**,**x**,**y**,**z**. The following table shows the pairwise similarity of the 4 points. You have to apply agglomerative clustering using complete link clustering to this data and draw the final dendogram. No marks will be given if no working or iterations are shown.

	w	**x**	**y**	**z**
w	10	2	1	3
x	2	10	5	0
y	1	5	10	4
z	3	0	4	10

SOLUTION

Iteration 1

	w	**x,y**	**z**
w	10	1	3
x,y	1	10	0
z	3	0	10

Iteration 2

	x,y	**w,z**
x,y	10	0
w,z	0	10

Iteration 3

One cluster x,y,w,z with similarity = 0

(make the final dendogram yourself)

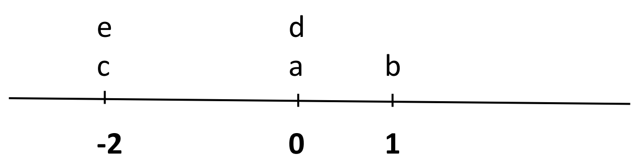
QUESTION 3 **(Marks: 5)**

The following table shows the pairwise Euclidean distance of 5 points. All the 5 points (named **a**,**b**,**c**,**d**,**e**) lie in a 50 dimensional space. You have to find a mapping of all these points in 1 dimensional space that preserves the distances perfectly. Use any method you like. Show all working/reasoning.

	a	**b**	**c**	**d**	**e**
a	0	1	2	0	2
b	1	0	3	1	3
c	2	3	0	2	0
d	0	1	2	0	2
e	2	3	0	2	0

SOLUTION

By observation we can solve this by looking at the distance between pairs of points



QUESTION 4 **(Marks 2+2+1)**

Suppose we have the following data (one dimensional): 10,12,12,13,13,13,13,14,14,14

a. The Naive density estimate at x=13.2 for h=2 is ____________________

b. Suppose the density is given by (for some constant c) :

$$p(x) = \frac{1}{c}\sum_{t=1}^{N} k\left(\frac{x - x^t}{h}\right) \quad \text{and } k(x) = \begin{cases} 2 \text{ if } |x| = 0 \\ 1 \text{ if } |x| < 3 \\ 0 \text{ otherwise} \end{cases}$$

The density at x = 12.2 for h=2 is _________________________

The density at x = 13.2 for h=1 is _________________________

c. If the label of the first 5 points is +1 and last five points is -1, then what is the classification of the point 12.6 when using 5 nearest neighbor algorithm.

SOLUTION

a. The Naive density estimate at x=13.2 for h=2 is ______7/20______________

b. Suppose the density is given by (for some constant c) :

$$p(x) = \frac{1}{c}\sum_{t=1}^{N} k\left(\frac{x - x^t}{h}\right) \text{ and } k(x) = \begin{cases} 2 \; if \; |x| = 0 \\ 1 \; if \; |x| < 3 \\ 0 \; otherwise \end{cases}$$

The density at x = 12.2 for h=2 is __________10/c______________

The density at x = 13.2 for h=1 is __________9/c______________

c. If the label of the first 5 points is +1 and last five points is -1, then what is the classification of the point 12.6 when using 5 nearest neighbor algorithm.

SOLUTION: +1