

Homework 2, 290D, Fall 07

Due November 2, 5 pm

1. (20 points) Compute the covariance matrix for the dataset below. Compute the eigenvalues and eigenvectors of the matrix. Reduce the dimensionality to 1 by projecting the points along the first eigenvector. Compute the maximum distortion and stress of the resulting embedding.

3	8	10
4	6	12
18	34	8
7	4	2
22	8	9
7	6	7
4	22	13
11	19	15
8	32	14
16	18	4

2. (10 points)
 - a. Prove the orthonormality of the basis vectors of DFT.
 - b. Prove the following for Haar wavelets.
 - i. V_{j+1} is the orthogonal sum of V_j and W_j .
 - ii. W_j and W_k are orthogonal for different values of k and j .
3. (30 points) Consider the time series dataset available at <http://kdd.ics.uci.edu/databases/synthetic/synthetic.data.html> . Transform the first 9 data sections using wavelets and DFT. Use data elements 1, 101, 201, ..., 9901 from the 10th section for querying. Use the first 8 columns from the dataset. Consider top-20 query with the L_2 distance metric. Compute and plot the average precision and query time for 1,2,4,8 dimensions. The ground-truth for precision is defined with respect to the full 8 dimensions.

Repeat the above analysis for the skyserver dataset (Details forthcoming).

4. (40 points) This problem is about the extraction of features and dimensionality reduction. You will be working with the image dataset located at CSIL in the “/cs/class/cs290d/sandbox/homework2/cortina_images” directory.
 - a. Extract the MPEG features of scalable color descriptor (256 bins in the HSV color space; no other transformation) and color structure descriptor (color histogram for 256 bins generated with an 8x8 block in HMMD color space) for the datasets.
 - b. Reduce the number of dimensions to 4, 8, 16, 32 using PCA, wavelets, and DFT.

- c. Plot graphs for quality versus number of dimensions and query time versus number of dimensions in each case. Consider top-20 query with the L_2 distance metric.