Question on Kernel_PCA: Sp24_567_VATSAL

sbmohant

April 2024

1 Introduction

OK

2 Kernel PCA (12 points)

Just like the other linear algorithms discussed in class, it can often be useful to do PCA in a higher-dimensional feature space. In fact, we can actually apply the kernel trick to PCA! In this problem we will derive some properties of *Kernel PCA*.

Suppose we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$, and some feature map $\phi(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^M$. Consider the data matrix $\boldsymbol{X}$ of the original data, and the feature matrix $\boldsymbol{\Phi}$ of the transformed datapoints defined as usual as follows:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \boldsymbol{\Phi} = \begin{bmatrix} \phi(\boldsymbol{x}_1)^\top \\ \phi(\boldsymbol{x}_2)^\top \\ \vdots \\ \phi(\boldsymbol{x}_n)^\top \end{bmatrix} \in \mathbb{R}^{n \times M}.$$

Assume that $\boldsymbol{\Phi}$ is centered, so $\boldsymbol{C} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ is the covariance matrix of the data in the high-dimensional feature space. Let $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top$ be the Kernel matrix for the data with the feature transformation $\phi(\cdot)$. Recall that for any matrix $\boldsymbol{A}$, $(\lambda, \boldsymbol{v})$ is any eigenpair of $\boldsymbol{A}$ if $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$. Since if $\boldsymbol{v}$ is an eigenvalue of $\boldsymbol{A}$ then $\alpha \boldsymbol{v}$ is also an eigenvector for any $\alpha \neq 0$, we usually normalize eigevectors to have unit norm, i.e. $\|\boldsymbol{v}\|_2 = 1$.

(a) For any $\lambda \neq 0$, show that if $(\lambda, \boldsymbol{v})$ is an eigenpair of $\boldsymbol{C}$, then $(\lambda, \boldsymbol{\Phi}\boldsymbol{v})$ is an eigenpair of $\boldsymbol{K}$. (2 points)

(b) Similarly for any $\lambda \neq 0$, show that if $(\lambda, \boldsymbol{a})$ is an eigenpair of $\boldsymbol{K}$, then $(\lambda, \boldsymbol{\Phi}^\top \boldsymbol{a})$ is an eigenpair of $\boldsymbol{C}$. (2 points)

(c) By the previous two parts, for any $\lambda \neq 0$, λ is an eigenvalue of $\boldsymbol{C}$ if and only if λ is an eigenvalue of $\boldsymbol{K}$. This also means that the largest magnitude eigenvalue of $\boldsymbol{K}$ is also the largest magnitude eigenvalue of $\boldsymbol{C}$.

Let $\boldsymbol{a}_1$ be the eigenvector of $\boldsymbol{K}$ corresponding to the largest magnitude eigenvalue λ_1. Show that the first principal component is given by,

$$\boldsymbol{v}_1 = \frac{\boldsymbol{\Phi}^\top \boldsymbol{a}_1}{\|\boldsymbol{\Phi}^\top \boldsymbol{a}_1\|_2}.$$

Assume that $\|\boldsymbol{a}_1\|_2 = 1$, and show that this expression can be simplified to,

$$\boldsymbol{v}_1 = \frac{\boldsymbol{\Phi}^\top \boldsymbol{a}_1}{\sqrt{\lambda_1}}. \tag{1}$$

(3 points)

(d) Given a datapoint $\boldsymbol{x}$, show that its projection α_1 onto the first principle component $\boldsymbol{v}_1$ is given by

$$\alpha_1 = \frac{1}{\sqrt{\lambda_1}} \sum_{i=1}^{n} k(\boldsymbol{x}, \boldsymbol{x}_i)(\boldsymbol{a}_1)_i.$$

Here $(\boldsymbol{a}_1)_i$ denotes the i-th coordinate of the vector $\boldsymbol{a}_1$. To derive this, use the expression for $\boldsymbol{v}_1$ from Eq. 1 above. (Hint: The projection of a vector $\boldsymbol{a}$ onto $\boldsymbol{b}$ is $\frac{\boldsymbol{b}^\top \boldsymbol{a}}{\|\boldsymbol{b}\|_2}$.) (3 points)

(e) Figure 8a represents a dataset that doesn't seem to be amenable to PCA. So we apply a non-linear transformation $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\phi\left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} \sqrt{x^2 + y^2} \\ \text{arctan2}(y, x) \end{bmatrix}$$

where the function arctan2 returns the angle from positive x-axis of the point (x,y) in the range $(0, 2\pi)$.

Similarly we define the inverse of ϕ as $\phi^{-1}\left(\begin{bmatrix} r \\ \theta \end{bmatrix} \right) = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}$

Locate the arrow(s) which best describe the following phenomena -

1) PCA does not always reduce dimensions but can find an alternative basis to represent the data.

2) Projecting the data onto the leading principal components are not always better for downstream classification tasks. Sometimes projecting onto the less significant principal components which contain lesser variance might be more suitable for downstream classification.

3) Kernel PCA can find a suitable non-linear basis to represent the original data on a manifold instead of a linear subspace as achieved by linear PCA.

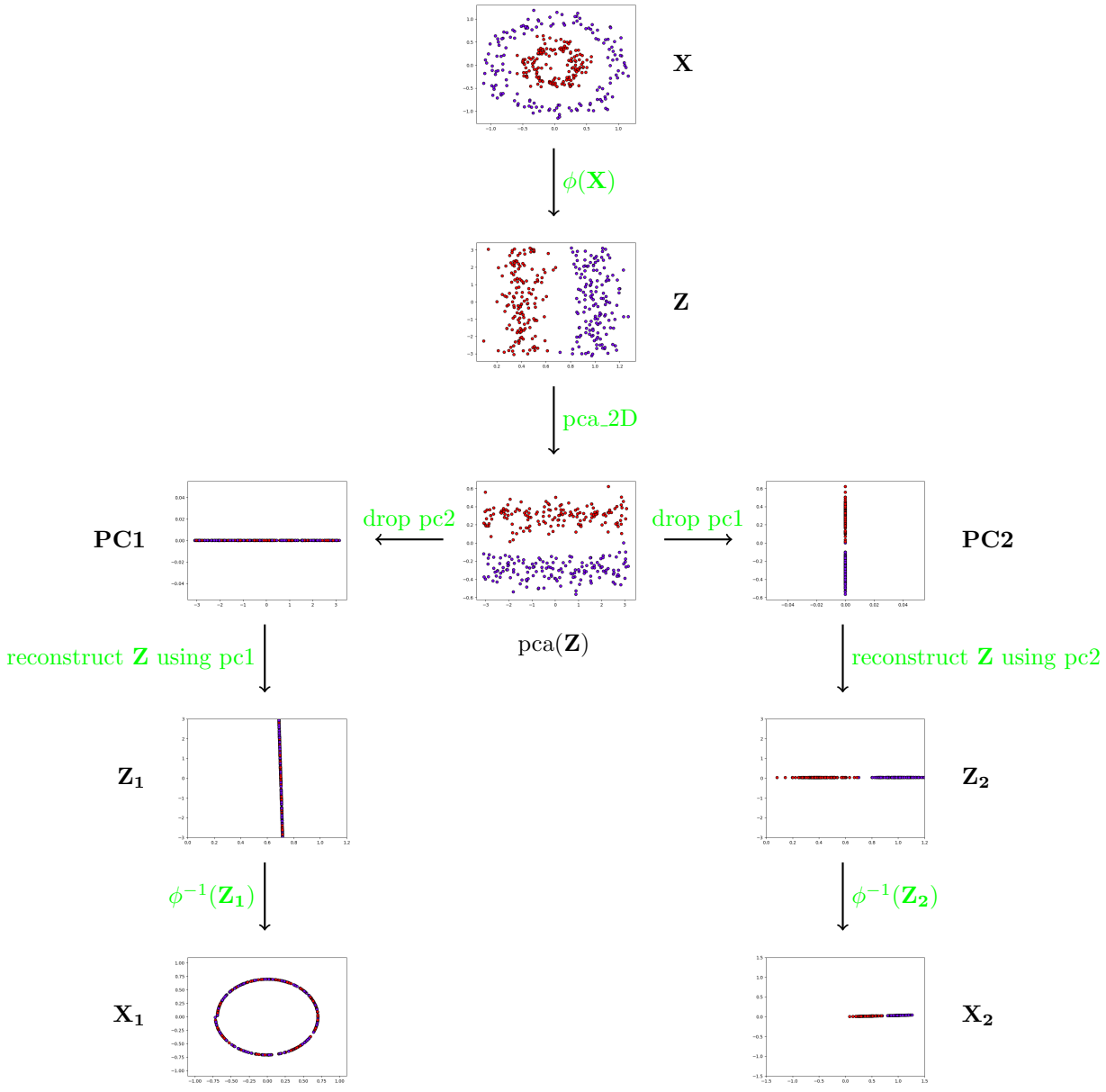

Figure 1: Kernel PCA step by step