Dimensionality Reduction

Sourav Karmakar

souravkarmakar29@gmail.com

Dimensionality Reduction

Problem with very high dimensional data: Curse of Dimensionality

- Data at very high dimensional space become sparse.
- For highly sparse data Machine Learning algorithm doesn't give satisfactory results always.
- This is simply because the available training dataset might not account for every possible combinations of the features. Hence, ML algorithms struggles to find any meaningful relationship among the features and the target variables.
- For example consider a dataset which has D many categorical features each with k many possible values. Then the number of possible combination of the datapoints is k^D, which grows exponentially as the dimension increases. Our training dataset may not contain all sorts of possible combinations.

Work around of this problem

Dimensionality Reduction

To find a suitable representation of the data in lower dimensional space which will eventually help us for machine learning (classification/ clustering/ regression/ data-visualization).

Dimensionality Reduction

- There are several types of dimensionality reduction techniques.
- However, all those techniques can be broadly classified into two categories:
    - Feature Selection techniques
    - Feature Extraction techniques

Feature Selection techniques

We select few features based on certain criteria and thus get rid of redundant or less important features to make our dataset less bulky. Note that feature selection depends on the dataset as well as on the task we want to perform with the dataset.

Feature Extraction techniques

We project the original dataset which belongs to a high dimensional space to a low dimensional space such that the inherent information inside the data is not lost. This projection could be either linear and non-linear based on which we have linear or non-linear feature extraction techniques.

Dimensionality Reduction

A very simple example of Feature Selection:

A school student in a particular school can have following features:

- x_1 = Name of the student
- x_2 = Contact details of the student
- x_3 = Class
- x_4 = Roll number
- x_5 = Marks in mathematics
- x_6 = Marks in language group
- x_7 = Marks in social science group
- x_8 = Marks in natural science group
- x_9 = Height
- x_{10} = Body weight
- x_{11} = Blood group

etc.

- Now let the job is to find suitable set of students for Science Olympiad. Which features play the important role in selecting the students for Science Olympiad?

  Obviously in this case the most important features are: Marks in mathematics and marks in natural science group.

- Now let us want to select the students for a sports tournament

  In this case we have to select the students based on the height, weight, fitness, and skill in that particular sports.

Each student can be represented by a vector $\vec{x} = [x_1, x_2, x_3, \dots]^T$

Based on the problem in hand we can discard some of the features and consider others.

Feature Selection

Objective:

To obtain a set of features k out of n many features in the dataset ($k < n$) such that, the selected features are best suitable for a required task T. Now the actual value of k depends on many things such as: the task T (if it's for visualization then $k = 2\ or\ 3$ is good enough), the amount of the data that machine can handle, number of available training data points so that the dataset becomes less sparse in the reduced feature space etc.

How to select the features:

- Use domain knowledge and experience to select set of best features suitable for particular task.
- **Exhaustive Search:** Select the best possible set of k many features out of all possible combinations of features. This method might work well if the number of features in the dataset is not very high. Otherwise this method doesn't scale well if the number of features in the dataset is too high.
- **Heuristic Search:** In this case we keep on adding features which will provide best value of some criterion function. We stop the algorithm till the required number of features are selected.

Other search strategies such as Randomized Search are also used for feature selection.

Feature Extraction

- Let the original datapoints X belong to a higher dimensional space of dimension D. i.e. $X \subseteq R^D$.
  We want to reduce the dimensionality of the dataset into d $(d < D)$.

- In feature extraction we shall project the dataset from dimension D to d, using some function $f : R^D \rightarrow R^d$ such that the information content inside the dataset is preserved significantly.

- Here we are not selecting or discarding any particular feature. The whole idea is to project the dataset into some smaller dimensional subspace, such that the inter-relationship of the datapoints are preserved.

  Let X is our original dataset in dimension D and $\vec{x} \in X$ is a datapoint.

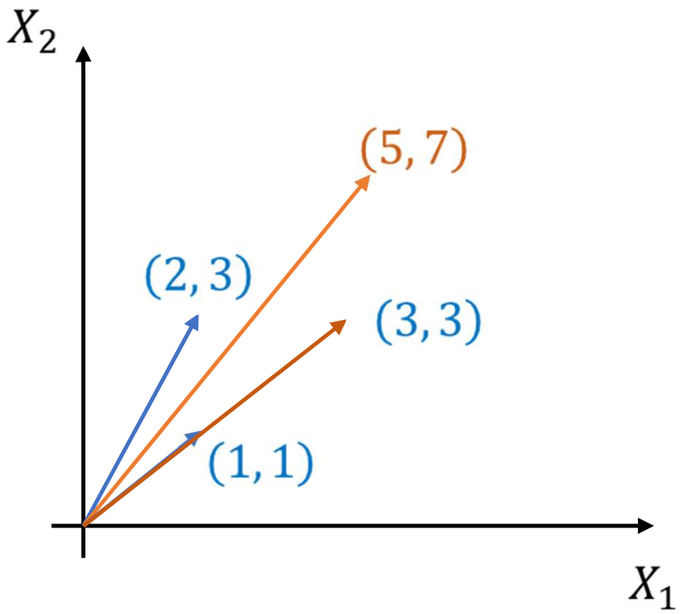
  Let, in the transformed space of dimension d, $\vec{x}'$ is the corresponding datapoint. i.e. $f(\vec{x}) = \vec{x}'$.

  If $\vec{x}_1$ & $\vec{x}_2$ are two neighbouring points in original dataset X. Then in the transformed space $\vec{x}_1'$ & $\vec{x}_2'$ should be also neighbours.

- There are various Feature Extraction techniques. Few techniques employ linear transformation / projection of the dataset. For example: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Some techniques involve non-linear transformation of the dataset. For example: Locally Linear Embedding (LLE) , Kernel PCA etc.

Eigenvalues & Eigenvectors

Matrix Vector Multiplication



- Consider the vector $\vec{x}_1 = [2\,, 3]^T$
- If the matrix $A = \begin{bmatrix} 4 & -1 \\ 2 & 1 \end{bmatrix}$ is post multiplied with $\vec{x}_1$, then we obtain another vector $\vec{x}_1'$.

$$\vec{x}_1' = A\vec{x}_1 \Rightarrow \vec{x}_1' = \begin{bmatrix} 4 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

- So we can say that multiplying a vector with a matrix changes the direction and magnitude of the vector.

- However for a given square matrix there are some vectors whose direction remain unchanged after multiplying with the matrix.
- Consider the vector $\vec{x}_2 = [1\,, 1]^T$. If we multiply A with $\vec{x}_2$ then, $\vec{x}_2' = \begin{bmatrix} 4 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- Notice that only the magnitude changed but direction remain unchanged. i.e. the transformed vector $\vec{x}_2'$ is a scalar multiplication of the original vector $\vec{x}_2$.

Eigenvalues And Eigenvectors

- For a given square matrix A if we find a vector $\vec{v}$. Such that multiplying A with $\vec{v}$ doesn't change the direction of $\vec{v}$ but scales it, then that vector $\vec{v}$ is known as the **eigenvector** of A and the scalar value by which it will be scaled after being multiplied with A is known as the corresponding **eigenvalue**.

- Mathematically, $if\ \ A\vec{v} = \lambda\vec{v}\ , then\ \lambda\ is\ called\ the\ eigenvalue\ and\ \vec{v}\ the\ eigenvector\ of\ matrix\ A$

- $A\vec{v} = \lambda\vec{v}\ \Rightarrow (A - \lambda I)\vec{v} = 0$, here I is the identity matrix of the same order of matrix A.

- For non-trivial solution of the eigenvectors, $|A\ - \lambda I| = 0$

- The above equation is a polynomial of λ of degree same as the order of the matrix A. This equation is also known as the characteristic equation of matrix A.

- We can calculate the eigenvalues by solving the characteristic equation.

- Putting the value of the eigenvalue one can calculate the corresponding eigenvectors.

Lagrange's Multiplier

Constrained Optimization

The unconstrained optimization problem can be stated as: **Find the extreme value of** $y = f(x)$

And to solve the problem we find the first derivative of y with respect to x and set that to zero. $\frac{dy}{dx} = 0$

In general a dependent variable could be functions of several independent variable. Then the unconstrained optimization problem is stated as: **Find the extreme value of** $y = f(x_1, x_2, x_3, \dots, x_n)$

And to solve the problem we find the first derivative of y with respect to ***all independent variables*** and set those to zero.

$$\frac{\partial y}{\partial x_1} = 0 \quad \frac{\partial y}{\partial x_2} = 0 \quad \frac{\partial y}{\partial x_3} = 0 \quad \bullet \bullet \bullet \quad \frac{\partial y}{\partial x_n} = 0$$

However in real life we come across with the situations where we have to optimize the function subject to certain conditions (called the constraints).

Find the extreme value of $y = f(x)$ **,**
subjected to the condition $g(x) = 0$

This type of problems are known as constrained optimization problem.

Lagrange's Multiplier

Find the extreme value of $y = f(x)$,
subjected to the condition $g(x) = 0$

To solve the constrained optimization problem we formulate the **Lagrangian** as following:

$$L(x, \lambda) = f(x) - \lambda g(x)$$

Where, λ is a dummy variable known as **Lagrange's Multiplier.** Then we take derivative of the Lagrangian and set that to zero to find the optimal value.

$$\frac{\partial L}{\partial x} = 0 \rightarrow \frac{\partial f}{\partial x} - \lambda \frac{\partial g}{\partial x} = 0 \rightarrow \frac{\partial f}{\partial x} = \lambda \frac{\partial g}{\partial x}$$

In general, there could be ***n*** many independent variables and ***m*** many constraints.

Find the extreme value of $y = f(x_1 , x_2 , x_3 , \dots , x_n)$,
subjected to: $g_1(x_1 , x_2 , x_3 , \dots , x_n) = 0 ; g_2(x_1 , x_2 , x_3 , \dots , x_n) = 0 ; \dots ; g_m(x_1 , x_2 , x_3 , \dots , x_n) = 0$

Then the **Lagrangian** is formulated as following.

$$L(x_1, x_2, \dots, x_n, \lambda_1, \lambda_2, \dots, \lambda_m) = f - \lambda_1 \, g_1 - \lambda_2 \, g_2 - \dots - \lambda_m \, g_m$$

Lagrange's Multiplier

Example: Maximize $y = 5x_1x_2$, subjected to $2x_1 + x_2 = 100$

Step-1: Formulate the Lagrangian. $L(x_1, x_2, \lambda) = 5x_1x_2 - \lambda\,(2x_1+x_2-100)$

Step-2: Take the partial derivatives of Lagrangian wrt x_1, x_2 and set them to zero.

$$\frac{\partial L}{\partial x_1} = 0 \rightarrow 5x_2 - 2\lambda = 0 \quad ---(1) \qquad \frac{\partial L}{\partial x_2} = 0 \rightarrow 5x_1 - \lambda = 0 \quad ---(2)$$

Step-3: Along with these we make use of the constrained equation. $2x_1 + x_2 - 100 = 0 \quad ---(3)$

Step-4: Solving these three equations we get:

$$\lambda = 125$$

$$x_1 = 25$$

$$x_2 = 50$$

Thank You