Problem Set K-nearest Neighbor - Key

1. Why does K-nearest neighbor algorithm suffer from curse of dimensionality?
   See chapter 1 of Bishop

2. Provide a reasonable approach for determining the value of K in the K-nearest neighbor algorithm.

   Try different K values with "leave one out" testing, select the one with least error.

3. CLASSIFICATION
   Imagine that you are given the following set of training examples. Each feature can take on one of three nominal values: a, b, or c.

F1	F2	F3	Category
a	c	a	+
c	a	c	+
a	a	c	-
b	c	a	-
c	c	b	-

Describe how a 3-nearest-neighbor algorithm would classify the test example given above. Use Hamming distance between two instances. Hamming distance is number of features that are same in two instances, so greater the hamming distance closer are they two instances.

F1 = a, F2 = c & F3 = b

The 3-nearest neighbor algorithm will use a distance measure to compute three nearest points to the test example and assign to it the majority class. Since the data is nominal, we will use hamming distance as the distance metric. Hamming distance is the number of attributes on which the two examples agree on.
Hamming distance of the test point to the first five examples is: 2, 0, 1, 1, 2.
Thus, we will assign to it the majority class of the first, third and fifth or that of first, fourth or fifth example. Both options yield majority class –.

4. Consider the training data given below, x is the attribute and y is the class variable.

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
y	A	A	A	A	B	A	A	A	A	B	B	B	B	A	B	B	B	B

a. What would be the classification of a test sample with x = 4.2 according to 1-NN?

B

b. What would be the classification of a test sample with x = 4.2 according to 3-NN?

A

c. What is the "leave-one-out" cross validation error of 1-NN. If you need to choose between two or more examples of identical distance, make your choice so that the number of errors is maximized.

8 out of 18

5. We have data from a questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here are the four training examples

X1 = Acid durability (in seconda)	X2 = Strength (Kg/sq meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new tissue that pass laboratory test X1 = 3 and X2 = 7. Without another expensive survey, can we guess the classification of the new tissue using K-nearest neighbor algorithm using k= 3?

Distance between data point X1=3 and X2=7 and point 1 = 4

Distance between data point X1=3 and X2=7 and point 2 = 5

Distance between data point X1=3 and X2=7 and point 3 = 3

Distance between data point X1=3 and X2=7 and point 4 = 3.605

So classification using k-nearest neighbor and k= 3 is Good.