Exercise Sheet 1, Business Analytics, SoSe 2012, 03.05.2011
Dr. Tomáš Horváth, Osman Akcatepe

1. Figure 1 shows classification data with two classes: square and circle. The two instances with dotted lines, which have been labeled 1 and 2, have not been classified yet. Which class labels would be assigned to them by k-nearest neighbour for k = 1, k = 3 and k = 5?

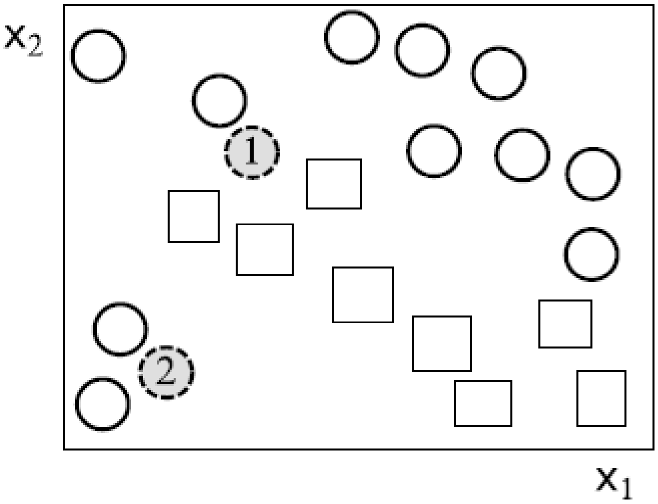

Figure 1: A classification data set

2. In a different data set, given in Table 1, the feature x_1 can take on three possible values: Square, Circle and Hexagon. The feature x_2 can take on any integer value, and you may assume that it makes sense to look at the difference between two of its values. What would be an appropriate way to represent these features for the k-nearest neighbour algorithm (assuming it uses Euclidean distance between instances)?

X1 ToyShape	X2 NumberOfSells	Y OfGoodQualityOrBadQuality
Square	1	Good
Square	24	Bad
Circle	0	Good
Hexagon	0	Good

Table 1: An artificial data set

3. Perhaps it would be better to measure NumberOfSells by the dozen. So suppose we would measure x_2 from Table 1 in units of twelve. For example, 1 => 1/12 and 24 => 2. How would this influence the k-nearest neighbour algorithm?

4. For each of the following learning problems, please indicate whether it is a prediction, regression or classification problem.:

(a) A search engine tries to determine whether a website is about sports based on the number of times the website contains the following words/phrases: ‘sports’, ‘football’, ‘tennis’, ‘hockey’, ‘elections’, ‘human rights’ and ‘party’.

(b) A farmer has used different amounts of fertilizer on different parts of his land. He has recorded the average height of his corn for each part. Now he wants to learn how the average height of his corn depends on the amount of fertilizer that he uses.

(c) Each spring a biologist counts the number of offspring in the same lion population. Based on her counts of the previous years she wants to estimate the number of offspring in the coming year.

5. Suppose that a sample of four numbers [1,2,13,15] is drawn from a population. In order to predict a fifth number not yet drawn from that same population, one can do “central tendency”: Which estimator will predict best: mean or median or midrange* ?

Good Luck!

* midrange is defined as M = (min X + max X) / 2