Dimension Reduction

Damien Benveniste

October 25, 2017

1 Regularization

Regularization is a process to prevent overfitting. In linear regression or logistic regression, it is usually achieved by adding a controlling parameter to the loss function

$$\epsilon(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + f(\beta). \tag{1}$$

The two main regularization schemes are Ridge regression

$$f(\beta) = \lambda \|\beta\|_2^2 \tag{2}$$

and Lasso regression

$$f(\beta) = \lambda \|\beta\|_1 \tag{3}$$

where λ is a free parameter. Although both Ridge and Lasso tend to reduce the magnitudes of the components of β and Lasso can zero out some less relevant features. As such Lasso can be used as a feature reduction process. Let's consider

$$\epsilon(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \tag{4}$$

We have

$$\nabla_\beta \lambda \|\beta\|_1 = \lambda \frac{\beta}{\|\beta\|_1} \tag{5}$$

therefore the Normal equations are transformed as

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\left(\mathbf{X}^T\mathbf{y} - \lambda\frac{\beta}{\|\beta\|_1}\right) \tag{6}$$

Let's consider the orthonormal case

$$\mathbf{X}^T\mathbf{X} = \mathbf{I} \tag{7}$$

then we have

$$\beta = \mathbf{X}^T\mathbf{y} - \lambda\frac{\beta}{\|\beta\|_1} \tag{8}$$

If we consider the component β_j we have then we have

$$\beta_j = \mathbf{X}^T\mathbf{y}\,|_j - \lambda \text{sign}(\beta_j) \tag{9}$$

where $\mathbf{X}^T\mathbf{y}|_j$ is the projection of $\mathbf{X}^T\mathbf{y}$ onto the j^{th} direction. We can see that there exist a value of λ that drives β_j to 0. For larger value of λ, sign(β_j) would change keeping β_j at 0.

2 Principal component analysis

The idea is to rotate the problem into an orthonormal basis (linear transformation) where each component captures as most variance as possible.

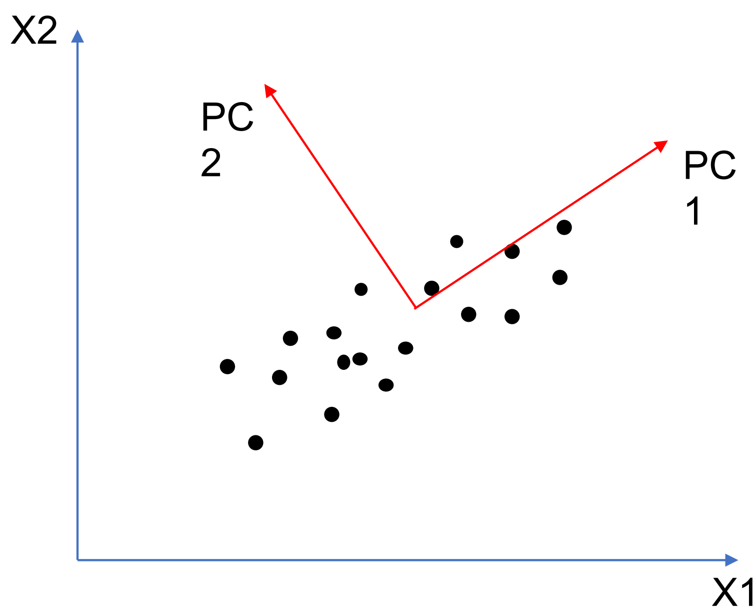

Figure 1

In the new basis, the components that captures the least variance are considered to be less important and are ignored.

Let's consider a feature matrix $\mathbf{X}$ that has been centered and normalized:

$$X_i = \frac{X'_i - \mu_{X'_i}}{\sigma_{X'_i}}. \tag{10}$$

As such the correlation matrix is simply

$$\rho = \mathbf{X}^T\mathbf{X} \tag{11}$$

We are interested to re-express this matrix in a basis such that the correlation is a diagonal matrix $\mathbf{\Lambda}$ (the eigen-vectors $\mathbf{W}$ of this basis are orthogonal)

$$\rho = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T \tag{12}$$

or

$$\mathbf{\Lambda} = \mathbf{W}^T \rho \mathbf{W}. \tag{13}$$

In this new basis the eigen-values of $\mathbf{\Lambda}$ represent the variance in each direction. The eigen-vectors or the principal components are ordered by these variances and the components with the lowest associated variances could be neglected. This effectively reduces the dimensionality of the problem.