Project DR06

Elhamsadat Hejazi (110084548) and Mohadeseh Heidarpour Roshan (110040890)
Email : {hejadi; heidarpm} @uwindsor.ca

Real-world data for many machine-learning or data-mining projects are usually high-dimensional or consist of a large number of features which is hard to be managed [1]. To avoid overfitting, wasting time and storage for saving data, and to improve the performance of the machine-learning model, there are some algorithms that scientists use to achieve low-dimensional data out of the original dataset provided that it retains most of the important attributes, called intrinsic, of the main dataset [2]. This transformation is called Dimensionality Reduction (DR).

The DR algorithms listed in the project DR06 are Probabilistic Principal Component Analysis (PPCA), Landmark ISOMAP [LIM], and Local Linear Embedding [LLE] that will be discussed in the following.

A. *Probabilistic PCA [PPCA]*

The major difference between the PPCA technique and the well-known PCA technique is that it deals with the missing data and in addition to it, the probabilistic approach enables the model to analyze a wide range of possible extensions in the future [3].

Algorithm 1 PPCA

Assume a centered dataset with $\boldsymbol{d}$ dimensions as $Y = [y_1, \dots, y_n]^T$ where $\boldsymbol{n}$ is the number of observations. Considering Gaussian noise, the relationship between this dataset and the embedded $\boldsymbol{q}$-dimensional latent-space data (i.e., X_n) will be as follows:

$$y_n = W_{X_n} + n \tag{1}$$

Where $\boldsymbol{y_n}$ is the set of a specific observation, $\boldsymbol{W}$ is a $\boldsymbol{d}\times\boldsymbol{q}$ matrix ($\boldsymbol{q}<\boldsymbol{d}$), and the noise is in Gaussian distribution with the mean of 0 and COV of $\boldsymbol{\beta^{-1}I}$.

Therefore, the likelihood will be calculated as follows:

$$P(y_n|X_n, W, \beta) = N\left(y_n|W_{X_n}, \beta^{-1}I\right) \tag{2}$$

Where $\boldsymbol{N}$ defines as a normal distribution function of multivariate of y_n with the mean of W_{X_n} and COV of $\beta^{-1}I$.

The next step would be defining the marginal likelihood which is as follows:

$$p(y_n|W, b) = \int p(y_n|X_n, W, b)p(X_n)dX_n \tag{3}$$

That necessitates identifying a prior distribution over X_n that for PPCA this prior distribution (i.e., $p(X_n)$) has a mean of 0, Gaussian distribution, and a unit COV which turns the equation (3) into the following:

$$p(y_n|W, \beta) = N(y_n|0, WW^T, \beta^{-1}I) \tag{4}$$

Then if the data be considered independent and identically distributed, the product of individual probabilities will result in the likelihood of the whole set:

$$p(Y|W, \beta) = \prod_{n=1}^{N} p(y_n|W, \beta) \tag{5}$$

The solution of $\boldsymbol{W}$ will be obtained by maximizing equation (5). Since the result of this model is calculated by maximizing the probability, it can be called the probabilistic PCA.

B. *Landmark Isomap [LIM]*

One of the well-established non-linear dimensionality reduction techniques is called Isoma. However, one of the issues of this technique is its scalability that has been improved with another technique named Landmark-Isomap (L-Isomap).

Algorithm 2 LIM

In the Isomap, the algorithm has input data space $X \in R^{D\times N}$ with all N data samples and D dimensions that it tries to embed the input samples into a lower-dimensional space $X \in R^{d\times N}$, In such a way that the geodetic distance between the points is maintained. In this algorithm, the starting point is to create an indirect neighborhood graph $G = (V, E)$, where each node V$i \in V$, corresponding to the point $x_i \in X$, is connected with its k-nearest neighbor, In such a way that each point node $v_i \in V$ is connected to the point $x_i \in X$ by its k-nearest neighbor and also each edge $e_{ij} \in E$ is appointed to weight D_{ij} that shows the Euclidean distance between points x_i and x_j. In the second, the Isomap algorithm calculates the shortest path between every two points in a graph to estimate the geodesic distance D_{ij}^{G} that uses the Dijkstra's or Floyd's shortest-path algorithm. matrix D_G is the geodesic distance between all the data points. Finally, Isomap runs MDS on matrix D_G to find the low-dimensional embedding. This method works well for many different fields, but unfortunately, if there is a large data space N, the time consumed by this algorithm to find the shortest path is not optimal in ($O(kN^2 \log N)$)

and the MDS eigenvalue decomposition ($O(N^3)$) [4,5]. To improve the speed of these two values, L-Isomap is presented. In L-Isomap, a series of points are considered random points. Instead of calculating the shortest path between all available data, the distance between points and random points, called Landmark points, is calculated. The classical MDS is then applied to the result obtained from the $n \times N$ geodesic distance matrix to embed the low dimensions of the Landmark points. The embeddings of the remaining points are achieved by their geodesic distance to the landmark points by a fixed linear change. In this way, the time complexities of calculating the shortest path and the MDS computation are decreased to $O(knN$ log $N)$ and $O(n^2N)$.

Landmark Candidate is a very significant procedure in the L-Isomap is to create the k-nearest neighborhood graph. Many points are similar because of shared data, so some points can be omitted to make the neighborhood graph and diagram easier. For this purpose, we select candidate landmark points using a simple neighborhood graph. This algorithm is summarized in Algorithm 1 and it can run in time complexities $O(N$ log $N)$ [6].

(1) Let $\Omega* = \Phi$. (2) If u_{S_i}= 0 for all i then stop: $\Omega*$ is the cover, where i = 1, 2, . . . ,N. Otherwise, find a subscript $i_r \in \{1, 2, . . . , N\}$, maximizing the ratio $u_{S_i}/c(S_i)$ and proceed to Step (3). (3) Add $S_{i_r}$to $\Omega*$, replace each Si by $S_i - S_{i_r}$and return to Step (1).

Algorithm 1: Selection of landmark candidate

B. Locally Linear Embedding [LLE]

LLE can be mentioned as one of the helpful techniques in the area of dimensional reduction. In comparison with the Isomap technique, there is a slight difference between them. While both compute the distance between points as small linear neighborhoods in the nonlinear manifold, Isomap uses the geodesic distance to obtain the graph traversal, but LLE finds the weights that execute the local linear interpolations [7].

Algorithm 3 LLE

The LLE algorithm consists of three main stages:

1) Discovering K Nearest Neighbors of each point via Euclidean distance. 2) Calculating the local interpolation weight matrix of each sample point. For this purpose, the error function must be minimized as follows:

$$\min \varepsilon(W) = \sum_{i=1}^{N} \left| x_i - \sum_{j=1}^{K} w_j^i x_{ij} \right|^2 \tag{6}$$

Such that.

$$\sum_{j=1}^{K} w_j^i = 1 \tag{7}$$

Where the k neighbor of x_i are defined as x_{ij} (j=$1,\ldots,k$), and the weights between x_i and x_{ij} are presented as w_j^i. A local COV matrix also should be created for the W matrix as follows:

$$Q_{jm}^i = \left(x_i - x_{ij}\right)^T (x_i - x_{im}) \tag{8}$$

By combining the last two equations, the local optimization reconstruction weight matrix can be obtained:

$$w_j^i = \frac{\sum_{m=1}^{k} \left(Q^i\right)_{jm}^{-1}}{\sum_{p=1}^{k} \sum_{q=1}^{k} \left(Q^i\right)_{pq}^{-1}} \tag{9}$$

3) computing the result of the sample point by its weight matrix and the nearest neighbors.

$$\min \varepsilon(Y) = \sum_{i=1}^{N} \left| y_i - \sum_{j=1}^{K} w_j^i y_{ij} \right|^2 \tag{10}$$

Such that.

$$\sum_{i=1}^{N} y_i = 0 \tag{11}$$

$$\frac{1}{N} \sum_{i=1}^{N} y_i y_i^T = I \quad ; (I \ is\ a\ unit\ matrix) \tag{12}$$

Reference

[1]: Zhao, J., & Jiang, Q. (2006). Probabilistic PCA for t distributions. Neurocomputing, 69(16-18), 2217-2226.

[2]: Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, *10*(66-71), 13.

[3]: Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3), 611-622.

[4]: E. W. Dijkstra, “A note on two problems in connexion with graphs,” *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[5]: R. W. Floyd, “Algorithm 97: shortest path,” *Communications of the ACM*, vol. 5, no. 6, article 345, 1962.

[6]: Shi, H., Yin, B., Kang, Y., Shao, C., & Gui, J. (2017). Robust l-Isomap with a novel landmark selection method. Mathematical Problems in Engineering, 2017.

[7]: Ventura, D. (2008). Manifold Learning Examples–PCA, LLE, and ISOMAP. Department of Computer Science, Brigham Young University, Provo, UT, USA.