BITS F464: MACHINE LEARNING

Lecture-07: Curse of Dimensionality, PCA and Eigenfaces

Dr. Kamlesh Tiwari
Assistant Professor
Department of Computer Science and Information Systems Engineering,
BITS Pilani, Rajasthan-333031 INDIA

Jan 24, 2018 (Campus @ BITS-Pilani Jan-May 2018)

Notes prepared by
Arnav Sailesh
2016A7PS0054P

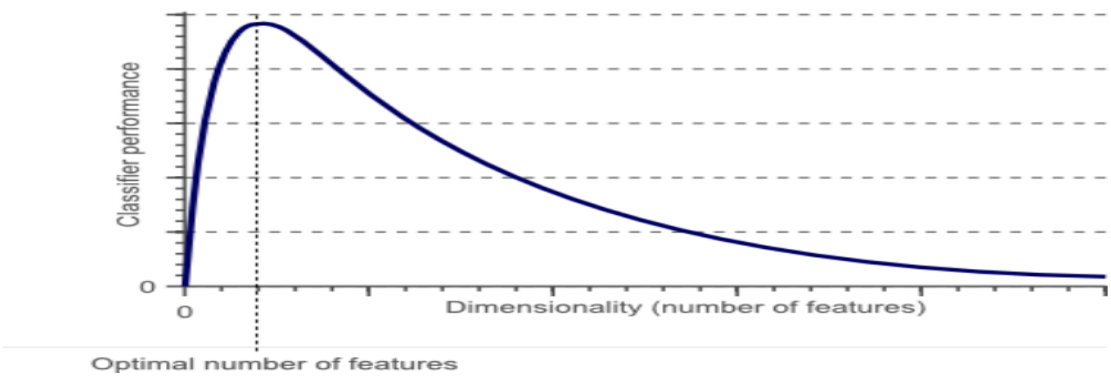
Curse of Dimensionality

The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.

It's typical Machine Learning implementation is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality.

Databases are generally of high dimension

- Images contain lot of pixels and text may contain lot of words (250×250 pixels, or 106 words)
- True dimensionality may be lot lower than the observed one.
- Data may be in some low dimensional manifold (sheet) in high dimensional space



Why it is a problem

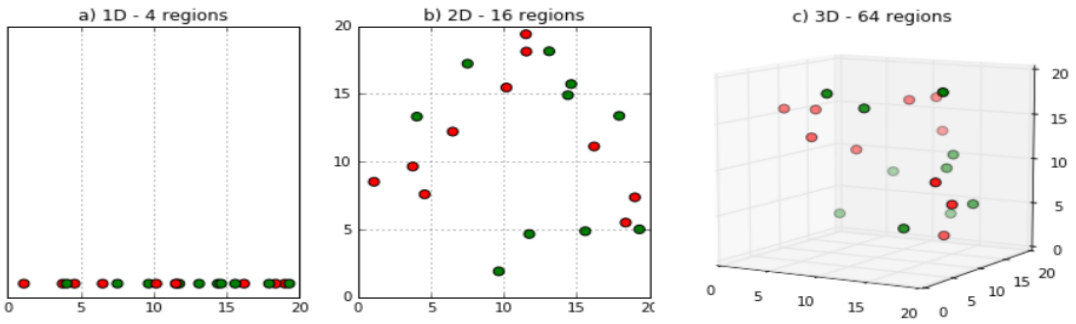
- Because most of the machine learning methods tends to count evidences
- Space grows very quickly but number example remains limited
- Density of data decreases sharply with the dimension. This lead to no observations for most of the cases

High Dimensional Data is difficult to work with since:

- Adding more features can increase the noise, and hence the error
- There usually aren't enough observations to get good estimates

This leads to-

- Increase in running time
- Overfitting
- Greater number of Samples Required



Some background math

Mean, Variance and Covariance

- Conceptually mean represents the center and variance the spread of data points
- Let $a = [a_1, a_2, ..a_n]$ and $b = [b_1, b_2..b_n]$ be two set of data (assume their mean be zero). And we want to find out whether these two are statistically independent? Covariance comes into picture

$$\sigma_a^2 = \frac{1}{n-1} a \times a^T \quad \sigma_b^2 = \frac{1}{n-1} b \times b^T \quad \sigma_{ab}^2 = \frac{1}{n-1} a \times b^T$$

- If a and b are of unit length then, the dot product a $\times$ b^T tells how much they are in same direction.
- If they are in same direction the value would be 1 and when they are orthogonal it would be 0.
- Consider covariance among all pair of data vectors:

$$C_X = \frac{1}{n-1} XX^T = \begin{bmatrix} \sigma^2_{a_1 a_1} & \sigma^2_{a_1 a_2} & \cdots & \sigma^2_{a_1 a_n} \\ \sigma^2_{a_2 a_1} & \sigma^2_{a_2 a_2} & \cdots & \sigma^2_{a_2 a_n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^2_{a_n a_1} & \sigma^2_{a_n a_2} & \cdots & \sigma^2_{a_n a_n} \end{bmatrix}$$

- Diagonal has variance and off diagonal covariance. It is a symmetric matrix
- If vectors are in same direction the covariance would be 1 and when they are orthogonal it would be 0.
- We want small covariance and large variance. i.e. large values at diagonal and small at rest of the places.

Diagonalization

- Let X be the data matrix, consider XX^T this is a symmetric matrix and self adjoint therefore one can always do eigen value decomposition.

  XX^T=SΛS^T

  where S is matrix of eigen vectors ($S^{-1} = S^T$) and Λ is diagonal matrix of eigen values

- Consider Y = S^TX, then

$$C_Y = \frac{1}{n-1} YY^T = \frac{1}{n-1} S^T X (S^T X)^T = \frac{1}{n-1} S^T X X^T S$$

$$C_Y = \frac{1}{n-1} S^T S \Lambda S^T S = \frac{1}{n-1} \Lambda$$

- Covariance matrix of S^TX is diagonal

PCA

- Similar operation can also be done by using SVD
- We know every matrix can be written as X = UΣV^T
- Let Y = U^TX

$$C_Y = \frac{1}{n-1} YY^T = \frac{1}{n-1} U^T X (U^T X)^T = \frac{1}{n-1} U^T X X^T U$$

$$C_Y = \frac{1}{n-1} U^T (U \Sigma^2 U^T) U = \frac{1}{n-1} \Sigma^2$$

Eigenfaces

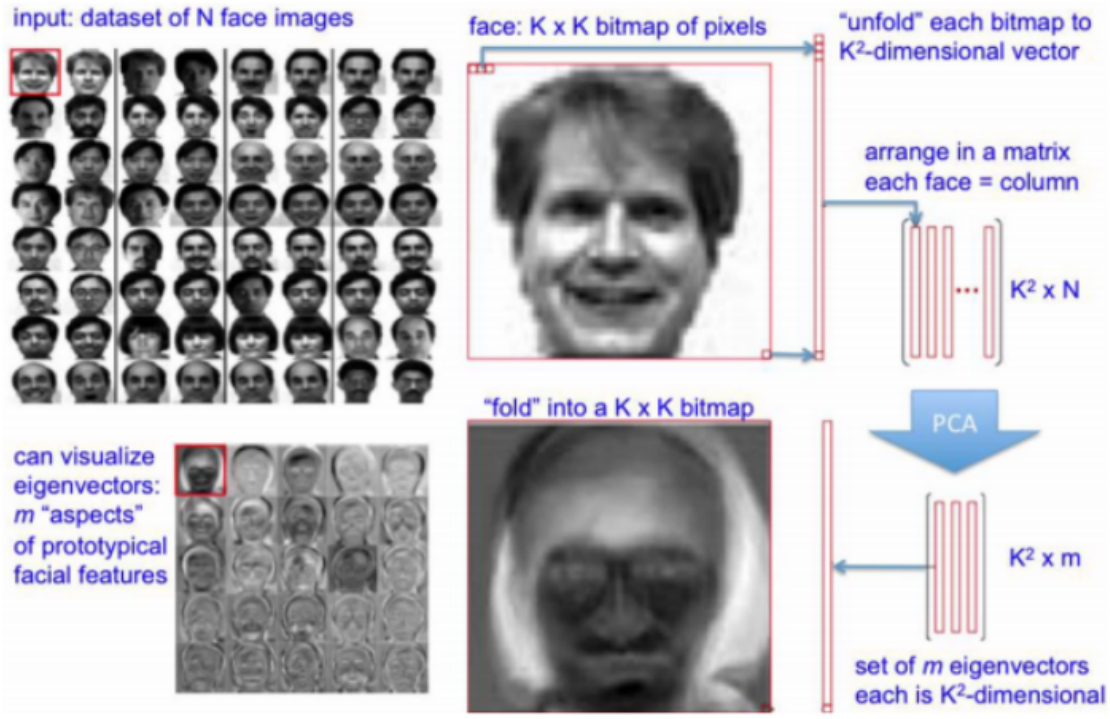


Figure 1: 2500 Face images, of 16 individuals; achieved 96 percent correct classification