Principal Component Analysis (PCA)

Aaryan
CO21BTECH11001

PCA is a dimensionality reduction technique which tries to identify the subspace in which the data approximately lies.
Suppose we are given a dataset $\{x^{(i)}; i = 1, \dots, m\}$.
Let $x^{(i)} \in R^n$ for each i

Usually, some features of the data are strongly co-related with each other, that the data really lies approximately on a lesser dimensional subspace. To detect and perhaps remove this redundancy, we use PCA.

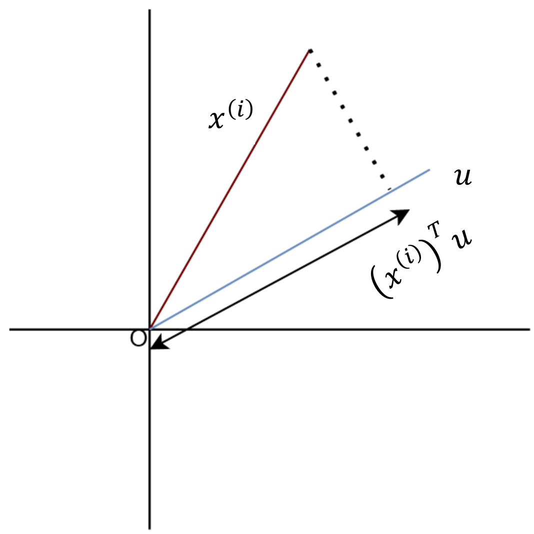
Prior to running PCA, we first pre-process the data to zero out the mean of the data and normalize the variance of the data –

1. Let $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
2. Replace each $x^{(i)}$ with $x^{(i)} - \mu$.
3. Let $\sigma_j^2 = \frac{1}{m}\sum_i \left(x_j^{(i)}\right)^2$
4. Replace each $x_j^{(i)}$ with $\frac{x_j^{(i)}}{\sigma_j}$

Let's suppose that we want to project (n dimensional) data to k dimensional data, where k<n.
Therefore, we should project the data on k vectors, say $u_1, u_2, \dots, u_k$, such that it represents maximum variance of data.

Let $u \in \{u_1, u_2, \dots, u_k\}$ and $x^{(i)}$ be a point in our dataset.

The length of projection of $x^{(i)}$ on u is given by ${x^{(i)}}^T u$.



To maximize the variance of the projections, we would like to choose a unit-length vector u so as to maximize:

$$\frac{1}{m}\sum_{i=1}^{m}\left({x^{(i)}}^T u\right)^2 = \frac{1}{m}\sum_{i=1}^{m} u^T x^{(i)} {x^{(i)}}^T u = u^T\left(\frac{1}{m}\sum_{i=1}^{m} x^{(i)} {x^{(i)}}^T\right)u$$

Maximizing this s.t. $||u|| = 1$, gives principal eigen vector of

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} {x^{(i)}}^T$$

which is empirical covariance matrix of the data.

Therefore $u_1, u_2, \ldots, u_k$ are top k eigen vectors of Σ i.e., eigen vectors corresponding to k dominant eigen values of Σ.

To represent $x^{(i)}$ in this basis,

$$y^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ . \\ . \\ u_k^T x^{(i)} \end{bmatrix} \in R^k$$

<u>Questions –</u>

1. PCA is a feature selection technique
   (a) True
   (b) False
   Ans. (b)
   In PCA, we obtain Principal Components axis, this is a linear combination of all the original set of feature variables which defines a new set of axes that explain most of the variations in the data. Therefore, it doesn't result in development of a model that relies upon a small set of the original features.

2. Why is it important to standardize the data before applying PCA?
   Ans. If we use features of different scales, we get misleading directions. So, we do standardization to assign equal weights to all the variables.

3. What is a good way to select how many dimensions to keep?
   Ans. Calculate the proportion of variance for each feature, pick a threshold (say 90%), and add features until you hit that threshold.

4. List 2 advantages of Dimensionality reduction.
   Ans. Less misleading data means model accuracy improves.
   Less data means less storage space required.

5. List 2 dis-advantages of Dimensionality reduction.
   Ans. Some information is lost.
   It makes the independent variables less interpretable.

6. What is the major dis-advantage of PCA.
   Ans. It doesn't work well for non linearly correlated data.