2.10 FEATURE ENGINEERING AND DIMENSIONALITY REDUCTION TECHNIQUES

Features are attributes. Feature engineering is about determining the subset of features that form an important part of the input that improves the performance of the model, be it classification or any other model in machine learning.

Feature engineering deals with two problems – Feature Transformation and Feature Selection. Feature transformation is extraction of features and creating new features that may be helpful in increasing performance. For example, the height and weight may give a new attribute called Body Mass Index (BMI).

Feature subset selection is another important aspect of feature engineering that focuses on selection of features to reduce the time but not at the cost of reliability.

The subset selection reduces the dataset size by removing irrelevant features and constructs a minimum set of attributes for machine learning. If the dataset has n attributes, then time complexity is extremely high as n dimensions need to be processed for the given dataset. For n attributes, there are 2^n possible subsets. If the value of n is high, the problem becomes intractable. This is called 'curse of dimensionality'. Since, as the number of dimensions increases, the time complexity increases. The remedy is that some of the components that do not contribute much can be deleted. This results in the reduction of dimensionality. Choosing optimal attributes becomes a graph search problem. Typically, the feature subset selection problem uses greedy approach by looking for the best choice at the time using locally optimal choice while hoping that it would lead to global optimal solutions.

The features can be removed based on two aspects:

1. Feature relevancy – Some features contribute more for classification than other features. For example, a mole on the face can help in face detection than common features like nose. In simple words, the features should be relevant. The relevancy of the features can be determined based on information measures such as mutual information, correlation-based features like correlation coefficient and distance measures. Distance measures are discussed in Chapter 13 of this book.
2. Feature redundancy – Some features are redundant. For example, when a database table has a field called Date of birth, then age field is not relevant as age can be computed easily from date of birth. This helps in removing the column age that leads to reduction of dimension one.

So, the procedure is:

1. Generate all possible subsets
2. Evaluate the subsets and model performance
3. Evaluate the results for optimal feature selection

Filter-based selection uses statistical measures for assessing features. In this approach, no learning algorithm is used. Correlation and information gain measures like mutual information and entropy are all examples of this approach.

Wrapper-based methods use classifiers to identify the best features. These are selected and evaluated by the learning algorithms. This procedure is computationally intensive but has superior performance.

Let us discuss some of the important algorithms that fall under this category.

2.10.1 Stepwise Forward Selection

This procedure starts with an empty set of attributes. Every time, an attribute is tested for statistical significance for best quality and is added to the reduced set. This process is continued till a good reduced set of attributes is obtained.

2.10.2 Stepwise Backward Elimination

This procedure starts with a complete set of attributes. At every stage, the procedure removes the worst attribute from the set, leading to the reduced set.

Combined Approach Both forward and reverse methods can be combined so that the procedure can add the best attribute and remove the worst attribute.

2.10.3 Principal Component Analysis

The idea of the principal component analysis (PCA) or KL transform is to transform a given set of measurements to a new set of features so that the features exhibit high information packing properties. This leads to a reduced and compact set of features. Basically, this elimination is made possible because of the information redundancies. This compact representation is of a reduced dimension.

Consider a group of random vectors of the form:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The *mean vector* of the set of random vectors is defined as:

$$m_x = E\{x\}$$

The operator E refers to the expected value of the population. This is calculated theoretically using the probability density functions (PDF) of the elements x_i and the joint probability density functions between the elements x_i and x_j. From this, the covariance matrix can be calculated as:

$$C = E\{(x - m_x)(x - m_x)^T\} \tag{2.52}$$

For M random vectors, when M is large enough, the mean vector and covariance matrix can be approximately calculated as:

$$m_x = \frac{1}{M}\sum_{k=1}^{M} x_k \tag{2.53}$$

$$A = \frac{1}{M}\sum_{k=1}^{M} x_k x_k^T - m_x m_x^T \tag{2.54}$$

This covariance matrix is real and symmetric. If e_i and λ_i (where, $i = 1, 2, \ldots, n$) be the set of eigen vectors and corresponding eigen values of the covariance matrix, the eigen values can be arranged in a descending order so that $\lambda_i \geq \lambda_{i+1}$ for $i = 1, 2, \ldots, n - 1$. The corresponding eigen vectors are calculated. Based on this, the transform kernel is constructed. Let the transform kernel be A. Then, the matrix rows are formed from the eigen vectors of the covariance matrix.

The mapping of the vectors x to y using the transformation can now be described as:

$$y = A(x - m_x) \tag{2.55}$$

This transform is also called as Karhunen-Loeve or Hoteling transform. The original vector x can now be reconstructed as follows:

$$x = A^T y + m_x \tag{2.56}$$

The goal of PCA is to reduce the set of attributes to a newer, smaller set that captures the variance of the data. The variance is captured by fewer components, which would give the same result as the original, with all the attributes. Here, instead of using all the eigen vectors of the covariance matrix, only a small set that exhibits the variance can be used. By using a small set, the KL transform can achieve maximum compression of the available data.

If K largest eigen values are used, the recovered information would be:

$$x = A_K^T y + m_x \tag{2.57}$$

The advantages of PCA are immense. It reduces the attribute list by eliminating all irrelevant attributes. The PCA algorithm is as follows:

1. The target dataset x is obtained
2. The mean is subtracted from the dataset. Let the mean be m. Thus, the adjusted dataset is $X - m$. The objective of this process is to transform the dataset with zero mean.
3. The covariance of dataset x is obtained. Let it be C.

4. Eigen values and eigen vectors of the covariance matrix are calculated.
5. The eigen vector of the highest eigen value is the principal component of the dataset. The eigen values are arranged in a descending order. The feature vector is formed with these eigen vectors in its columns.

   Feature vector = {eigen vector$_1$, eigen vector$_2$, ... , eigen vector$_n$}
6. Obtain the transpose of feature vector. Let it be A.
7. PCA transform is $y = A \times (x - m)$, where x is the input dataset, m is the mean, and A is the transpose of the feature vector.

The original data can be retrieved using the formula given below:

$$\text{Original data } (f) = \{(A)^{-1} \times y\} + m \tag{2.58}$$

$$= \{(A)^{T} \times y\} + m \tag{2.59}$$

The new data is a dimensionally reduced matrix that represents the original data. Therefore, PCA is effective in removing the attributes that do not contribute. If the original data is required, it can be obtained with no loss of information. Scree plot is a visualization technique to visualize the principal components or variables that play a more important role as compared to other attributes. Scree plot is a visualization technique to visualize the principal components visually. For a randomly selected dataset of 246 attributes, PCA is applied and its scree plot is shown in Figure 2.15. The scree plot indicates that only 6 out of 246 attributes are important.

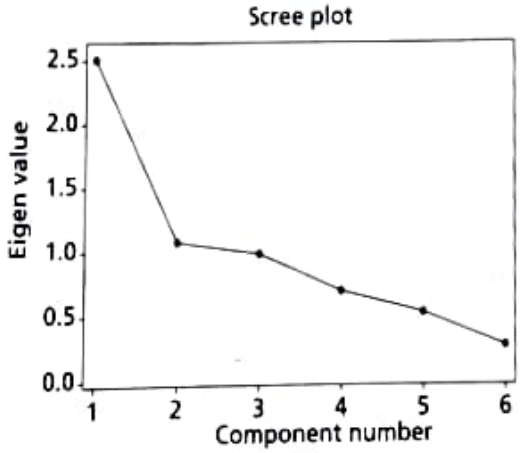

Figure 2.15: Scree Plot

From Figure 2.15, one can infer the relevance of the attributes. The scree plot indicates that the first attribute is more important than all other attributes.

Example 2.12: Let the data points be $\begin{pmatrix}2\\6\end{pmatrix}$ and $\begin{pmatrix}1\\7\end{pmatrix}$. Apply PCA and find the transformed data. Again, apply the inverse and prove that PCA works.

Solution: One can combine two vectors into a matrix as follows:

The mean vector can be computed as Eq. (2.53) as follows:

$$\mu = \begin{pmatrix}\frac{2+1}{2}\\ \frac{6+7}{2}\end{pmatrix} = \begin{pmatrix}1.5\\6.5\end{pmatrix}$$

As part of PCA, the mean must be subtracted from the data to get the adjusted data:

$$x_1 = \begin{pmatrix} 2 - 1.5 \\ 6 - 6.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$$

$$x_2 = \begin{pmatrix} 1 - 1.5 \\ 7 - 6.5 \end{pmatrix} = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}$$

One can find the covariance for these data vectors. The covariance can be obtained using Eq. (2.54):

$$m_1 = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix}$$

$$m_2 = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix} \begin{pmatrix} -0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.25 \\ -0.25 & 0.25 \end{pmatrix}$$

The final covariance matrix is obtained by adding these two matrices as:

$$C = \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

The eigen values and eigen vectors of matrix C can be obtained (left as an exercise) as $\lambda_1 = 1$, $\lambda_2 = 0$. The eigen vectors are $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The matrix A can be obtained by packing the eigen vector of these eigen values (after sorting it) of matrix C. For this problem, $A = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$. The transpose of A, $A^T = \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$ is also the same matrix as it is an orthogonal matrix. The matrix can be normalized by diving each elements of the vector, by the norm of the vector to get:

$$A = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

One can check that the PCA matrix A is orthogonal. A matrix is orthogonal is $A^{-1} = A$ and $AA^{-1} = I$.

$$AA^T = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The transformed matrix y using Eq. (2.55) is given as:

$$y = A \times (x - m)$$

Recollect that $(x-m)$ is the adjusted matrix.

$$y = A(x-m) = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \left(\text{for convenience } 0.5 = \frac{1}{2} \right)$$

$$= \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix}$$

One can check the original matrix can be retrieved from this matrix as:

$$\{(A)^T \times y\} + m$$

$$x = A^T y + m = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1.5 \\ 6.5 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} + \begin{pmatrix} 1.5 \\ 6.5 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 6 & 7 \end{pmatrix}$$

Therefore, one can infer the original is obtained without any loss of information.

2.10.4 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is also a feature reduction technique like PCA. The focus of LDA is to project higher dimension data to a line (lower dimension data). LDA is also used to classify the data. Let there be two classes, c_1 and c_2. Let μ_1 and μ_2 be the mean of the patterns of two classes. The mean of the class c_1 and c_2 can be computed as:

$$\mu_1 = \frac{1}{N_1} \sum_{x_i \in c_1}^{n} x_i \text{ and } \mu_2 = \frac{1}{N_2} \sum_{x_i \in c_2}^{n} x_i$$

The aim of LDA is to optimize the function:

$$J(V) = \frac{V^T \sigma_B V}{V^T \sigma_W V} \tag{2.60}$$

where, V is the linear projection and σ_B and σ_W are class scatter matrix and within scatter matrix, respectively. For the two-class problem, these matrices are given as:

$$\sigma_B = N_1(\mu_1 - \mu)(\mu_1 - \mu)^T + N_2(\mu_2 - \mu)(\mu_2 - \mu)^T \tag{2.61}$$

$$\sigma_W = \sum_{x_i \in c_1} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{x_i \in c_2} (x_i - \mu_2)(x_i - \mu_2)^T \tag{2.62}$$

The maximization of $J(V)$ should satisfy the equation:

$$\sigma_B V = \lambda \sigma_W V \text{ or } \sigma_W^{-1} \sigma_B V = \lambda V \tag{2.63}$$

As $\sigma_B V$ is always in the direction of $(\mu_1 - \mu_2)$, V can be given as:

$$V = \sigma_W^{-1}(\mu_1 - \mu_2) \tag{2.64}$$

Let $V = \{v_1, v_2, \cdots, v_d\}$ be the generalized eigen vectors of σ_B and σ_W, where, d is the largest eigen values as in PCA. The transformation of x is then given as:

$$y = V_d^T x \tag{2.65}$$

Like in PCA, the largest eigen values can be retained to have projections.

2.10.5 Singular Value Decomposition

Singular Value Decomposition (SVD) is another useful decomposition technique. Let A be the matrix, then the matrix A can be decomposed as:

$$A = USV^T \tag{2.66}$$

Here, A is the given matrix of dimension $m \times n$, U is the orthogonal matrix whose dimension is $m \times n$, S is the diagonal matrix of dimension $n \times n$, and V is the orthogonal matrix. The procedure for finding decomposition matrix is given as follows:

1. For a given matrix, find AA^T
2. Find eigen values of AA^T
3. Sort the eigen values in a descending order. Pack the eigen vectors as a matrix U.
4. Arrange the square root of the eigen values in diagonal. This matrix is diagonal matrix, S.
5. Find eigen values and eigen vectors for A^TA. Find the eigen value and pack the eigen vector as a matrix called V.

Thus, $A = USV^T$. Here, U and V are orthogonal matrices. The columns of U and V are left and right singular values, respectively. SVD is useful in compression, as one can decide to retain only a certain component instead of the original matrix A as:

$$a_{ij} = \sum_{k=1}^{n} u_{ik} s_k v_{jk} \tag{2.67}$$

Based on the choice of retention, the compression can be controlled.

Example 2.13: Find SVD of the matrix:

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix}$$

Solution: The first step is to compute:

$$AA^T = \begin{pmatrix} 1 & 2 \\ 4 & 9 \end{pmatrix} \begin{pmatrix} 1 & 4 \\ 2 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 22 \\ 22 & 97 \end{pmatrix}$$

The eigen value and eigen vector of this matrix can be calculated to get U. The eigen values of this matrix are 0.0098 and 101.9902.

The eigen vectors of this matrix are:

$$u_1 = \begin{pmatrix} 0.2268 \\ 1 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} -4.4086 \\ 1 \end{pmatrix}$$

These vectors are normalized to get the vectors respectively as:

$$u_1 = \begin{pmatrix} 0.2212 \\ 0.9752 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} -0.9752 \\ 0.2212 \end{pmatrix}$$

The matrix U can be obtained by concatenating the above vector as:

$$U = [u_1, u_2] = \begin{pmatrix} 0.2212 & -0.9752 \\ 0.9752 & 0.2212 \end{pmatrix}$$

The matrix V can be obtained by finding A^TA. It is $\begin{pmatrix} 17 & 38 \\ 38 & 85 \end{pmatrix}$. The eigen values are 0.0098 and 101.9902. The eigen vectors can be found as follows:

$$v_1 = \begin{pmatrix} 0.447 \\ 1 \end{pmatrix} \text{ when } \lambda = 101.99$$

$$v_2 = \begin{pmatrix} -2.236 \\ 1 \end{pmatrix} \text{ when } \lambda = 0.0098$$

The above can be normalized as follows:

$$v_1 = \begin{pmatrix} 0.4082 \\ 0.9129 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} -0.9129 \\ 0.4082 \end{pmatrix}$$

The matrix V can be obtained by concatenating the above vector as:

$$V = [v_1 \ v_2] = \begin{pmatrix} 0.4081 & -0.9129 \\ 0.9129 & 0.4082 \end{pmatrix}$$

The matrix S can be found as the diagonal matrix as:

$$S = \begin{pmatrix} \sqrt{101.9902} & 0 \\ 0 & \sqrt{0.0098} \end{pmatrix} = \begin{pmatrix} 10.099 & 0 \\ 0 & 0.099 \end{pmatrix}$$

Therefore, the matrix decomposition $A = U\,SV^T$ is complete.

The main advantage of SVD is compression. A matrix, say an image, can be decomposed and selectively only certain components can be retained by making all other elements zero. This reduces the contents of image while retaining the quality of the image. SVD is useful in data reduction too.