CS685: Data Mining
Homework 2

Total Marks: 50

Due date: 11:00pm, 30th September, 2012

Submit all parts of a single question as a single zip file named rollno_hw1_q.zip. Replace rollno with your roll number (please omit 'Y', if present) and q with the question number.

1 Curse of Dimensionality

Q 1 (4 + 8 + 8 + 3 = 23)

Generate n d-dimensional points uniformly randomly within a $[0, 1)$ d-dimensional cube. Compute all pairwise Euclidean distances among these points. Assume $n = 10^5$. Vary d as $1, 2, 5, 10, 25, 50, 100$.

(a) Compute the standard deviation of the distances for each dimension.

(b) Draw histograms with 50 equal width bins. Use gnuplot with properly labeled axes.

(c) Repeat the above exercise after normalizing the distances by dividing them with $\sqrt{d}$.

(d) What are your conclusions?

2 Principal Component Analysis

Q 2 (6 + 4 + 2 + 2 + 1 = 15)

Perform PCA on the points given in pca_data.txt.

(a) Submit the code (in any language including Octave) for performing PCA. (You may use already existing code or function for SVD.)

(b) Plot the points. In the same graph, also show the two principal component axes.

(c) What are the principal component vectors and the corresponding eigenvalues?

(d) Reduce the dimensionality to 1. Submit the points with reduced dimensionality.

(e) How much energy is retained with dimensionality 1?

3 Sampling

Q 3 (3 + 3 + 6 = 12)

Generate a dataset of size $N = g \times k$. Divide the numbers into k equal groups of size g. Pick a sample of size n. Assume sampling with replacement. See if the sample contains at least one representative from each group. Repeat picking the sample for t times. Also, count the number of successes s. The empirical probability is, thus, s/t.

(a) Plot these empirical probabilities across different n for a fixed $g = 10^3$, $k = 5$ and $t = 500$. Take reasonable values of n.

(b) Plot these empirical probabilities across different k for a fixed $g = 10^3$, $n = 10^2$ and $t = 500$. Take reasonable values of k.

(c) Do you think the empirical probability depends on g and t? Justify.