Randomized Kernel PCA

Deepak
Leaandro Rodricks

Supervisor: Sivananthan Sampath

May 12, 2023

Table of Contents

Table of Contents

Abstract

Problem Statement

Given a set of n data points, each lying in a d-dimensional space, find the most computationally efficient method that can be used to store them efficiently.

Abstract

Problem Statement

Given a set of n data points, each lying in a d-dimensional space, find the most computationally efficient method that can be used to store them.

Methods:

- Representation Learning
- Principal component analysis (PCA)
- Kernel PCA
- Nystrom sampling

Table of Contents

Representation Learning

- Representation learning refers to a type of machine learning approach that involves discovering or learning a feature space, or representation, that allows raw input data to be transformed into a more informative and useful format.

- The main goal of representation learning is to create a more compact and efficient representation of the data that captures the underlying patterns and structure in the input. This will lead to reduce the number of elements used in the representation of data.

Representation Learning

Example

Given $x_1, x_2, x_3, x_4 \in \mathbb{R}^2$ such that
$x_1 = [-7, -14]^T, x_2 = [3, 6]^T, x_3 = [-4, 8]^T, x_4 = [1, 2]^T$.

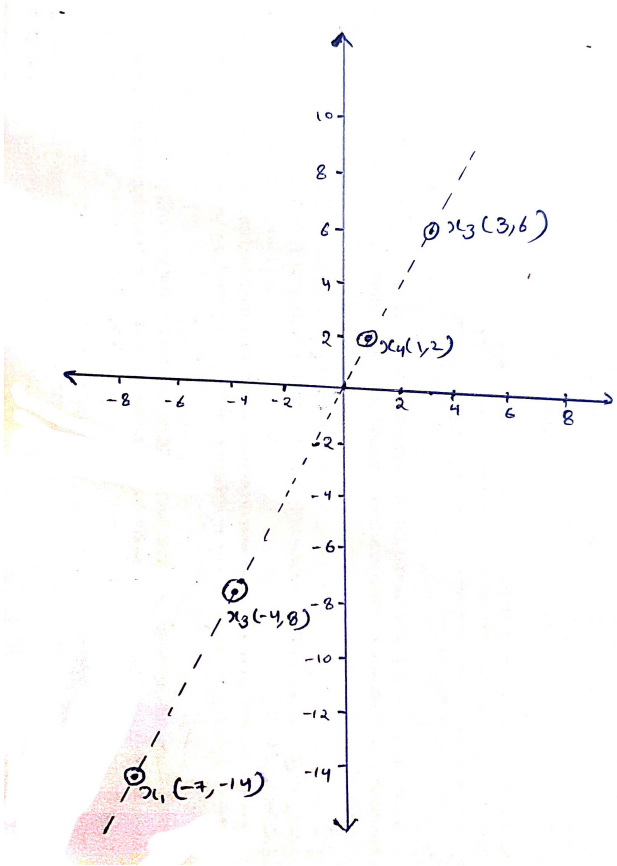
Sol.

- To store this input as it is, we would require 8 elements for each coordinate of the 4 points.

Example

Given $x_1, x_2, x_3, x_4 \in \mathbb{R}^2$ such that
$x_1 = [-7, -14]^T, x_2 = [3, 6]^T, x_3 = [-4, 8]^T, x_4 = [1, 2]^T$.

Plotting the points:



Representation Learning

Example

Given $x_1, x_2, x_3, x_4 \in \mathbb{R}^2$ such that
$x_1 = [-7, -14]^T, x_2 = [3, 6]^T, x_3 = [-4, 8]^T, x_4 = [1, 2]^T$.

Sol.

- To store this input as it is, we would require 8 elements for each coordinate of the 4 points.
- **Observation:**
  All the 4 points lie on a straight line with with the equation $y = 2x$.
  $\Rightarrow$ We can use a representative vector $r = [1, 2]^T$ and a coefficient vector $c = [-7, 3, -4, 1]$ to store the 4 points.
  In this representation, we need only $2 + 4 = 6$ elements instead of 8.

Representation Learning

Question

What if the dataset has another point $x_5(5, 5)$?

Sol. PCA

Table of Contents

Principal component analysis (PCA)

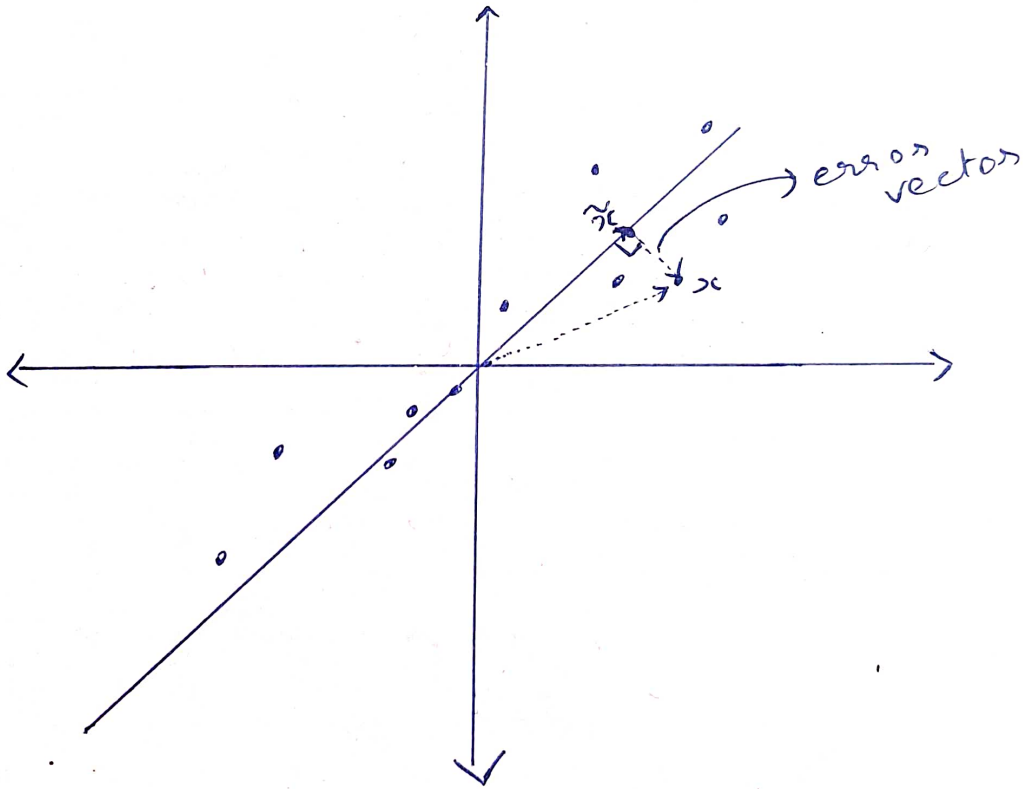
Def.

PCA is a statistical technique used to reduce the dimensionality (number of variables in the dataset) of the data while retaining the most important information, i.e., PCA is a linear dimension-reduction technique that finds new axes that maximize the variance in the data.

Principal component analysis (PCA)

Basic Intuition

Assume a lower dimensional direction($\hat{w}$), project all data points(x_i) onto the assumed direction and minimize the the total error to get the direction, i.e., $min \sum \|x_i - (x_i^T w)w\|^2$.



Principal component analysis (PCA)

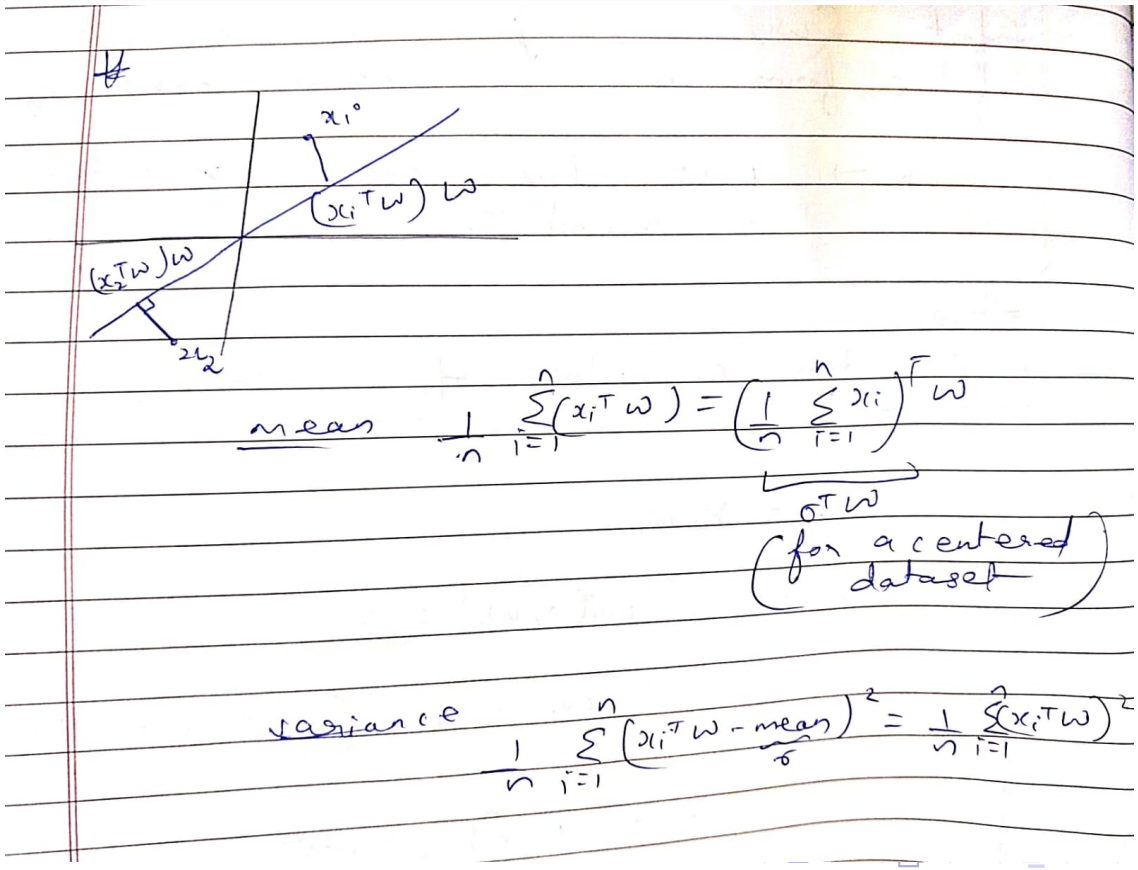
Algorithm

- Given $D = \{x_1, x_2, ..., x_n\}$ where $x_i \in \mathbb{R}^d$; compute the covariance matrix $C = 1/n \sum_{i=1}^{n} x_i x_i^T$.
- Given w which is a random unit vector ($\|w\|_2{}^2 = 1$), find another direction vector such that $w_1 = argmax\,(w^t C w)$.
- Compute the projections of the data points onto the w_1 vector to get the new data points x_i^1 like
  $x_1^1 \Rightarrow x_1 - (x_1^T w_1) w_1, ...x_n^1 \Rightarrow x_n - (x_n^T w_1) w_1$
- Similarly compute the covariance matrix $C^1 = 1/n \sum_{i=1}^{n} x_i^1 (x_i^1)^T$ for the projected data set obtained and find a new direction vector such that $w_2 = argmax\,(w^T C^1 w)$
- Repeat these iterations till all the errors associated with x1, x2, ..., xn become approximately 0.

Principal component analysis (PCA)

Applying PCA implies getting a direction w that maximizes the variance in the data. How?

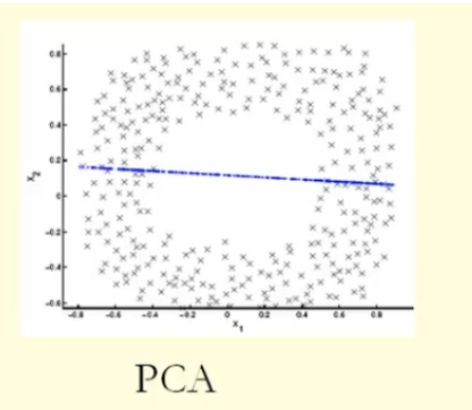
- As discussed in the algorithm, PCA deals with finding a direction w_1 such that $w_1 = argmax\,(w^t C w)$



Principal component analysis (PCA)

Limitations

- **High Time Complexity:** The covariance matrix of the features would be a d*d matrix, and we need to find the k largest eigenvalues and corresponding eigenvectors. The time complexity for this process would be $O(d^3)$.
- **Features having non-linear relationships:** The data may not lie in a low-dimensional **linear** subspace of the feature space and thus, we may need non-linear principal components.


PCA

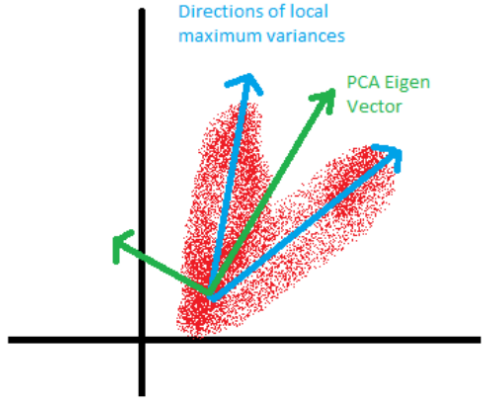


Table of Contents

Kernel

Def.

Any function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel function if:

- K is symmetric, i.e., $K(x, \tilde{x}) = K(\tilde{x}, x)$
- For any dataset $\{x_1, x_2, ..., x_n\}$, the matrix $K \in \mathbb{R}^d$ (where $K_{ij} = K(x_i, x_j)$) is positive semi-definite, i.e., all eigenvalues of K are non-negative. A matrix K is positive semi-definite if and only if for all vectors x, the following inequality holds:
  $x^T Ax \geq 0$, where x^T denotes the transpose of x.
  Another way to check if a matrix is positive semi-definite is to check that all its principal sub-matrices have non-negative determinants.

Kernel

Example of Kernel Functions

- **Linear Kernel:** If there are two vectors named x_1 and x_2, the linear kernel can be defined by the dot product of the two vectors: $K(x_1, x_2) = x_1 \cdot x_2$
- **Polynomial Kernel:** We can define a polynomial kernel as: $K(x_1, x_2) = (x_1 \cdot x_2 + 1)^d$. Here, d represents the degree of the polynomial.
- **Gaussian Kernel:** The Gaussian kernel can be represented with this equation: $k(x_i, x_j) = exp(-||x_i - x_j||^2/2\sigma^2)$.

Table of Contents

Kernel PCA

Def.

Kernel PCA is a nonlinear extension of traditional linear PCA that uses kernel functions to project the dataset into a higher dimensional feature space, where it is linearly separable.

Kernel PCA

Algorithm

- Compute the gram matrix $K \in R^{n*n}$ where $K_{ij} = k(x_i, x_j)\ \forall i, j$ and centralize it to get $\tilde{K}$.
- Compute the eigenvalues $n\lambda_1, n\lambda_2, ..., n\lambda_l$ and the eigenvectors $\beta_1, \beta_2, ..., \beta_l$ of $\tilde{K}$. Normalize the eigenvectors by the relation $\alpha_i = \beta_i / \sqrt{n\lambda_i}$ and obtain the set of vectors $\{\alpha_1, \alpha_2, ..., \alpha_l\}$
- If we write $w_k = \phi(x)\alpha_k$, we would have to compute $\phi(x)$ and reconstruct the eigenvectors of the covariance matrix. Instead compute $\sum_{j=1}^{N} \alpha_{kj} K_{ij}$ since the equation $\phi(x_i)^T w_k = \sum_{j=1}^{N} \alpha_{kj} K_{ij}$ holds true. Now we can use the following compressed representation $x_i \Rightarrow [\sum_{j=1}^{N} \alpha_{1j} K_{ij}, \sum_{j=1}^{N} \alpha_{2j} K_{ij}, ..., \sum_{j=1}^{N} \alpha_{lj} K_{ij}]$.
- Thus we have changed the dimension from d to l, where l can be greater or lesser than d, but the non-linear structure of the original data set would be retained.

Kernel PCA

Limitations

- **Higher time complexity for large data sets:** It involves finding the eigenvectors of the $N * N$ matrix $\tilde{K}$ rather than the $D * D$ matrix S of conventional linear PCA giving it a time complexity of $O(n^3)$.
- **Absence of an inverse mapping:** PCA provides an inverse mapping from the low-dimensional space back to the input space. So, input points can be approximately reconstructed from their low-dimensional images. Kernel PCA doesn't inherently provide an inverse mapping since the projection of points in feature space onto the linear PCA subspace in that space will typically not lie on the nonlinear D-dimensional manifold.

Table of Contents

Background

Best k-dimensional subspace

- **Theorem**- Let A be an nXd matrix where $v_1, v_2, ..., v_r$ are its singular vectors. For $1 <= k <= r$, let V_k be the subspace spanned by $v_1, v_2, ..., v_k$. Then for each k, V_k is the best-fit k-dimensional subspace for A.

Background

Best k-dimensional subspace

- **Theorem**- Let A be an nXd matrix where $v_1, v_2, ..., v_r$ are its singular vectors. For $1 <= k <= r$, let V_k be the subspace spanned by $v_1, v_2, ..., v_k$. Then for each k, V_k is the best-fit k-dimensional subspace for A.
- We can prove this theorem using induction. We can see that the statement is trivially true for k=1. for $k > 1$, by the IH, V_{k-1} is the best $(k-1)$ dimensional subspace. Suppose W is a best-fit k-dimensional subspace. Choose a basis $w_1, w_2, ..., w_k$ of W so that w_k is perpendicular to $v_1, v_2, ..., v_{k-1}$. Then $\Sigma_{i=1}^{k}|Aw_i|^2 <= \Sigma_{i=1}^{k-1}|Av_i|^2 + |Aw_k|^2$. Since w_k is perpendicular to $v_1, v_2, ..., v_{k1}$, by the definition of v_k, $|Aw_k|^2 <= |Av_k|^2$. Thus $\Sigma_{i=1}^{k}|Aw_i|^2 <= \Sigma_{i=1}^{k}|Av_i|^2$ proving that V_k is at least as good as W and hence is optimal.

Singular Value Decomposition

- Let A be an *nXd* matrix with singular vectors $v_1, v_2, ..., v_r$ and corresponding singular values $\sigma_1, \sigma_2, ..., \sigma_r$. Then $u_i = (1/\sigma_i)Av_i$, for i = 1, 2, . . . , r, are the left singular vector. Then, A can be decomposed into a sum of rank one matrices $A = \Sigma_{i=1}^{r} \sigma_i u_i v_i^T$.

Singular Value Decomposition

- Let A be an *nXd* matrix with singular vectors $v_1, v_2, ..., v_r$ and corresponding singular values $\sigma_1, \sigma_2, ..., \sigma_r$. Then $u_i = (1/\sigma_i)Av_i$, for i = 1, 2, . . . , r, are the left singular vector. Then, A can be decomposed into a sum of rank one matrices $A = \Sigma_{i=1}^{r} \sigma_i u_i v_i^T$.
- In matrix notation $A = U\Sigma V^T$ where the columns of U and V consist of the left and right singular vectors, respectively, and Σ is a diagonal matrix whose diagonal entries are the singular values of A. Since the vectors in U and V are orthogonal to each other, we can say that U and V are orthonormal matrices.
- Suppose A is a symmetric positive semi-definite (SPSD) matrix. Then U and V would be equal .i.e. $A = U\Sigma U^T$ and thus, $A^{-1} = U\Sigma^{-1}U^T$

Background

Matrix Norms

- The norm of a square matrix A is a non-negative real number,which is a measure of the magnitude of the matrix, denoted $||A||$. Similar to vector norms, it follows the properties of positivity, homogeneity, submultiplicativity and the triangle inequality among others.

Background

Matrix Norms

- The norm of a square matrix A is a non-negative real number,which is a measure of the magnitude of the matrix, denoted $||A||$. Similar to vector norms, it follows the properties of positivity, homogeneity, submultiplicativity and the triangle inequality among others.
- **Examples**

  1) **The 1-Norm** - $||A||_1 = max_{1<=j<=n} \Sigma_{i=1}^{n} |a_{ij}|$; is the maximum of the absolute column sums.

  2) **The 2-Norm (Spectral)** - $||A||_2 = max_{|v|=1} |Av|$; equals the largest singular value of the matrix.

  3) **The Frobenius Norm** - $||A||_F = \sqrt{\Sigma_{j,k} a_{jk}^2}$. However $\Sigma_{j=1}^{n} \Sigma_{k=1}^{d} a_{jk}^2$, which is in turn equal to $\Sigma_{i=1}^{r} \sigma_i^2(A)$; thus, is equal to root of the sum of the squares of singular values.

Background

Matrix Norms

- The norm of a square matrix A is a non-negative real number,which is a measure of the magnitude of the matrix, denoted $||A||$. Similar to vector norms, it follows the properties of positivity, homogeneity, submultiplicativity and the triangle inequality among others.
- **Examples**
  1) **The 1-Norm** - $||A||_1 = max_{1<=j<=n} \Sigma_{i=1}^{n} |a_{ij}|$; is the maximum of the absolute column sums.
  2) **The 2-Norm (Spectral)** - $||A||_2 = max_{|v|=1} |Av|$; equals the largest singular value of the matrix.
  3) **The Frobenius Norm** - $||A||_F = \sqrt{\Sigma_{j,k} a_{jk}^2}$. However $\Sigma_{j=1}^{n} \Sigma_{k=1}^{d} a_{jk}^2$, which is in turn equal to $\Sigma_{i=1}^{r} \sigma_i^2(A)$; thus, is equal to root of the sum of the squares of singular values.
- We are going to use the Spectral and Frobenius norms to compare the approximations of the Kernel Matrix in the following analysis.

Background

Kernel Approximations

- Let $K \in R^{nXn}$ be a symmetric positive semidefinite (SPSD) kernel or Gram matrix with $rank(K) = r <= n$. We will write the SVD of K as $K = U\Sigma U^T$ and the pseudo-inverse of K as $K^+ = \Sigma_{t=1}^{r} \sigma_t^{-1} U^{(t)} U^{(t)T}$, with $K^+ = K^{-1}$ when K is full rank. For $k < r$, $K_k = \Sigma_{t=1}^{k} \sigma_t^{-1} U^{(t)} U^{(t)T} = U_k \Sigma_k U_k^T$ is the 'best' rank-k approximation to K,i.e., $K_k = argmin_{K' \in R^{nxn}, rank(K')=k} ||K - K'||_{\xi \in \{2,F\}}$ where Σ_k contains the top-k singular values of X and U contains their associated singular vectors.

Background

Kernel Approximations

- Let $K \in R^{nXn}$ be a symmetric positive semidefinite (SPSD) kernel or Gram matrix with $rank(K) = r <= n$. We will write the SVD of K as $K = U\Sigma U^T$ and the pseudo-inverse of K as $K^+ = \Sigma_{t=1}^{r} \sigma_t^{-1} U^{(t)} U^{(t)T}$, with $K^+ = K^{-1}$ when K is full rank. For $k < r$, $K_k = \Sigma_{t=1}^{k} \sigma_t^{-1} U^{(t)} U^{(t)T} = U_k \Sigma_k U_k^T$ is the 'best' rank-k approximation to K,i.e., $K_k = argmin_{K' \in R^{nxn}, rank(K')=k} ||K - K'||_{\xi \in \{2,F\}}$ where Σ_k contains the top-k singular values of X and U contains their associated singular vectors.
- We can calculate the difference between the exact Kernel matrix and its best k-rank approximation using the matrix norms -
  1) $||K - K_k||_2 = \sigma_{k+1}$
  2) $||K - K_k||_F = \sqrt{\Sigma_{t=k+1}^{r} \sigma_t^2}$

Table of Contents

Nystrom method

Basic notion

The general idea in Nystrom sampling is to obtain a low-rank approximation to the Gram matrix K, and replace K by this approximation in kernel algorithms, resulting in computational speedup

Nystrom method

Process

- Let us assume that the sample of l columns is given to us, C denotes the *nxl* matrix formed by these columns and W the *lxl* matrix consisting of the intersection of these l columns with the corresponding l rows of K. Since K is SPSD, W is also SPSD. Thus, K and C can be written as- $K = \begin{bmatrix} W & K_{21}^T \\ K_{21} & K_{22} \end{bmatrix}$ and $C = \begin{bmatrix} W \\ K_{21} \end{bmatrix}$.

Nystrom method

Process

- Let us assume that the sample of l columns is given to us, C denotes the *nxl* matrix formed by these columns and W the *lxl* matrix consisting of the intersection of these l columns with the corresponding l rows of K. Since K is SPSD, W is also SPSD. Thus, K and C can be written as- $K = \begin{bmatrix} W & K_{21}^T \\ K_{21} & K_{22} \end{bmatrix}$ and $C = \begin{bmatrix} W \\ K_{21} \end{bmatrix}$.
- The Nystrom method generates a rank-k approximation $\tilde{K}$ of K for $k < n$ defined by $\tilde{K}_k^{nys} = CW_k^+C \approx K$, where W_k is the best k-rank approximation of W with respect to the spectral or Frobenius norm and W_k^+ denotes the pseudo-inverse of W_k. When $k = l$ (or $k >= rank(C)$), this approximation perfectly reconstructs three blocks of K, and K_{22} is approximated by the Schur Complement of W in K. $\tilde{K}_k^{nys} = \begin{bmatrix} W & K_{21}^T \\ K_{21} & K_{21}W^+K_{21}^T \end{bmatrix}$.

Nystrom method

Various techniques

- The most common class of sampling techniques that select columns using a fixed probability distribution. Their types are-
  1) **Plain Nystrom**- The columns are sampled uniformly at random without any replacement.
  2) **Diagonal Nystrom**- The i^{th} column can be sampled non-uniformly with weight proportional to its corresponding diagonal element K_{ii}
  3) **Column-Norm Nystrom**- The i^{th} column can be sampled non-uniformly with weight proportional to the L_2 norm of the column.

Nystrom method

Various techniques

- The most common class of sampling techniques that select columns using a fixed probability distribution. Their types are-
  1) **Plain Nystrom**- The columns are sampled uniformly at random without any replacement.
  2) **Diagonal Nystrom**- The i^{th} column can be sampled non-uniformly with weight proportional to its corresponding diagonal element K_{ii}
  3) **Column-Norm Nystrom**- The i^{th} column can be sampled non-uniformly with weight proportional to the L_2 norm of the column.
- **Adaptive sampling**- Instead of sampling all l columns from a fixed distribution, we can alternate between selecting a set of columns and updating the distribution over all the columns. Starting with an initial distribution over the columns, $s < l$ columns are chosen to form a submatrix C'. The probabilities are then updated and s new columns are sampled and incorporated in C' and this continues until l columns have been selected.

Table of Contents

Theoretical Analysis (Method 1)

We now present some theoretical results that help us to determine the quality of the plain Nystrom approximation. We use 2 methods to do this- in one method, we form a comparison with the best k-rank approximation and achieve an inequality between the 2 difference norms, while in the second method, we achieve a Nystrom-error bound which varies according to the choice of Kernel taken.

Theoretical Analysis (Method 1)

Theorem 1

Let $Z_1, ..., Z_l$ be a sequence of random variables sampled uniformly without replacement from a fixed set of $l + u$ elements Z, and let $\phi : Z^l \rightarrow R$ be a symmetric function such that for all $i \in [1, l]$ and for all $z_1, ..., z_l \in Z$ and $z_1', ..., z_l' \in Z, |\phi(z_1, ..., z_l) - \phi(z_1, ..., z_{i-1}, z_i', z_{i+1}, ..., z_l)| \leq c$.
Then, for all $\in > 0$, the following inequality holds:
$Pr[[\phi] \geq \in] <= exp(-2 \in^2 / \alpha(l, u)c^2)$, where,
$\alpha(l, u) = (lu)/(l + u - 1/2) * 1/(1 - 1/(2max\{l, u\}))$

Theoretical Analysis (Method 1)

Proof Idea

Let X be a real random variable such that $E(X) = 0$ and $a <= X <= b$ for some $a, b \in R$. Then, for all $s \in R$, $log(Ee^{sX}) <= s^2(b-a)^2/8$. On substituting X as $\phi - E[\phi]$, we can calculate the PDF of this function by taking the Inverse Fourier Transform of the Characteristic Function (CF) of X, where the CF can be obtained from MGF by replacing X with iX.

Theoretical Analysis (Method 1)

Some definitions

We define the selection matrix corresponding to a sample of l columns as the matrix $S \in R^{nxl}$ defined by $S_{ii} = 1$ if the i^{th} column of K is among those sampled, $S_{ij} = 0$ otherwise. Thus, C=KS is the matrix formed by the columns sampled. Since K is SPSD, there exists $X \in R^{Nxn}$ such that $K = X^T X$. We shall denote by K_{max} the maximum diagonal entry of K,$K_{max} = max_i K_{ii}$, and by d^K_{max} the distance $max_{ij}\sqrt{K_{ii} + K_{jj} - 2K_{ij}}$.

Theoretical Analysis (Method 1)

Theorem 2

Let $Z = \sqrt{n/l}XS$. Then the following inequality holds:
$||K - \tilde{K}||_2 <= ||K - K_k||_2 + 2||XX^T - ZZ^T||_2$ where $\tilde{K}$ denotes the rank-k Nystrom approximation of K based on l columns sampled uniformly at random without replacement from K, and K_k the best rank-k approximation of K.

Theoretical Analysis (Method 1)

Proof Idea

Combining the results of the following 2 lemmas-

- If $\tilde{G}_k = CW_k^+ C^T$ then $||G - \tilde{G}_k||_2 = ||X - U_k U_k^T X||_2^2$.
- Suppose $A \in R^{mxn}$ and let H_k be the mxk matrix whose columns consist of the top k singular vectors of the mxc matrix C. Then, for every k : $0 <= k <= rank(C)$,
  $||A - H_k H_k^T A||_2^2 <= ||A - A_k||_2^2 + 2||AA^T - CC^T||_2$.

Theoretical Analysis (Method 1)

Theorem 3

Let $\tilde{K}$ denote the rank-k Nystrom approximation of K based on l columns sampled uniformly at random without replacement from K, and K_k the best rank-k approximation of K. Then, with probability at least $1 - \delta$, the following inequalities hold for any sample of size l:

1) $||K - \tilde{K}||_2 \leq ||K - K_k||_2 + 2n/\sqrt{l} K_{max}[1+ \sqrt{(n-l)/(n-1/2) * log(1/\delta)/\beta(l,n)} d^K_{max}/K^{0.5}_{max}]$

2) $||K - \tilde{K}||_F \leq ||K - K_k||_F + (64k/l)^{1/4} n K_{max}[1+ \sqrt{(n-l)/(n-1/2) * log(1/\delta)/\beta(l,n)} d^K_{max}/K^{0.5}_{max}]^{0.5}$

Theoretical Analysis (Method 1)

Proof Idea

Let $\phi(S) = ||XX^T - ZZ^T||_2$. Let S' be a sampling matrix selecting the same columns as S except for one, and let Z denote $\sqrt{n/l}XS$. Let z and z denote the only differing columns of Z and Z', then
$|\phi(S') - \phi(S)| <= ||z'z'^T - zz^T||_2 = ||(z' - z)z'^T + z(z' - z)^T||_2 \rightarrow$
$|\phi(S') - \phi(S)| <= 2||z' - z||_2 max(||z||_2, ||z'||_2)$.
Columns of Z are those of X scaled by $\sqrt{n/l}$. The norm of the difference of two columns of X can be viewed as the norm of the difference of two feature vectors associated to K and thus can be bounded by d^K_{max}.
Similarly, the norm of a single column of X is bounded by $K^{1/2}_{max}$. This leads to the following inequality:
$|\phi(S') - \phi(S)| <= (2n/l)d^K_{max}K^{1/2}_{max}....(9)$.
The expectation of ϕ can be bounded as follows: $E[\phi] = E[||XX^T - ZZ^T||_2] <= E[||XX^T - ZZ^T||_F] <= (n/\sqrt{l})K_{max}.....(10)$.

Theoretical Analysis (Method 2)

Theorem 4

Given the data set $X = (x_i)_{i=1}^{n}$, and the landmark point set $Z = (z_j)_{j=1}^{m}$. Then the Nystrom reconstruction of the kernel entry $K(x_i, x_j)$ will be exact if there exist two landmark points such that $z_p = x_i$, and $z_q = x_j$.

Theoretical Analysis (Method 2)

Proof Idea

Let $K_{x_k,Z} \in R^{1 \times m}$ be the similarity between x_k and the landmark points Z. Then the Nystrom reconstruction of the kernel entry will be $K_{x_i,Z} W^{-1} K_{x_j,Z}$, where $W \in R_{m \times m}$ is the kernel matrix defined on the landmark set Z. Let W(k) be the k^{th} row of W, then we have $K_{x_i,Z} = W(p)$ and $K_{x_j,Z} = W(q)$ since $x_i = z_p$, and $x_j = z_q$. As a result, the reconstructed entry will be $W^{(p)} W^{-1} (W^{(q)}) = W_{pq} = K(z_p, z_q) = K(x_i, x_j)$.

Theoretical Analysis (Method 2)

Some definitions

We first define a partial approximation error for a sub-kernel Matrix. Suppose that the landmark set is $Z = z_{i=1}^{m}$, and the whole sample set X is partitioned into m disjoint clusters S_k's. Suppose each cluster has T samples. Repeat the following sampling process T times: at each time t, pick one sample from each cluster and denote the set of samples chosen at time t as X_{I_t}. Then $X = X_{I_1} \cup X_{I_2} \cup ... \cup X_{I_T}$, and the whole kernel matrix will be correspondingly decomposed into T^2 blocks, each of size m $\times$ m. Let K_{I_i,I_j}, and $E_{I_i,Z}$ be the m$\times$m similarity matrices defined on (X_{I_i}, X_{I_j}) and (X_{I_i}, Z), respectively, and $W \in R^{m \times m}$ the kernel matrix defined on Z. The partial approximation error is the difference between K_{I_i,I_j} and its Nystrom approximation under the Frobenius norm given by $E_{I_i,I_j} = ||K_{I_i,I_j} - E_{I_i,Z} W^{-1} E'_{I_j,Z}||_F$

Theoretical Analysis (Method 2)

Theorem 5

We assume the kernel k satisfies the following property: $(k(a,b)-k(c,d))^2 <= C_X^k(||a-c||^2+||b-d||^2, \forall a,b,c,d$. For any kernel k, the partial approximation error E_{I_i,I_j} is bounded by

$$E_{I_i,I_j} <= \sqrt{2mC_X^k(e_{I_i}+e_{I_j})}+\sqrt{mC_X^k e_{I_i}}+\sqrt{mC_X^k e_{I_j}}+mC_X^k\sqrt{e_{I_i}e_{I_j}}||W^{-1}||_F,$$

where $e_{I_i}=\Sigma_{x_i\in X_{I_i}}||x_i-z_{c(i)}||^2$.

Theoretical Analysis (Method 2)

Proof Idea

We first define a partial approximation error for a sub-kernel Matrix. Suppose that the landmark set is $Z = z_{i=1}^{m}$, and the whole sample set X is partitioned into m disjoint clusters S_k's. Suppose each cluster has T samples. Repeat the following sampling process T times: at each time t, pick one sample from each cluster and denote the set of samples chosen at time t as X_{I_t}. Then $X = X_{I_1} \cup X_{I_2} \cup ... \cup X_{I_T}$, and the whole kernel matrix will be correspondingly decomposed into T^2 blocks, each of size m $\times$ m. Let K_{I_i,I_j}, and $E_{I_i,Z}$ be the m$\times$m similarity matrices defined on (X_{I_i}, X_{I_j}) and (X_{I_i}, Z), respectively, and $W \in R^{m \times m}$ the kernel matrix defined on Z. The partial approximation error is the difference between K_{I_i,I_j} and its Nystrom approximation under the Frobenius norm given by $E_{I_i,I_j} = ||K_{I_i,I_j} - E_{I_i,Z} W^{-1} E'_{I_j,Z}||_F$.

Theoretical Analysis (Method 2)

Theorem 6

The error of the Nystrom approximation is bounded by $E <= 4T\sqrt{mC_X^k eT} + mC_X^k Te||W^{-1}||_F$ where $T = max|S_k|$, and $e = \Sigma_{i=1}^{n}||x_i - z_{c(i)}||^2$.

Theoretical Analysis (Method 2)

Proof Idea

$\Sigma_{i,j=1}^{T}\sqrt{2mC_X^k(e_{I_i}+e_{I_j})} = \sqrt{2mC_X^k}\Sigma_{i=1}^{T}(\Sigma_{j=1}^{T}\sqrt{e_{I_i}+e_{I_j}})$

$\leq \sqrt{2mC_X^k}\Sigma_{i=1}^{T}(\sqrt{T}\sqrt{Te_{I_i}+\Sigma_{j=1}^{T}e_{I_j}} = 2T\sqrt{mC_X^k Te}.$

Similarly the second and the third terms can be simplified as:

$\Sigma_{i,j=1}^{T}\sqrt{mC_X^k e_{I_i}} = \sqrt{mC_X^k}\Sigma_{i=1}^{T}(\Sigma_{j=1}^{T}\sqrt{e_{I_i}}) <= T\sqrt{mC_X^k Te}.$

The last term of the expression can be summarized as:

$\Sigma_{i,j=1}^{T} mC_X^k\sqrt{e_{I_i}e_{I_j}}||W^{-1}||_F = mC_X^k||W^{-1}||_F(\Sigma_{i=1}^{T}\sqrt{e_{I_i}})^2 <=$

$mC_X^k||W^{-1}||_F Te.$

Theoretical Analysis (Method 2)

Note

The constant C_X^k for the kernel property depends on the choice of kernel taken in the underlying KPCA. Let us take the Gaussian Kernel $G(\alpha) = exp(-\alpha^2)$. By using the mean value theorem and triangular inequality, we have, for any $a, b, c, d \in R$, $(k(a,b)k(c,d))^2 = (G(||a-b||/\sigma)G(||c-d||/\sigma))^2 = (G'(\epsilon)/\sigma)^2(||a-b|| - ||c-d||)^2$. We also have by $(||ab||||cd||)^2 <= (||ac|| + ||bd||)^2 <= 2(||ac||^2 + ||bd||^2)$. Thus we can choose C_X^k as $max((2G'(\epsilon)/\sigma)^2)$, which is often bounded. (C_X^k is $1/(2\sigma^2)$ for the Gaussian kernel)

Table of Contents

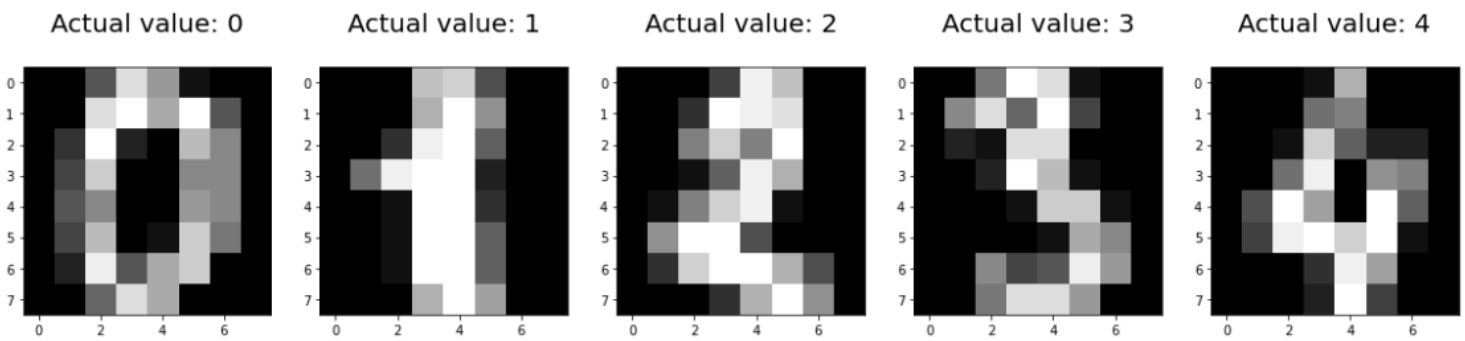
Computational Results

Dataset Information

```
Number of data points:  1797
Number of features:     64

X:
[[ 0.  0.  5. ...  0.  0.  0.]
 [ 0.  0.  0. ... 10.  0.  0.]
 [ 0.  0.  0. ... 16.  9.  0.]
 ...
 [ 0.  0.  1. ...  6.  0.  0.]
 [ 0.  0.  2. ... 12.  0.  0.]
 [ 0.  0. 10. ... 12.  1.  0.]]


-----------------------------------------Images in dataset as example-----------------------------
```



Computational Results

	Method	Accuracy	Time Taken
1	PCA	60.44%	0.009 s
2	Kernel PCA	78.96%	15.43 s
3	Nystrom KPCA	76.09%	0.22 s

Table of Contents

References

- Bishop, Christopher M. Pattern Recognition and Machine Learning. New York : Springer, 2006.
- Tripathy, BK., S Anveshrithaa, and Shrusti Ghela. Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization. 1st ed. CRC Press, 2021
- Alessandro Rudi, Raffaello Camoriano, Lorenzo Rosasco. Less is More: Nystr¨om Computational Regularization, arXiv, 2016.
- Kai Zhang, Ivor W. Tsang, James T. Kwok. Improved Nystrom Low-Rank Approximation and Error Analysis, 2008.
- Sanjiv Kumar, Mehryar Mohri, Ameet Talwalkar. Sampling Methods for the Nystrom Method, 2012.

Thank You