Foundation of Machine Learning

(CSE4032)

Lecture 05: Dimensionality Reduction Methods

Dr. Kundan Kumar
Associate Professor
Department of ECE

Faculty of Engineering (ITER)
S'O'A Deemed to be University, Bhubaneswar, India-751030

Outline

Principal Component Analysis

Principal Components Analysis

- Principal Component Analysis (PCA) is a method of dimension reduction. This is not directly related to prediction problem, but several regression methods are directly dependant on it.
- The regression methods (PCR and PLS) will be considered later.
- Principal component analysis is one of the most common methods used for linear dimension reduction.
- The motivation behind dimension reduction is that, the process gets unweildy with a large number of variables while the large number does not add any new information to the process.
- A linear combination of variables is then considered which are orthogonal to one another, but the total variability within the sample is preserved as much as possible.

Principal Component Analysis

- Suppose the data is 10-dimensional but needs to be reduced to 2-dimensional. The idea of principal component analysis is to use two directions that capture the variation in the data as much as possible.
- An analogy may be drawn with variance inflation factors (VIF) in multiple regression. If VIF corresponding to any predictor is large, that predictor is not included in the model, as that variable does not contribute any new information.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

- On the other hand, because of linear dependence, the regression matrix may become singular. In a multivariate situation, it may well happen that, a few (or a large number of) variables have high interdependence.

Singular Value Decomposition (SVD)

- Singular value decomposition is the key part of principal components analysis.
- Assume that the columns of $\mathbf{X}$ are zero-centered, i.e., the estimated column mean is subtracted from each column.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{pmatrix}$$

- The SVD of the $N \times p$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

Singular Value Decomposition (SVD)

- where
  - $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N)$ is an $N \times N$ orthogonal matrix. $\mathbf{u}_j,\ j = 1, \ldots, N$, form an orthonormal basis for the space spanned by the column vectors of $\mathbf{X}$.
  - $\mathbf{V} = (v_1, v_2, \ldots, v_p)$ is an $p \times p$ orthogonal matrix. $v_j,\ j = 1, \ldots, p$, form an orthonormal basis for the space spanned by the row vectors of $\mathbf{X}$.
  - $\mathbf{D}$ is a $N \times p$ rectangular matrix with nonzero elements along the first $p \times p$ submatrix diagonal. $\operatorname{diag}(d_1, d_2, \ldots, d_p), d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ are the singular values of $\mathbf{X}$ with $\mathrm{N} > \mathrm{p}$.
  - The columns of $\mathbf{V}$ (i.e., $v_j, j = 1, \ldots, p$) are the eigenvectors of $\mathbf{X}^T\mathbf{X}$. They are called principal component direction or eigenvectors of $\mathbf{X}$.
  - The diagonal values in $\mathbf{D}$ (i.e., $d_j, j = 1, \cdots, p$) are the square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$.

Eigendecomposition

- The sample covariance matrix of $\mathbf{X}$ is given as:

$$\mathbf{S} = \mathbf{X}^T\mathbf{X}/N$$

- If you do the Eigendecomposition of $\mathbf{X}^T\mathbf{X}$:

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \left(\mathbf{U}\mathbf{D}\mathbf{V}^T\right)^T\left(\mathbf{U}\mathbf{D}\mathbf{V}^T\right)\\ &= \mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}^2\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}\end{aligned}$$

- It turns out that if you have done the singular value decomposition then you already have the Eigendecomposition for $\mathbf{X}^T\mathbf{X}$.

Eigendecomposition

- The $\mathbf{\Lambda}$ is the diagonal part of matrix $\mathbf{D}$ with every element on the diagonal squared.
- Also, we should point out that we can show using linear algebra that $\mathbf{X}^T\mathbf{X}$ is a semi-positive definite matrix. This means that all of the eigenvalues are guaranteed to be nonnegative. The eigen values are in matrix $\mathbf{\Lambda}$. Since these values are squared, every diagonal element is non-negative.
- The eigenvectors of $\mathbf{X}^T\mathbf{X}$, v_j, can be obtained either by doing an Eigen decomposition of $\mathbf{X}^T\mathbf{X}$, or by doing a singular value decomposition from $\mathbf{X}$.
- These v_j^Ts are called principal component directions of $\mathbf{X}$. If you project $\mathbf{X}$ onto the principal components directions you get the principal components.

Principle Components

- It's easy to see that $\mathbf{z}_j = \mathbf{X} v_j = \mathbf{u}_j d_j$. Hence $\mathbf{u}_j$ is simply the projection of the row vectors of $\mathbf{X}$, i.e., the input predictor vectors, on the direction v_j, scaled by d_j. For example:

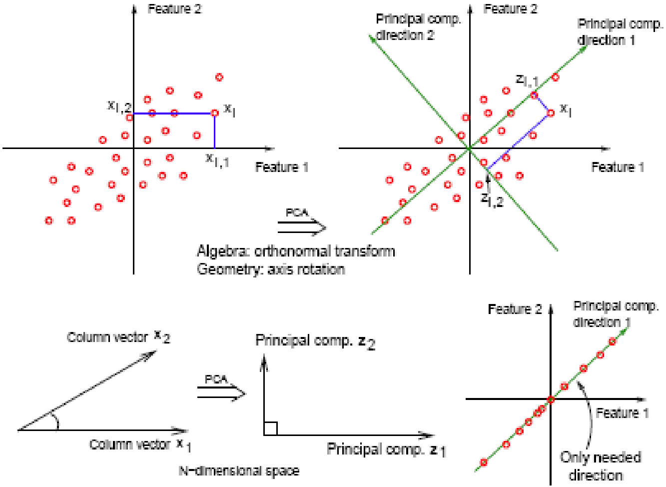
$$\mathbf{z}_1 = \begin{pmatrix} X_{1,1}v_{1,1} + X_{1,2}v_{1,2} + \ldots + X_{1,p}v_{1,p} \\ X_{2,1}v_{1,1} + X_{2,2}v_{1,2} + \ldots + X_{2,p}v_{1,p} \\ \vdots \\ X_{N,1}v_{1,1} + X_{N,2}v_{1,2} + \ldots + X_{N,p}v_{1,p} \end{pmatrix}$$

  The principal components of $\mathbf{X}$ are $\mathbf{z}_j = d_j \mathbf{u}_j, j = 1, \ldots, p$.
- The first principal component of $\mathbf{X}, \mathbf{z}_1$, has the largest sample variance amongst all normalized linear comninations of the coulmns of $\mathbf{X}$.

$$\mathrm{Var}\left(\mathbf{z}_1\right) = d_1^2 / N$$

Principal Components

- Subsequent principal components $\mathbf{z}_j$ have maximum variance d_j^2/N, subject to being orthogonal to the earlier ones.



Principal Components Analysis (PCA)

- Objective: Capture the intrinsic variability in the data. Reduce the dimensionality of a data set, either to ease interpretation or as a way to avoid overfitting and to prepare for subsequent analysis.
- The sample covariance matrix of $\mathbf{X}$ is $\mathbf{S} = \mathbf{X}^T\mathbf{X}/N$, since $\mathbf{X}$ has zero mean. Eigendecomposition of $\mathbf{X}^T\mathbf{X}$:

$$\mathbf{X}^T\mathbf{X} = \left(\mathbf{UDV}^T\right)^T \left(\mathbf{UDV}^T\right) = \mathbf{VD}^T\mathbf{U}^T\mathbf{UDV}^T = \mathbf{VD}^2\mathbf{V}^T$$

- The eigenvectors of $\mathbf{X}^T\mathbf{X}$ (i.e., $v_j,\ j = 1, \ldots, p$) are called principal component directions of $\mathbf{X}$.
- The first principal component direction v_1 has the following properties that
  - v_1 is the eigenvector associated with the largest eigenvalue, d_1^2, of $\mathbf{X}^T\mathbf{X}$.

Principal Components Analysis (PCA)

- $\mathbf{z}_1 = \mathbf{X}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of X.
- $\mathbf{z}_1$ is called the first principal component of $\mathbf{X}$.
- we have $\mathrm{Var}(\mathbf{z}_1) = d_1^2/N$.

- The second principal component direction v_2 (the direction orthogonal to the first component that has the largest projected variance) is the eigenvector corresponding to the second largest eigenvalue, d_2^2, of $\mathbf{X}^T\mathbf{X}$, and so on.
- The eigenvector for the kth largest eigenvalue corresponds to the kth principal component direction v_k).
- The kth principal component of $\mathbf{X}, \mathbf{z}_k$, has maximum variance d_k^2/N, subject to being orthogonal to the earlier ones.

Principal Components Regression

Principal Components Regression (PCR)

- Principal component regression forms the derived input columns $\mathbf{z}_m = \mathbf{X} v_m$, and then regresses $\mathbf{y}$ on $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M$ for some $M \leq p$. Since the $\mathbf{z}_m$ are orthogonal, this regression is just a sum of univariate regressions:

$$\hat{\mathbf{y}}_{(M)}^{\mathrm{pcr}} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

where $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$. Since the $\mathbf{z}_m$ are each linear combinations of the original $\mathbf{x}_j$, we can express the solution in terms of coefficients of the $\mathbf{x}_j$:

$$\hat{\beta}^{\mathrm{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

Principal Components Regression (PCR)

- As with ridge regression, principal components depend on the scaling of the inputs, so typically we first standardize them.
- Note that if $M = p$, we would just get back the usual least squares estimates, since the columns of $Z = \mathbf{UD}$ span the column space of $\mathbf{X}$.
- For $M < p$ we get a reduced regression. We see that principal components regression is very similar to ridge regression: both operate via the principal components of the input matrix.
- Ridge regression shrinks the coefficients of the principal components, shrinking more depending on the size of the corresponding eigenvalue; principal components regression discards the $p - M$ smallest eigenvalue components.

Partial Least Squares

Partial Least Squares (PLS)

- This technique also constructs a set of linear combinations of the inputs for regression, but unlike principal components regression it uses $\mathbf{y}$ (in addition to $\mathbf{X}$) for this construction.
- Like principal component regression, partial least squares (PLS) is not scale invariant, so we assume that each $\mathbf{x}_j$ is standardized to have mean 0 and variance 1.

Algorithm PLS

- Partial Least Squares

1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\mathbf{y}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j, j = 1, \ldots, p$.
2. For $m = 1, 2, \ldots, p$
   - $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} \mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \left\langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \right\rangle$.
   - $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.
   - $\hat{\mathbf{y}}^{(m)} = \mathbf{y}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$
   - Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m : \mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - \left[\left\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \right\rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle \right] \mathbf{z}_m, j = 1, 2, \ldots, p$.
3. Output the sequence of fitted vectors $\left\{ \hat{\mathbf{y}}^{(m)} \right\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\mathbf{y}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

References

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Hastie, Tibshirani, and Friedman, Springer.
- In Introduction to Statistical Learning with Application in R, Second Edition, James, Witten, Hastie, and Tibshirani, Springer.

Thank you!