Dimensionality and PCA

Dimensionality reduction is a crucial step in data preprocessing, especially when dealing with high-dimensional datasets. Principal Component Analysis (PCA) is a widely used technique for reducing the dimensionality of data while preserving as much variance as possible. PCA works by identifying the directions (principal components) in which the data varies the most. These directions are orthogonal to each other and can be used to project the data into a lower-dimensional space.

To apply PCA, these are the steps:

1. Standardize the data.
2. Compute the covariance matrix.
3. Calculate the eigenvalues and eigenvectors.
4. Sort eigenvectors by eigenvalues in descending order.
5. Select the top k eigenvectors.
6. Project the data onto the new subspace.

Derivation of PCA

If I have an axis w and I want to project a point X onto that axis, I can use the formula:

$$Z = X \cdot w$$

where Z is the projected point in the new space.

W is a $dx1$ vector where d is the original dimensionality of the data.

To best align this axis with the data, we want to maximize the variance of the projected data points. This can be achieved by selecting the eigenvector corresponding to the largest eigenvalue of the covariance matrix of the data.

We will set up an objective function:

$$J = \text{Var}(Z) = \text{Var}(X \cdot w) = \frac{(Xw)^T(Xw)}{n-1} = \frac{w^T X^T X w}{n-1}$$

Note that in the last step, we distribute the transpose so the term Xw^T becomes $w^T X^T$.

Since $X^T X$ is proportional to the covariance matrix of the data, we can rewrite the objective function as:

$$J = w^T \Sigma w$$

where Σ is the covariance matrix of the original data (not the summation symbol or something like that).

We will then use calculus to maximize this objective function with respect to w.

$$\frac{\partial J}{\partial w} = 2\Sigma w = 0$$

We arrived here because the derivative of a quadratic form $w^T A w$ with respect to w is $2Aw$, just as the derivative of x^2 is $2x$. Really the derivative is $\Sigma w + (\Sigma w)^T$ but since Σ is the covariance matrix, it is symmetric, so these two terms are equal.

The next step is to set this equal to zero and solve for w:

$$w = \Sigma^{-1} zeros$$

... which isn't particularly useful. Instead, we will use a constraint that the length of w is equal to 1 (i.e., $w^T w = 1$). This constraint prevents the trivial solution of $w = 0$.

To obtain a more useful solution, we will add in a constraint using a Lagrange multiplier λ:

$$J = w^T \Sigma w - \lambda(w^T w - 1)$$

Note that if $w^T w = 1$, then the term in parentheses is zero (meaning there is no penalty), so we are not changing the value of the objective function.

We want to maximize the variance ($w^T \Sigma w$) while minimizing the penalty ($\lambda(w^T w - 1)$). We will take the derivative of this new objective function with respect to w and set it equal to zero:

$$\frac{\partial J}{\partial w} = 2\Sigma w - 2\lambda w = zeros$$

Pulling out the 2, we have:

$$\Sigma w - \lambda w = 0$$

This can be rearranged to:

$$(\Sigma - \lambda I)w = 0$$

where I is the identity matrix.

We can bring

$$(\Sigma - \lambda I)$$

to the other side by multiplying both sides by the inverse of

$$(\Sigma - \lambda I)$$

:

$$w = 0$$

which is again not useful.

However, the step

$$\Sigma w - 2\lambda w = zeros$$

is significant because it is known as an eigendecomposition problem because it is some matrix times an unknown vector equals a scalar times that same unknown vector. The solutions to this problem are the eigenvalues and eigenvectors of the matrix Σ. The eigenvectors represent the directions of maximum variance in the data, and the corresponding eigenvalues indicate the amount of variance captured by each eigenvector.

Rearrangiing this part, it can be written as:

$$\Sigma w = \lambda w$$

where Σ is the covariance matrix, w is the eigenvector, and λ is the eigenvalue.

so

$$[W, \lambda] = \text{eigenDecomposition}(\Sigma)$$

Where W is a matrix whose columns are the eigenvectors of Σ, and λ is a diagonal matrix containing the corresponding eigenvalues.

Choosing Eigenvectors

The next step is to sort the eigenvalues in descending order and select the top k eigenvectors to form a new matrix W_k. The original data can then be projected onto this new subspace using:

$$Z = XW_k$$

Where Z is the transformed data in the lower-dimensional space as long as k is less than the original dimensionality. So if $k = 3$ then we are projecting the data down to 3 dimensions.

Another approach could be to capture some percentage a of the variance. This can be expressed as follows:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq a$$

Where d is the original dimensionality of the data, and λ_i are the eigenvalues. This means we select the smallest k such that the ratio of the sum of the top k eigenvalues to the sum of all eigenvalues is at least a.

Note that the denominator here represents the sum of all the eigenvalues and therefore the total variance in the data. The numerator represents the variance captured by the top k principal components. By selecting k such that this ratio meets or exceeds a, we ensure that we retain a specified proportion of the total variance in the reduced-dimensional representation of the data. Typically values for a are around 0.90 or 0.95 meaning we want to capture 90% or 95% of the variance in the data.

Now the matrix W should be of size dxk. We can project the original data X (of size nxd) onto this new subspace to get the reduced data Z (of size nxk):

Example

Consider the following dataset with 2 features:

$$X = \begin{bmatrix} 2.5 & 2.4 \\ 0.5 & 0.7 \\ 2.2 & 2.9 \end{bmatrix}$$

1. Standardize the data (mean center):

$$\text{Mean} = \begin{bmatrix} 1.733 & 1.333 \end{bmatrix}$$

$$X_{centered} = X - \text{Mean} = \begin{bmatrix} 0.767 & 1.067 \\ -1.233 & -0.633 \\ 0.467 & 1.567 \end{bmatrix}$$

2. Compute the covariance matrix:

$$\Sigma = \frac{1}{n-1} X_{centered}^T X_{centered} = \begin{bmatrix} 1.163 & 1.165 \\ 1.165 & 1.333 \end{bmatrix}$$

3. Calculate the eigenvalues and eigenvectors of the covariance matrix (we will not need to know how to do this for the exam): Eigenvalues: $\lambda_1 = 2.414$, $\lambda_2 = 0.082$
Eigenvectors:

$$w_1 = \begin{bmatrix} 0.677 \\ 0.736 \end{bmatrix}, \quad w_2 = \begin{bmatrix} -0.736 \\ 0.677 \end{bmatrix}$$

4. Sort eigenvectors by eigenvalues in descending order:

$$W = \begin{bmatrix} 0.677 & -0.736 \\ 0.736 & 0.677 \end{bmatrix}$$

5. Select the top k eigenvectors (for $k = 1$):

$$W_k = \begin{bmatrix} 0.677 \\ 0.736 \end{bmatrix}$$

6. Project the data onto the new subspace:

$$Z = X_{centered} W_k = \begin{bmatrix} 1.827 \\ -1.777 \\ 2.992 \end{bmatrix}$$