PCA Kernel Notes

Lorenzo D'Isidoro

1 Overview

Many machine learning algorithms assume that the training dataset is linearly separable, if not, the PCA kernel can be used to transform it into a smaller linearly separable one.

The aim is to perform a non-linear mapping of n dataset features, that cannot be linearly separated, in a new larger space and then apply the standard PCA on it to project the samples into a smaller subspace, which is linearly separable.

N points will be transformed in a new subspace k-dimensional using the map function Φ with $(k >> d)$

$$x_i \in \mathbb{R}^d \Rightarrow \Phi(x_i) : \mathbb{R}^d \rightarrow \mathbb{R}^k$$

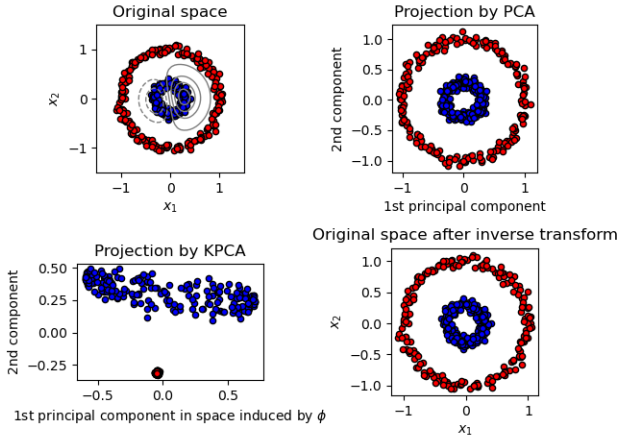

Figure 1: a plot which show that Kernel PCA is able to find a projection of the data that makes data linearly separable. Source scikit-learn.org

2 Covariance Matrix

The elements of the covariance matrix C for the features k and j are

$$COV(x_j, x_k) = \sigma_{j,k} = \frac{1}{n}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)(x_k^{(i)} - \mu_k)$$

$$C_{n,d} = \begin{pmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & c_{n,2} & \cdots & c_{n,d} \end{pmatrix}$$

If the features are standardized the avarage μ_j and μ_k will be 0.

At this point the equation which is used to calculate the covariance matrix is

$$C = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T$$

Apply the map function called Φ to the combination of nonlinear features

$$C_{map} = \frac{1}{n}\sum_{i=1}^{n}\Phi(x^{(i)})\Phi(x^{(i)})^T = \frac{1}{n}\Phi(X)\Phi(X)^T$$

The eigenvalues and eigenvectors can be calculated starting from the following equation

$$C_{map}v = \lambda v$$

Where λ is the vector of the eigenvalues to be used to calculate the eigenvectors and substituting C_{map}

$$C_{map}v = \lambda v \Rightarrow (\frac{1}{n}\sum_{i=1}^{n}\Phi(x^{(i)})\Phi(x^{(i)})^T)v = \lambda v \Rightarrow v = v\frac{1}{n\lambda}\sum_{i=1}^{n}\Phi(x^{(i)})\Phi(x^{(i)})^T$$

$$\Rightarrow v = \frac{1}{n}\sum_{i=1}^{n}\Phi(x^{(i)})a^{(i)} = \Phi(X)^T a$$

$$Where: Ka = \lambda a$$

Then

$$C_{map}v = \lambda v \Rightarrow (\frac{1}{n}\Phi(X)\Phi(X)^T)(\Phi(X)^T a) = \lambda(\Phi(X)^T a)$$