Cheat Sheet – PCA Dimensionality Reduction

What is PCA?

- Based on the dataset find a new set of orthogonal feature vectors in such a way that the data spread is maximum in the direction of the feature vector (or dimension)
- Rates the feature vector in the decreasing order of data spread (or variance)
- The datapoints have maximum variance in the first feature vector, and minimum variance in the last feature vector
- The variance of the datapoints in the direction of feature vector can be termed as a measure of information in that direction.

Steps

1. Standardize the datapoints $$X_{new} = \frac{X - mean(X)}{std(X)}$$
2. Find the covariance matrix from the given datapoints $$C[i,j] = cov(x_i, x_j)$$
3. Carry out eigen-value decomposition of the covariance matrix $$C = V \Sigma V^{-1}$$
4. Sort the eigenvalues and eigenvectors $$\Sigma_{sort} = sort(\Sigma) \quad V_{sort} = sort(V, \Sigma_{sort})$$

Dimensionality Reduction with PCA

- Keep the first m out of n feature vectors rated by PCA. These m vectors will be the best m vectors preserving the maximum information that could have been preserved with m vectors on the given dataset

Steps:

1. Carry out steps 1-4 from above
2. Keep first m feature vectors from the sorted eigenvector matrix $$V_{reduced} = V[:, 0:m]$$
3. Transform the data for the new basis (feature vectors) $$X_{reduced} = X_{new} \times V_{reduced}$$
4. The importance of the feature vector is proportional to the magnitude of the eigen value

Figure 1

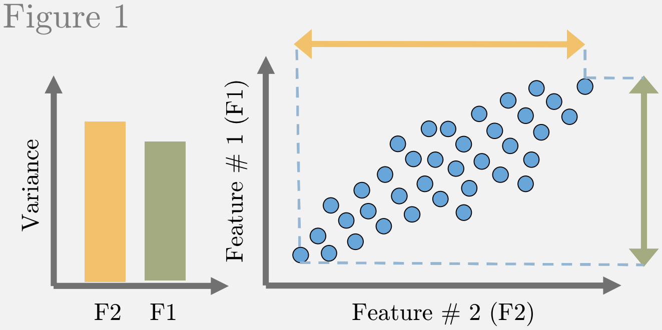

Figure 2

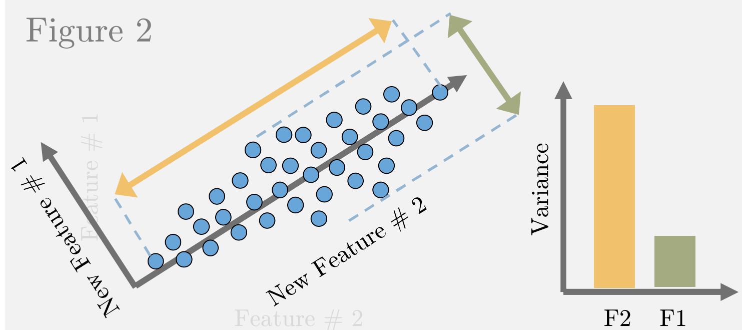

Figure 3

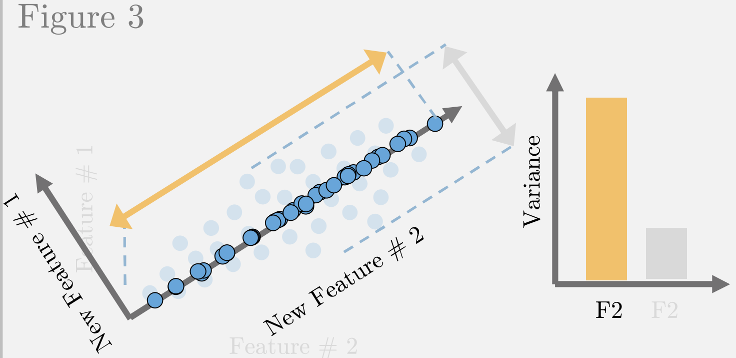

Figure 1: Datapoints with feature vectors as x and y-axis

Figure 2: The cartesian coordinate system is rotated to maximize the standard deviation along any one axis (new feature # 2)

Figure 3: Remove the feature vector with minimum standard deviation of datapoints (new feature # 1) and project the data on new feature # 2

Source: https://www.cheatsheets.aqeel-anwar.com