Iran University of Science and Technology

Dimensionality Reduction

Framework setup

Ali Golestani,Alireza Nikpoosh,Nargess Seifi

2023

Homework N° 1

Understanding the Concept

Large datasets are increasingly common and are often difficult to interpret. Dimensionality Reduction provides researchers many new techniques to reduce the dimensionality of such datasets, increasing the interpretability but at the same time minimizing information loss.

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of datasets. It creates new uncorrelated variables that successively maximize variance.

Initial Framework

Almost all machine learning/pattern recognition problems work in identical frameworks. In this section, we will review the elements of this framework.

Review:

Let $D = \{(X_i, y_i)_1^n\}$ be an arbitrary dataset with its labels.
Every dataset is a subset of a linear space $\mathbb{V}$ over a field F.
As already mentioned, there exists (just accept it!) a function $C : \mathbb{V} \to Y$ that for every entry X returns a label from the set of all possible labels Y. The goal of machine learning (at least most of it) is to determine this function.
Note that the function C is defined over the set D and not $\mathbb{V}$ necessarily; meaning, $C(D) \subset Y$.

A very very very short note on Isomorphisms

The concept of Isomorphism is of great importance in linear algebra. We also need some of its notions to tackle the dimensionality problem further ahead.

Review: Isomorphism and Isomorphic structures

Let $\mathbb{V}_1$ and $\mathbb{V}_2$ be two vector spaces over fields $\mathbb{K}_1$ and $\mathbb{K}_2$ respectively.
An Isomorphism is a linear transformation $\alpha : \mathbb{V}_1 \to \mathbb{V}_2$ such that $\alpha(kv_1 + v_2) = k\alpha(v_1) + \alpha(v_2)$.

Query: Let $\mathbb{V}$ be a finite dimensional space over the field $\mathbb{K}$

- Show that $\mathbb{V}$ is isomorphic to $\mathbb{K}^n$ where $dim(\mathbb{V}) = n$
- Find an isomorphism $\alpha : \mathbb{V} \to \mathbb{K}^n$ where $ker(\alpha) = \{0\}$
- Is α a Homomorphism?
- What can one say about Isomorphisms, studying only the properties of kernels?

Recenter Data

The goal of this task is to relocate the data in a way that maximizes the interpretability with minimum effort.

Before we dive into the main problem, we shall review some fundamental concepts.

Review: Center of Data

The "center" of a data set is a way of describing location. The two most widely used measures of the "center" of the data are the mean (average) and the median. In this project, let the center of data be the mean $\bar{X}_n = \frac{1}{n}\sum_1^n X_i$.

Query: For an arbitrary dataset of $dim = 2$

- Find the center of the data $\bar{X}_n$.
- Shift the Center in the way that it lies on the origin $\mathbb{O}(0, 0)$.
- Find a Homomorphism $T : \mathbb{V} \to \mathbb{V}$ such that $T(\bar{X}_n) = \mathbb{O}(0, 0)$ and all other points are positioned respectively.
  Call the relocated dataset D'.
- Find $ker(T)$
- Is T an Isomorphism?

From now on, for more interpretability, we continue working on D' instead of D.

Projective Spaces

In the previous section, you have proved that every finite-dimensional space $\mathbb{V}$ over the field K is Isomorphic to K^n. Projective spaces are very simmilar to subspaces of K^n.

Query: Projective Spaces

Let $K_1, K_2, ..., K_n$ be n different fields.

- Show that $\Pi_1^n K_i$ is also a field where ΠK_i is the Cartesian product of K_is.
- Let $(k_1, k_2, ..., k_n)$ be a vector in $\Pi_1^n K_i$. Define $\alpha_j : \Pi_1^n K_i \to K_j$ where $\alpha_j(k_1, k_2, ..., k_n) = k_j$. Show that α_j is an isomorphism.
- Find $ker(\alpha_j)$
- Is α_j a Homomorphism?

Best Fit

In this section, we are going to find the line that best fits the dataset distance-wise.

Query: Linear semi-Regression

- Consider an arbitrary line $C = \alpha X$.
- Show that the line C is a subspace of $\mathbb{V}$ containing D'
- Project all points on the line C using projective transformation mentioned in previous section.
- Find a basis $\mathbb{B}$ for the projective space C
- Show that every vector on line C is of the form $\sum_1^n t_i b_i$ where $b_i \in \mathbb{B}$ and $t_i \in F$.
- Find a basis for $\mathbb{V}$

Query:

- Show that for every two arbitrary vector spaces $\mathbb{V}_1$ and $\mathbb{V}_2$ with $\mathbb{B}_1$ and $\mathbb{B}_2$ as basis respectively, if $\mathbb{B}_1$ and $\mathbb{B}_2$ are homomorphic, then $\mathbb{V}_1$ and $\mathbb{V}_2$ are homomorphic.
- Find an isomorphism from $\mathbb{B}'$ to $\mathbb{B}$

- Show that V is isomorphic to C.

Quantify How Good it Fits

To quantify how good a line fits the data, PCA projects all the data points on the line.

Query:

- Consider the vector space C in previous section. Show that $\mathbb{V}$ is isomorphic to C.
- Let $d_C(X_i)$ be the distance between point X_i and line C. Find $S = \sum_1^n d_c(X_i)$
- Find a line C^* minimizing the sum S, using only rotation about the origin.
- is this line unique?