Weekly Report 8

Utkarsh Patel
18EC30048

Topics Covered

1 PRINCIPAL COMPONENTS ANALYSIS

It is a linear projection method. In this approach, we seek to find a mapping from the $d-$dimensional space to a $k-$dimensional space ($k < d$). This mapping must ensure that there is a minimum loss of information. Given a vector x and directional vector w, the projection of x on w is given as

$$z = w^T x, z \in \mathbb{R}, w, x \in \mathbb{R}^d$$

PCA is unsupervised learning algorithm. In this approach, the only thing we care about is *variance*. The principal component, denoted as w_1 is defined as direction for which the projection $z_1 = w_1^T x$ has maximum variance. Assuming x to be a random vector (composed of random variables) i.e. $x = (x_1, x_2, \dots, x_d)$. We define covariance matrix of x to be Σ, where $\Sigma[i,j] = cov(x_i, x_j)$. It is known fact that covariance matrix of a random vector is a positive semi-definite matrix i.e. all the eigen values of covariance matrix are non-zero. In this setting, the variance of projection z_1 is given as $var(z_1) = w_1^T \Sigma w_1$. To find the principal component, we have to maximize this variance under constraint that $\|w_1\| = 1$. We transform this into Lagrange problem.

$$w_1 = \underset{w}{\operatorname{argmax}}(w^T \Sigma w - \alpha(w^T w - 1))$$

After differentiating and putting the derivative to zero, we get

$$\Sigma w_1 = \alpha w_1$$
$$var(z_1) = w_1^T \Sigma w_1 = \alpha w_1^T w_1 = \alpha$$

The equations obtained implies that the principal component is an eigenvector of the covariance matrix Σ corresponding to eigenvalue α, which is also the variance of the projection. Hence, as principal component gives the direction of maximum variance, it can be concluded that the eigenvector of covariance matrix Σ corresponding to largest eigenvalue $\lambda_1 = \alpha$ is the principal component.

The second principal component w_2 must be the direction of maximum variance given that it is orthogonal to the first principal component w_1. Therefore, transforming this to Lagrange problem we have,

$$w_2 = \underset{w}{\operatorname{argmax}}\left(w^T \Sigma w - \alpha(w^T w - 1) - \beta(w^T w_1 - 0)\right)$$

Differentiating wrt to w and putting the derivative to zero, we get

$$2\Sigma w_2 - 2\alpha w_2 - \beta w_1 = 0$$

Pre-multiplying with w_1^T, we have,

$$2w_1^T \Sigma w_2 - 2\alpha w_1^T w_2 - \beta w_1^T w_1 = 0$$

Now, $w_1^T w_2 = 0$ as both the directions are orthogonal. Consider $s = w_1^T \Sigma w_2$. As s is a scalar, we have

$$w_1^T \Sigma w_2 = (w_1^T \Sigma w_2)^T = w_2^T \Sigma w_1 = \lambda_1 w_2^T w_1 = 0$$

This implies $\beta = 0$. Hence, we have

$$\Sigma w_2 = \alpha w_2$$
$$var(z_2) = w_2^T \Sigma w_2 = \alpha w_2^T w_2 = \alpha$$

Therefore, it can be concluded that the second principal component is the eigenvector of covariance matrix Σ corresponding to second largest eigenvalue $\lambda_2 = \alpha$. It is guaranteed that w_1 and w_2 will be orthogonal as they are eigenvectors of two distinct eigenvalue of a symmetric matrix Σ. Similarly, we can find other principal components. If the input dimensions are uncorrelated already, then the rank of covariance matrix will be d. Therefore, we will have d distinct eigenvalues hence eigenvectors. In such a scenario, using PCA is wasteful. However, if it is known that the input dimensions are correlated to an extent (like in image and speech processing tasks), then we will have only k ($< d$) principal components and thus a reduction in dimensionality is achieved. We can also choose number of principal components we want depending on how much variance we want to lose. Suppose we have k non-zero eigenvalues (sorted in decreasing order), if we select the first m eigenvalues (hence first m principal components) only, then fraction of variance retained is given as

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_m}{\lambda_1 + \lambda_2 + \cdots + \lambda_m + \cdots + \lambda_k}$$

2 LINEAR DISCRIMINANT ANALYSIS

It is a linear projection, supervised method for classification problem. For sake of simplicity, let's have only two classes C_1 and C_2. We are interested in the direction w such that after projection in this direction, the instances of the two classes are as well separated as possible. In this case, we have a dimensionality reduction from d to 1. Let w be the required direction. Then, projection of instance x is given as $z = w^T x$. Let μ_1, m_1 be mean of instances belonging to class C_1 before and after the projection and μ_2, m_2 for class C_2. Let the training set be given as $\{x^{(i)}, y^{(i)}\}$ where $y^{(i)} = 1$ if $x^{(i)}$ belongs to class C_1 and $y^{(i)} = 0$, if $x^{(i)}$ belongs to C_2.

$$m_1 = \frac{\sum_{i=1}^N w^T x^{(i)} y^{(i)}}{\sum_{i=1}^N y^{(i)}} = w^T \mu_1$$
$$m_2 = \frac{\sum_{i=1}^N w^T x^{(i)} (1 - y^{(i)})}{\sum_{i=1}^N (1 - y^{(i)})} = w^T \mu_2$$

The scatter of sample (representing the spread) from C_1 and C_2 after the projection is defined as

$$s_1 = \sum_{i=1}^N \left((w^T x^{(i)} - m_1)^2 \cdot y^{(i)} \right) = \sum_{i=1}^N \left(w^T (x^{(i)} - m_1)(x^{(i)} - m_1)^T y^{(i)} w \right) = w^T S_1 w$$
$$s_2 = \sum_{i=1}^N \left((w^T x^{(i)} - m_2)^2 \cdot (1 - y^{(i)}) \right) = \sum_{i=1}^N \left(w^T (x^{(i)} - m_1)(x^{(i)} - m_1)^T (1 - y^{(i)}) w \right) = w^T S_2 w$$

Here, S_1 and S_2 are the within-class scatter matrix for C_1, C_2 respectively. Now, several implementations of *LDA* exists depending on how we define the criterion of being well-separated. One variation is the *Fisher's linear discriminant* in which we expect mean of the classes to be as far as possible and scatter of each class to be as less as possible. Therefore, we maximize the following function,

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

We have, $(m_1 - m_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = w^T S_B w$, where $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ is the between-class scatter matrix. For the denominator part, we have $s_1^2 + s_2^2 = w^T S_1 w + w^T S_2 w = w^T (S_1 + S_2) w = w^T S_W w$. Hence, the objective function reduces to

$$J(w) = \frac{w^T S_B w}{w^T S_W w} = \frac{|w^T(\mu_1 - \mu_2)|^2}{w^T S_W w}$$

Differentiating wrt w and putting the derivative to 0, we get $w = S_W^{-1}(\mu_1 - \mu_2)$.

Novel Ideas

In my research project on 'Classification of Autism Disorder using Deep Learning', I used both *PCA* and *LDA*: *PCA* for identifying important modules of the brain and *LDA* for classification purpose. In that, we got test accuracy of about 59.43% which was not the best (best test accuracy was observed for *SVM* classifier). I think this is because in LDA, the dimensions are reduced to unity and hence model becomes quite simpler. You can see the project at https://github.com/utkarsh512/fMRI-classification-of-ASD/