

A Study of the ISOMAP Algorithm and Its Applications in Machine Learning

Lucas Oliveira David

Department of Computing
Universidade Federal de São Carlos

Supervisor: Dr. Alexandre Luis Magalhães Levada

December 2015

Declaration

I hereby declare that I carried out the work entitled “A Study of the ISOMAP Algorithm and Its Applications in Machine Learning”, which is being submitted to Universidade Federal de São Carlos, under the supervision of Dr. Alexandre Luis Magalhães Levada, as the partial fulfillment of the requirements for the award of the Degree of Bachelor in Computer Science in the Department of Computing. I solemnly declare that to the best of my knowledge, no part of this report has been submitted here or elsewhere in a previous application for award of a degree. All sources of knowledge used have been duly acknowledged.

.....

Lucas Oliveira David

407917

Approval

“A Study of the ISOMAP Algorithm and Its Applications in Machine Learning,” written by Lucas Oliveira David as the partial fulfillment of the requirements for the award of the Degree of Bachelor in Computer Science in the Department of Computing, has been approved and accepted by the following:

.....
Dr. Alexandre Luis Magalhães Levada

.....
Date

.....
Dr. ...

.....
Date

.....
Dr. ...

.....
Date

Abstract

This project aims to study the foundations of nonlinear dimensionality reduction through manifold learning with the algorithm known as Isometric Feature Mapping (ISOMAP) and observe the application of the algorithm in practical experiments.

Firstly, the relevant background will be presented in order to contextualize the reader and provide concepts that are closely related to dimensionality reduction, manifold learning and, of course, the ISOMAP algorithm. We will then proceed to study linear dimensionality reduction: algorithms, applications and limitations. Finally, manifold learning with ISOMAP will be covered, from the algorithm's concept to its extensions and limitations.

In chapters 4 and 5, experiments created during the project are presented for both observation and comparison between the methods. The experiments were developed and executed in the following computational environment: linux Ubuntu 15.10 64 bits, Intel Core i7-4700MQ CPU 2.40GHz × 8, 16 GB of RAM computational environment. All the artifacts, (e.g., source-code, docs, experiments) can be found in the repository <https://github.com/lucasdavid/manifold-learning> and are licensed under the MIT License.

Contents

1	Introduction	1
2	Relevant Background	2
2.1	Data set	2
2.1.1	Data set as a collection of vectors in the \mathbb{R}^n	3
2.1.2	Modern Problems and Applications	4
2.2	Probability Theory	5
2.2.1	Feature Standardization	5
2.2.2	Centering Matrix	5
2.2.3	Variance	6
2.2.4	Covariance	7
2.2.5	Correlation	8
2.3	Numerical Analysis	8
2.3.1	Eigenvalues and Eigenvectors of a Matrix	8
2.3.2	Spectral Decomposition of a Matrix	9
2.3.3	Singular Value Decomposition	9
2.4	Topology	9
2.4.1	Manifolds	9
2.5	Graph Theory	11
2.5.1	Graphs	11
2.5.2	Related Problems	13
2.6	Machine Learning	17
2.6.1	Machine Learning Algorithms	17
2.6.2	Support Vector Machine: An Example of Supervised Learning	18
2.6.3	Multi-class Classification	22

2.6.4	Evaluating learners	23
2.6.5	Examples of Learning	25
3	Linear Dimensionality Reduction	27
3.1	Principal Component Analysis	28
3.1.1	Study of the PCA Algorithm	29
3.1.2	Formalization of the PCA Algorithm	30
3.2	Multidimensional Scaling	31
3.2.1	Study of the MDS	31
3.2.2	Formalization of the Multidimensional Scaling Method	33
3.3	Evaluating Reductions	34
3.4	Classification and Regression Over Linearly Reduced Data Sets	35
3.4.1	K Data Set	35
3.4.2	The Iris Flower Data Set	36
3.4.3	The Digits Data Set	37
4	Nonlinear Dimensionality Reduction	39
4.1	The ISOMAP Algorithm	40
4.1.1	Study of the ISOMAP Algorithm	41
4.1.2	Formalization of the ISOMAP Algorithm	43
4.1.3	Computational Complexity	43
4.1.4	Extensions	45
4.1.5	Evaluating Reductions	49
4.1.6	Applicability and Limitations	50
4.2	Classification and Regression Over Data Sets Reduced with ISOMAP	53
4.2.1	The Swiss Roll Data Set	53
4.2.2	The Iris Flower Data Set	54
4.2.3	The Glass Data Set	55
4.2.4	The Digits Data Set	56
4.2.5	The Dermatology Data Set	57
4.2.6	The Leukemia Data Set	58
4.2.7	The WDBC Data Set	59
4.2.8	Diabetes Data Set	60
Final Considerations		61

List of Figures

2.1	The data set ILPD mapped onto \mathbb{R}^n , where each of its features is an axis, except for $S := \{1, 2\}$, which was represented by the vertices' colors.	4
2.2	Stereographic projection applied to Earth.	10
2.3	The Les Miserables graph.	12
2.4	A tree extracted (a subgraph) from the Les Miserables graph.	13
2.7	A SVM classifier projecting a hyperplane that perfectly separates two classes of samples. [1]	19
2.8	Projection of samples from the \mathbb{R} to the \mathbb{R}^2 , allowing SVM to find a hyperplane that perfectly separates both classes. [2]	22
2.9	graphic representation of a data set generalization by a linear (orange) and a nonlinear model (green).	25
2.10	Confusion matrix of a SVM with $C = 100$, $gamma = .01$ and rbf kernel when predicting samples from the Iris flower data set.	26
3.1	The data set $K \in \mathbb{R}^2$	27
3.2	The principal components of K	29
3.3	The PCA algorithm reducing K to 2 and 1 dimension, respectively.	35
3.4	The Iris Flower data set and its reductions to 3, 2, and 1 dimensions, respectively.	36
3.5	Digits data set reduced to 10, 3, 2 and 1 dimension, respectively.	37
4.1	The Swiss-roll manifold and its reductions to 3, 2 and 1 dimensions, respectively, using the PCA algorithm.	39
4.2	The data set S , consisting of 1000 samples and 3 features.	41
4.3	The graph H	42

4.5	The embedded swiss-roll data set by the original and scikit-learn's implementation.	46
4.7	The ISOMAP applied on a noisy data set.	52
4.8	The Swiss Roll data set and its reductions to 2 and 1 dimensions, respectively.	53
4.9	The Iris data set and its reductions to 3, 2 and 1 dimensions, respectively.	54
4.10	The Glass data set and its reductions to 3, 2 and 1 dimensions, respectively.	55
4.11	Digits data set and its reductions to 3, 2 and 1 dimension, respectively.	56
4.12	The Dermatology data set and its reductions to 3, 2 and 1 dimensions, respectively.	57
4.13	The Leukemia data set and its reduction to 3, 2 and 1 dimensions, respectively.	58
4.14	The WDBC data set and its reduction to 3, 2 and 1 dimensions, respectively.	59
4.15	The Diabetes data set and its reduction to 3, 2 and 1 dimensions, respectively.	60

List of Tables

2.1	The first three samples of the Iris flower data set.	3
2.2	The first three samples of the Indian Liver Patient Dataset (ILPD)	3
2.3	A data set with 8200 samples and 100 features.	5
2.4	Example of confusion matrix for a data-set with four different classes.	24
3.1	Covariance between the components of K	28
3.2	Description of predictions and reduction performance for K	36
3.3	Description of predictions and reduction performance for Iris flower.	37
3.4	Description of predictions and reduction performance for Digits. . .	38
4.1	Regression accuracy and reduction performance for the Swiss-roll data set.	40
4.2	Listing of time spent on each step of the ISOMAP algorithm. . . .	44
4.3	Timing the implemented ISOMAP and scikit-learn's implementation.	44
4.4	Description of predictions and reduction performance for Swiss Roll.	53
4.5	Description of predictions and reduction performance for Iris. . . .	54
4.6	Description of predictions and reduction performance for Glass. . .	55
4.7	Description of predictions and reduction performance for Digits. . .	56
4.8	Description of predictions and reduction performance for Dermatology.	57
4.9	Description of predictions and reduction performance for Leukemia data set.	58
4.10	Description of predictions and reduction performance for Leukemia data set.	59
4.11	Description of predictions and reduction performance for Diabetes data set.	60

Chapter 1

Introduction

Throughout the years, machine learning techniques have grown popular between both academical and the corporative sectors. Their vast applications and promising results [3] indubitably contributed to our current scenario where not only computer scientists or mathematicians, but engineers, psychologists and many other groups have taken interest [4] on how to adapt and apply these studies to their own problems.

Machine learning can help us to understand large amount of data and take decisions based on it. To achieve this, however, we must first find ways to effectively (and efficiently) extract the information that lies within the data.

Many different machine learning algorithms have been developed during the this and the last century. Between those, many could successfully generalize low dimensional data. [5] In the other hand, problems of our world are often too complex and may be represented by high dimensional data. For example, images, sounds or text documents can be expressed as vectors of the \mathbb{R}^n , where each element corresponds to a pixel, wave signal or character, respectively. When analyzing these problems, we observed that many of the algorithms would often become unstable. **Dimensionality reduction** (or DR) then quickly became a key concept for minimizing the data size while maintaining its meaning.

Finally, dimensionality reduction has evolved into an extensive area. Nowadays, DR is not only applied to reduce the data size, specifically, but often employed in data visualization, noise reduction and for many other purposes.

Chapter 2

Relevant Background

2.1 Data set

In the context of Computer Science, very often our goal is to develop machines that can assist or automate the process of solving real-world problems. Firstly, however, we must find ways to express these problems numerically.

Although the recurrent usage of the term in scientific work, there is not a clear definition established. It is possible, however, to observe the regular presence of four related features: grouping, content, relatedness e purpose. [6] For the scope of this work, the term data set is invariably associated with the idea of a collection of samples. Each sample is a sequence of features, where the i -th feature of all instances belong to a same set of symbols f_i .

Succinctly, consider S a sequence of samples and $F := \{f_i \mid f_i \text{ is a set of symbols}\}$. Then, the dataset D is defined as:

$$D := [d_{ij}] \mid d_{ij} \in f_j, \forall i \in [1, |S|], \forall j \in [1, |F|]$$

Example of a canonical data set

The table bellow illustrates an examples of data set, where each row represents a **sample**, and each column a **feature**.

	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.1	3.5	1.4	0.2	I. setosa
2	4.9	3.0	1.4	0.2	I. setosa
3	4.7	3.2	1.3	0.2	I. setosa
...

Table 2.1: The first three samples of the Iris flower data set.

Iris flower is an example of data set broadly used in the machine learning demonstrations, being usually interpreted as a classification problem where the feature *Species* will be learned from its adjacent features. In that scenario, *Species* is denominated **target feature**.

2.1.1 Data set as a collection of vectors in the \mathbb{R}^n

A data set can have each one of its nominal features enumerated. I.e., mapped to an element of \mathbb{N} . Such set could then be expressed as a collections of vectors in \mathbb{R}^n . Consider the data set bellow:

	Age	Gen.	TB	DB	Alk.	Sgpt	Sgot	TP	ALB	A/G	S
1	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
2	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
3	62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
...

Table 2.2: The first three samples of the Indian Liver Patient Dataset (ILPD)

Composed by 583 samples and 11 features, the data set ILPD has a nominal feature *Gender* := {Male, Female}. *Gender* can, of course, be mapped on {0, 1}. ILPD can finally be expressed by the figure bellow:

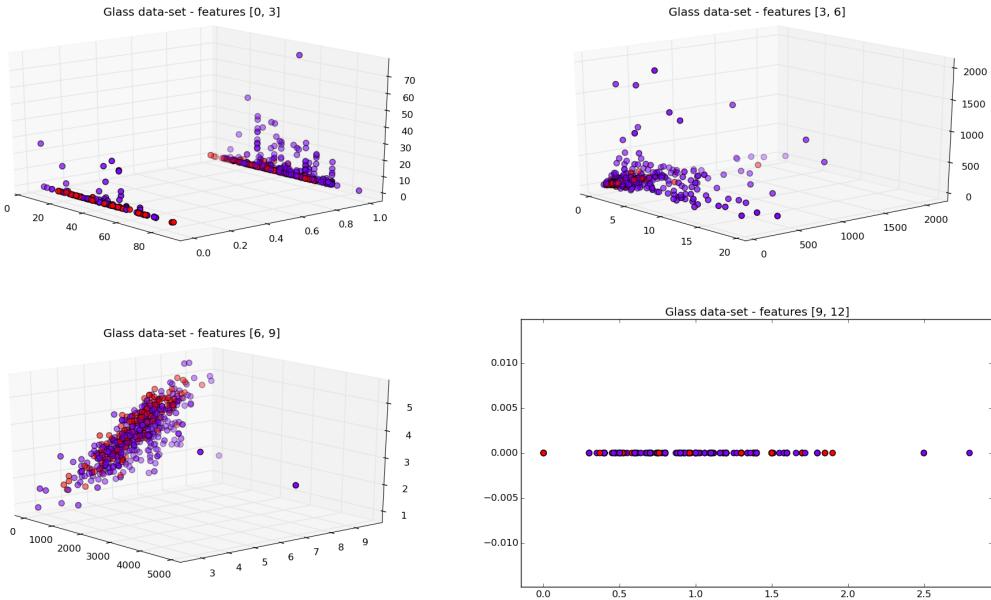


Figure 2.1: The data set ILPD mapped onto \mathbb{R}^n , where each of its features is an axis, except for $S := \{1, 2\}$, which was represented by the vertices' colors.

Considering the many graphs required to display the data set, it is quite difficult to identify a plausible distribution for ILPD. We define here our first encouragement towards the study of dimensionality reduction: the identification of the most significant features and plotting of those might result on simpler and more intuitive representations.

2.1.2 Modern Problems and Applications

Differently from Iris flower or ILPD data set, data sets associated with modern problems are often very dense, i.e., data sets containing many samples and/or features. Although the high number of samples is essentially benefic, a high number of features might be irrelevant or even unconstructive to the learning process. [7]

	A	B	...	AAAV
1	0.0111486888670454	-0.01541263850539861	...	0.007440367302352156
2	0.03016080450207878	0.1772161342899135	...	0.01309011094914101
...
8200	0.02680808496910305	-0.0320375843317954	...	0.1772161342899135

Table 2.3: A data set with 8200 samples and 100 features.

In many cases, there are indicatives that the data set lie near a lower-dimensional manifold embedded in the \mathbb{R}^n . [8] For the data set above, in special, the \mathbb{R}^{100} . A second encouragement can then be set: it is possible that the data set might be shrunk by combining similar (linearly dependent) features or eliminating the ones that poorly contribute towards the learning process. In order to do this, one must be able to qualify the “contribution” of each feature or even identify dependencies between features.

2.2 Probability Theory

2.2.1 Feature Standardization

Many of the methods ahead will require the data set to be centered in the origin. To center a data set X , we build a data set X' s.t. each column has zero mean and it is contracted by its standard deviation:

$$X'_{.j} = \frac{X_{.j} - \mu}{\sigma}, \text{ where}$$

1. $X_{.j}$ is the j-th column of the matrix X .
2. μ is the mean of $X_{.j}$.
3. σ is the standard deviation of $X_{.j}$.

2.2.2 Centering Matrix

The symmetric matrix H is named the **centering matrix** when the multiplication of it by a vector X produces the same effect of subtracting the mean of the

components from each component of X . H is defined as:

$$H = I_n - \frac{1}{n} \mathbf{1}\mathbf{1}^T, \text{ where:}$$

1. I_n is the identity matrix of order n .
2. $\mathbf{1}$ is the column vector of 1's.

2.2.3 Variance

Variance is the measure which describes how far the samples in a given set X vary. For the scope of this project, only discrete probabilities will be considered. That is, if X represents a variate with known distribution $P(x)$, where $P(x) = k \in \mathbb{R}^+, \forall x \in X$ and $\sum_{x \in X} P(x) = 1$, and μ is the population mean of X , then: [9]

$$\text{Var}(X) = \frac{1}{n} (X - \mu) \cdot (X - \mu) = \frac{1}{n} \sum_{x \in X} (x - \mu)^2$$

Remark 2.2.1 If $\text{var}(X) = 0$, then it can be easily asserted that all variables in X assume the exact same value by using the elementary properties of the inner product.

Example 2.2.1 If $X = \{1, 2, -2, 4\}$ and $\mu = \frac{1}{4} \sum X_i = \frac{1+2-2+4}{4} = 1.25$, then

$$\begin{aligned} \text{var}(X) &= \frac{1}{4} \sum (X_i - \mu)^2 \\ &= \frac{(1 - 1.25)^2 + (2 - 1.25)^2 + (-2 - 1.25)^2 + (4 - 1.25)^2}{4} \\ &= 4.6875 \end{aligned}$$

Example 2.2.2 The variance of the Sepal length feature in the Iris flower data set can be calculated as:

$$\begin{aligned} \text{var}(X) &= \frac{1}{150} \sum (X_i - \mu)^2 \\ &= \frac{1}{150} [(5.1 - 5.84)^2 + (4.9 - 5.84)^2 + \dots + (5.9 - 5.84)^2] \\ &= \frac{102.17}{150} = .681122 \end{aligned}$$

2.2.4 Covariance

The covariance measures the variance of two random variates in respect to each other. Formally, if X and Y are two given random variates with known mean population distribution μ_X and μ_Y , respectively, then

$$\sigma(X, Y) = \frac{1}{n}(X - \mu_X) \cdot (Y - \mu_Y)$$

Simply putting, the covariance of two random variables X and Y can be interpreted as one of the following behaviors:

$$\sigma(X, Y) = \begin{cases} \sigma_{xy} > 0 \implies X \text{ tends to increase as } Y \text{ increases.} \\ \sigma_{xy} < 0 \implies X \text{ tends to increase as } Y \text{ decreases.} \\ \sigma_{xy} = 0 \implies X \text{ and } Y \text{ are completely unrelated.} \end{cases}$$

Remark 2.2.2 For a random variate X , $\text{var}(X) = \sigma(X, X)$.

Covariance Matrix of Features in a Centered Data Set

Let D be a data set, H the centering matrix and $HD_{\cdot i}$ the i -th feature column of the centered data set HD , the covariance between each one of its features can be represented by the matrix:

$$\begin{aligned} \Sigma &= [\sigma_{xy}]_{n \times n} \\ &= \begin{bmatrix} \sigma(HD_{\cdot 0}, HD_{\cdot 0}) & \sigma(HD_{\cdot 0}, HD_{\cdot 1}) & \cdots & \sigma(HD_{\cdot 0}, HD_{\cdot n-1}) \\ \sigma(HD_{\cdot 1}, HD_{\cdot 0}) & \sigma(HD_{\cdot 1}, HD_{\cdot 1}) & & \sigma(HD_{\cdot 1}, HD_{\cdot n-1}) \\ \vdots & & & \vdots \\ \sigma(HD_{\cdot n-1}, HD_{\cdot 0}) & \sigma(HD_{\cdot n-1}, HD_{\cdot 1}) & \cdots & \sigma(HD_{\cdot n-1}, HD_{\cdot n-1}) \end{bmatrix} \\ &= \frac{1}{n}(HD)^T HD \\ &= \frac{1}{n}D^T H^T HD \\ &= \frac{1}{n}D^T HD \end{aligned}$$

Where H is the **centering matrix**.

2.2.5 Correlation

The correlation is a measure of the direction and strength of a linear relationship among variables. Let X and Y be two random variables, $r(X, Y)$, i.e., the correlation between X and Y is defined as: [9]

$$r(X, Y) = \frac{\sigma(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of the variables X and Y , respectively. Of course, $r(X, Y) \in [-1, 1]$.

Remark 2.2.3 If X and Y are two random variables and $\sigma(X, Y) = 1$, $X = Y$.

Correlation Matrix of Features in a Centered Data Set

Let D be a data set, H the centering matrix and $HD_{\cdot i}$ the i -th feature column of the centered data set HD . The correlation between each one of its features can be represented by the matrix:

$$\text{corr}(D) = \begin{bmatrix} 1 & \frac{\sigma(HD_{\cdot 0}, HD_{\cdot 1})}{\sigma_{HD_{\cdot 0}} \sigma_{HD_{\cdot 1}}} & \dots & \frac{\sigma(HD_{\cdot 0}, HD_{\cdot n-1})}{\sigma_{HD_{\cdot 0}} \sigma_{HD_{\cdot n-1}}} \\ \frac{\sigma(HD_{\cdot 1}, HD_{\cdot 0})}{\sigma_{HD_{\cdot 1}} \sigma_{HD_{\cdot 0}}} & 1 & \dots & \frac{\sigma(HD_{\cdot 1}, HD_{\cdot n-1})}{\sigma_{HD_{\cdot 1}} \sigma_{HD_{\cdot n-1}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma(HD_{\cdot n-1}, HD_{\cdot 0})}{\sigma_{HD_{\cdot n-1}} \sigma_{HD_{\cdot 0}}} & \frac{\sigma(HD_{\cdot n-1}, HD_{\cdot 1})}{\sigma_{HD_{\cdot n-1}} \sigma_{HD_{\cdot 1}}} & \dots & 1 \end{bmatrix}$$

2.3 Numerical Analysis

2.3.1 Eigenvalues and Eigenvectors of a Matrix

Given a matrix $A \neq 0 \in \mathbb{R}^{2n}$, a vector $v \in \mathbb{R}^n$ is said to be an **eigenvector** of A if the multiplication Av does not change the direction of v ; that is:

$$\exists \lambda \in \mathbb{R} \mid Av = \lambda v, \text{ where}$$

λ is the **eigenvalue** associated to the eigenvector v .

2.3.2 Spectral Decomposition of a Matrix

If $[M]_{n \times n}$ is a symmetric matrix of rank n and admits n pairs of eigenvalues λ and eigenvectors $[V]_{n \times n} = [v_0, v_1, v_2, \dots, v_{n-1}]$, such that $V^T V = 1$, then

$$MV = V\lambda$$

$\lambda = \text{diag}(\sigma_0, \sigma_1, \dots, \sigma_{n-1})$ is the diagonal matrix, where σ_i is the eigenvalue associated to the eigenvector v_i . [10] Furthermore, V 's columns are linear independent, hence V is invertible.

$$\begin{aligned} MV &= V\lambda \\ MVV^T &= V\lambda V^T \\ M &= V\lambda V^T \end{aligned}$$

2.3.3 Singular Value Decomposition

If $M \in \mathbb{R}^{m \times n}$, then $\exists U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ and $\Sigma = \text{diag}(\sigma_0, \dots, \sigma_{n-1})$ conditioned to $\sigma_i \geq \sigma_{i+1} \geq 0, \forall i \in [0, n)$ s.t. [11]

$$M = U\Sigma V^T$$

Theorem 2.3.1 If $A = U\Sigma V^T$, $AA^T = U\Sigma^2 U^T$ and $A^T A = V\Sigma^2 V^T$.

Proof

$$\begin{aligned} A^T A &= (U\Sigma V^T)^T (U\Sigma V^T) \\ &= V\Sigma^T U^T U\Sigma V^T \\ &= V\Sigma\Sigma V^T \\ &= V\Sigma^2 V^T \end{aligned}$$

Proving $AA^T = U\Sigma^2 U^T$ is analogous to the above.

2.4 Topology

2.4.1 Manifolds

Intuitively, n-dimensional topological manifolds are sets that are “locally Euclidean”. [12] In other words, they can be decomposed into sub sets that can

be mapped to the \mathbb{R}^n .

Formally, a set M is said to be a **n-dimensional topological manifold** $\iff M$ is a paracompact Hausdorff topological space $| \forall p \in M, p \in U_p$, where U_p is an open set that is homeomorphic to an open set V_p of the Euclidean space \mathbb{R}^n . [12]

From here, we will use the word manifold to refer to the n-dimensional topological manifold.

Charts

The pair (U_i, ϕ_i) is called a **coordinate chart** or **chart** on M if $U \in M$ and ϕ_i is a **homeomorphism** such that $\phi_i(U_i) = V_i \subseteq \mathbb{R}^n$ [?].

Atlas

A set $A = \{(U_i, \phi_i)\}_{i \in A}$ is said to be an **atlas** on a manifold M if $\cup_{i \in A} U_i = M$ [?].

Example 2.4.1 *The \mathbb{R}^n is, directly, a manifold.*

Example 2.4.2 *A n-dimensional sphere is a manifold. Earth, in special, is a 3-dimensional sphere and its stereographic projection is its mapping to the \mathbb{R}^2 .* [13]

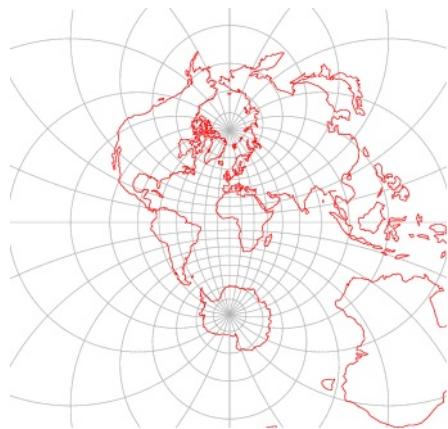


Figure 2.2: Stereographic projection applied to Earth.

2.5 Graph Theory

2.5.1 Graphs

Let G be the pair (X, U) . G is defined as a **graph**, [14] where

1. X is a set of objects called **vertices**.
2. U is a family of elements $u_i \in X \times X$ called arcs.

Basic Concepts [14]

Multiplicity If $G = (X, U)$ and $(x, y) \in X \times X$, the multiplicity $m_g^+(x, y)$ of x, y is defined to be the number of arcs with initial endpoint x and terminal endpoint y . Furthermore:

1. $m_g^-(x, y) = m_g^+(y, x)$
2. $m_G(x, y) = m_g^+(x, y) + m_g^-(x, y)$

Degree If $G = (X, U)$ is a graph and $x \in X$, the degree $d(x)$ of x is defined [15] as $d(x) = 2n_s + n_n$, where n_s is the number of arcs self-incident at x (i.e., $\{(x, x)\}$) and n_n is the number of arcs incident at x .

Adjacency Matrix If $G = (X, U)$, $X = \{x_1, x_2, \dots, x_n\}$, define the adjacency matrix $A = [a_{ij}]_{n \times n}$ associated with graph G , where $a_{ij} = m_g^+(x_i, x_j)$.

Further Specifications

Undirected graph “All graphs are directed, but sometimes the direction need not be specified [14]”.

Let $G = (X, U)$ be a graph and $u_k \in U \mid u_k = (a, b)$. The element $e_i = [a, b]$ can be defined as the **edge** that links a to b without specifying direction. Finally, define the **undirected graph** H as (X, E) , where $E = \{e_i\}$ is the set of edges created from U .

Figure 2.3 shows the graph **Les Misérables**, where each node is a character and each arc links two characters that have shared stage at some point during the play.

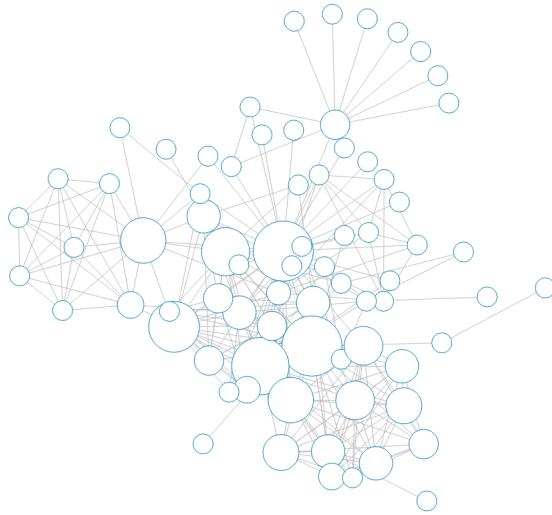


Figure 2.3: The **Les Miserables** graph.

Complete graph A graph $G = (X, U)$ is said to be complete if

$$\forall(x, y) \in X, x \neq y, m_G(x, y) \geq 1$$

Remark 2.5.1 Let $G = (X, U), G_1 = (X, E)$ and $n = |X|$. G is called the complete graph K_n if

$$\forall(x, y) \in X, \exists e \in E \mid e = [x, y]$$

Weighted graph Let $G = (X, U)$ be a graph and $w: U \rightarrow \mathbb{R} \mid w(u) = w_u$ be the weighted associated with arc u . G is said to be a weighted graph.

Euclidean graph If $G = (X, U)$ and $W = \{w_u, \forall u \in U\} \subset \mathbb{R}$, G is said to be an euclidean graph if w_u corresponds to the euclidean distance between the vertices connected by u in a specified embedding.

Tree Let G be the graph (V, E) such that G is connected and admits no cycles. G is said to be a **tree** [14].

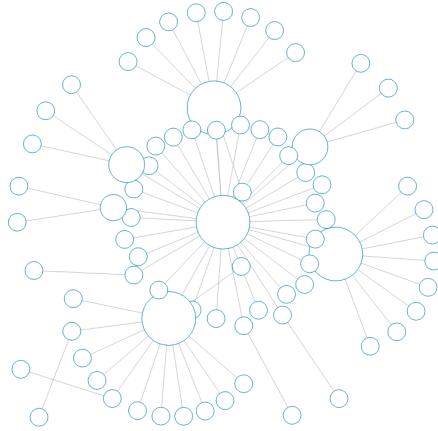


Figure 2.4: A tree extracted (a subgraph) from the **Les Miserables** graph.

2.5.2 Related Problems

Nearest-Neighbor Search

Let $G = (X, U)$ be a weighted graph, where $\forall u \in U, \exists w_u \in \mathbb{R}$, i.e., the weight or **length** of the arc u , and $n: U \rightarrow \{0, 1\}$ a definition of **nearness** in G . The nearest-neighbor search is a optimization problem that consists of finding a subgraph $H = (X, F \subseteq U) \mid f \in F \iff n(f) = 1$.

A Generic Algorithm for NN Search

1. $F_x := \emptyset, \forall x \in X$
2. $F_x := F_x \bigcup_{\forall p \in U_x} \begin{cases} \{p\}, & \text{if } n(p) = 1 \\ \emptyset, & \text{if } n(p) = 0 \end{cases}$
3. $H := (X, F), F = \bigcup_{x \in X} F_x$

Let's consider two (between many) specifications of this algorithm:

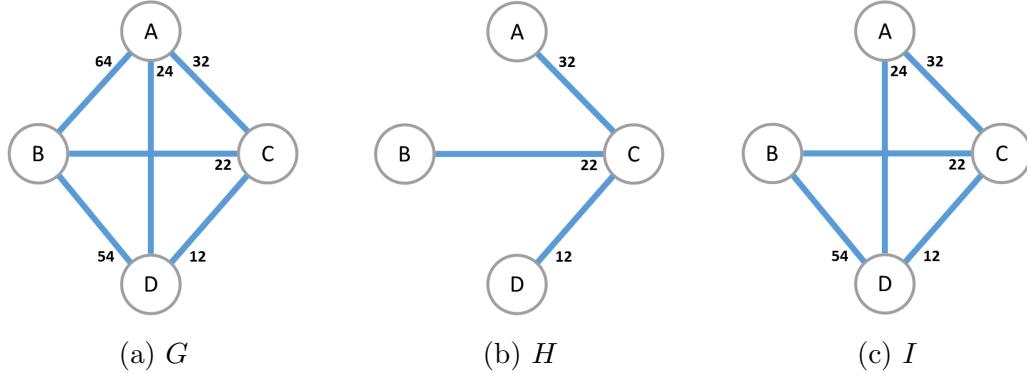
K-Nearest Neighbor Search (K-NN) Fixed $k \in \mathbb{N}$ and $U_x \subseteq U$, where U_x is the set of all arcs with x as initial endpoint, define:

$$n(u_x) := \begin{cases} 1, & \text{if } w_{u_x} \leq w_p, \forall p \in U_x - F_x \text{ and } |F_x| \leq k \\ 0, & \text{otherwise.} \end{cases}$$

ϵ -Nearest Neighbor (ϵ -NN) Fixed $\epsilon \in \mathbb{R}$, define:

$$n(f) := \begin{cases} 1, & \text{if } w_f \leq \epsilon \\ 0, & \text{otherwise} \end{cases}$$

Example 2.5.1 If $k = 1$ and $\epsilon = 60$, the graph G , the sub-graph H found from K -Nearest neighbor algorithm and the sub-graph I found from the ϵ -Nearest neighbor are defined as follows:



Single-pair Shortest-path Problem [16]

Let $G = (X, U)$ be a weighted graph, a path $p(x, y) = (u_0, u_1, u_2, \dots, u_{p-2}, u_{p-1}, u_p) \mid u_0 = (x, -), u_p = (-, y), u_i \in U, \forall i \in [0, p]$, and the weight of p be given by $w(p) = \sum_i^p w(u_i)$. The shortest-path weight between two vertices x and y is defined as:

$$\sigma(x, y) = \begin{cases} \min\{w(p(x, y))\}, & \text{if } p(x, y) \text{ exists,} \\ \infty, & \text{otherwise.} \end{cases}$$

while $p(x, y)$ is the **shortest path** from x to y .

The shortest-path between a vertex $x_0 \in X$ and all other vertices can be expressed by the tree $S = (X, F), F \subseteq U$ named **shortest-path tree**.

Dijkstra's Algorithm [16]

Let $G = (V, E)$ be a directed weighted graph and $v_0 \in V$ the initial vertex, the shortest-path from v_0 to all other vertices can be found through Dijkstra's algorithm:

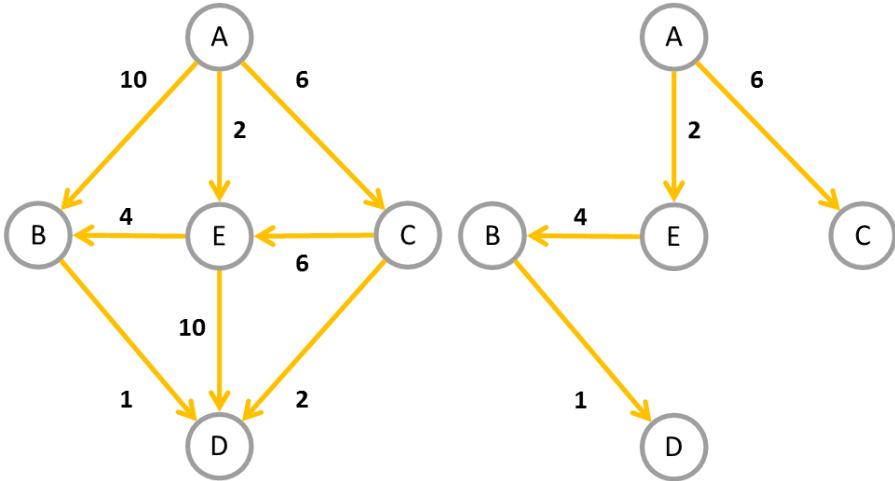
```

Dijkstra(V, E, w) :
    for v ∈ V
        predecessor(vi) ← None
        d(vi) ← ∞
    d(v0) ← 0
    S ← ∅
    Q ← priority-queue (V, d)
    while Q ≠ ∅
        u ← pop_min(Q)
        for each v ∈ V | (u, v) ∈ E
            if d(v) > d(u) + w(u, v)
                d(v) ← d(u) + w(u, v)
                predecessor(v) = u

```

Remark 2.5.2 Dijkstra's algorithm is based on two strategies: greedy, when extracting the vertex u such that $d(u)$ is minimum, and dynamic programming, during edge relaxation (comparison and update of the $d(v)$ and $\text{predecessor}(v)$ values).

Example 2.5.2 Let G be the weighted graph as defined in figure 2.6a. The shortest-path between A and all other vertices is described by the tree S illustrated in figure 2.6b.



(a) The weighted graph G .

(b) The shortest-path tree S .

All-pairs Shortest-path Problem

Similarly from the previous problem, all-pairs shortest-path is also a search problem that aims to find the predecessors of each vertex $v_i \in V$. The difference is that there is no constraint regarding the initial vertex, i.e., one is interested in finding the shortest routes from every vertex to every other vertex in the graph.

Clearly, all-pairs shortest-path can be solved by executing the Dijkstra's Algorithm $|V|$ times, using every $v_i \in V$ as initial vertex.

Floyd-Warshall Algorithm [17]

Another solution for the all-pairs shortest-path problem is Floyd-Warshall algorithm, which is entirely based on dynamic programming:

```

 $Floyd - Warshall(w, n):$ 
    for  $i \in V$ 
        for  $j \in [1, n]$ 
             $d(i, j) = w(i, j)$ 
             $predecessor(i, j) \leftarrow None$ 

    for  $k \in [1, n]$ 
        for  $j \in [1, n]$ 

```

```

for i ∈ [1, n]
    if d(i, j) > d(i, k) + d(k, j)
        d(i, j) ← d(i, k) + d(k, j)
        predecessor(i, j) ← k

```

Now, matrix d contains the weights sum for the paths between all vertices of the graph. These paths can be constructed by:

```

Path(i, j):
    if predecessor(i, j) = None
        return [i, j]
    return Path(i, pred[i, j]) ∪ Path(pred[i, j], j)

```

2.6 Machine Learning

“Learning is the improvement of performance in some environment through the acquisition of knowledge resulting from experience in that environment.” [18]

“Machine learning is the area in Computer Science that has as goal the projection and implementation of machines that have the ability to learn autonomously”. [19] In other words, the creation of machines that are able to recognize patterns in an environment and interpret those using concepts related to artificial intelligence. Such interpretation can create a model which might eventually be used to predict new patterns, take actions and/or solve domain problems.

2.6.1 Machine Learning Algorithms

Based on how the learning phase of a problem is, most of the ML algorithms may be divided into one of the following categories:

Supervised A direct feedback is presented during the learning phase. [18] For the instances where the ME problem relies on a data set, the learning task uses the labeled training samples (i.e., samples that present the **target feature**) to synthesize the model that attempts to generalize the relationship between the feature vectors and the target variable. [20]

Unsupervised Infer hidden structures from the data set without direct feedback, such as known labels for the samples. [20]

Semi-supervised Most commonly, it is given by the extension of either supervised or unsupervised learning to include the other paradigm, resulting in a combination of both. [21]

Reinforcement Through iterative exploration, the learner is positively or negatively reinforced for its actions. The learner's goal is, ultimately, maximize the cumulative reward gained. [20]

2.6.2 Support Vector Machine: An Example of Supervised Learning

The Support Vector Machine algorithm (i.e., SVM) is a powerful tool, often used in classification problems.

Intuitively, the SVM algorithm attempts to find a hyperplane that separates the samples in a data set into two different groups: positives and negatives. Furthermore, the hyperplane is placed such that the distance between the **support vectors** (the closest samples) and it are maximized.

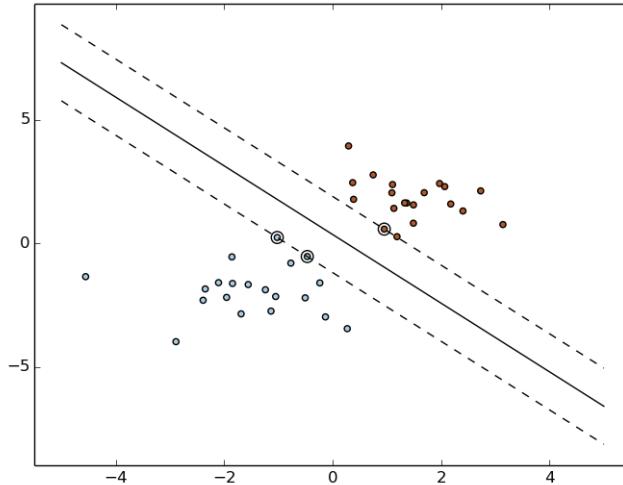


Figure 2.7: A SVM classifier projecting a hyperplane that perfectly separates two classes of samples. [1]

More elaborately, given any linearly separable data set X containing samples from two distinguished classes $\{-1, +1\}$, consider the vector w a hyperplane $d = \{x \mid w \cdot x + b = 0\}$ (represented in fig. 2.7 by the contiguous line) s.t.

$$\text{decision rule} \begin{cases} w \cdot u + b \geq 0 \implies y_u = +1 \\ w \cdot u + b < 0 \implies y_u = -1 \end{cases}$$

To prevent samples from falling into the margin or being misclassified, [22] reinforce that for any positive sample x_+ , $w \cdot x_+ + b \geq 1$. Similarly for x_- samples, $w \cdot x_- + b \leq -1$. These both constraints can be expressed as

$$y_i(w \cdot x_i + b) - 1 \geq 0 \quad (2.1)$$

Notice that $y_i(w \cdot x_i + b) - 1 = 0 \iff x_i$ is a **support vector**.

The width (the distance between the two margins) of the street is the vector $(x_+^0 - x_-^0)$ projected onto the vector w , where x_-^0 is a positive support vector and

x_+^0 is a negative one.

$$\begin{aligned}
width &= (x_+^0 - x_-^0) \cdot \frac{w}{\|w\|} \\
&= \frac{x_+^0 \cdot w - x_-^0 \cdot w}{\|w\|} \\
&= \frac{1 - b - (-1 - b)}{\|w\|} = \frac{2}{\|w\|}
\end{aligned} \tag{2.2}$$

As the goal is to maximize the width, while still respecting the constraint (2.1).

$$\max width = \max \frac{2}{\|w\|} \equiv \min \frac{1}{2} \|w\|^2 \tag{2.3}$$

Which can be solved using standard quadratic programming.

SVM for non-separable data sets (soft margins)

To deal with non-separable data sets, it is possible to introduce the variables ξ_i , [22] which represent a trade-off between maximum margin/distance of the misclassified samples from the decision boundary.

$$\begin{aligned}
&\min \frac{1}{2} \|w\|^2 + C \sum_1^m \xi_i, \text{ constrained to:} \\
&y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0
\end{aligned}$$

Such trade-off can be adjusted through the parameter C . Notice that small values for C might result in misclassification, whereas high values can produce overfitting.

Dependency over the dot product

Equation 2.3 does not explicitly illustrates how the model generated depents on the dot product between the training samples. The implications of such fact will be discussed in the next section. For now, let us convert 2.3 to its dual form. By the Lagrange multipliers method: [23]

$$L = \frac{1}{2} \|w\|^2 - \sum \alpha_i [y_i(w \cdot x_i + b) - 1] \tag{2.4}$$

$$\frac{\partial L}{\partial w} = w - \sum \alpha_i y_i x_i = 0 \implies w = \sum \alpha_i y_i x_i \tag{2.5}$$

$$\frac{\partial L}{\partial b} = - \sum \alpha_i y_i = 0 \implies \sum \alpha_i y_i = 0 \tag{2.6}$$

Applying (2.5) and (2.6) on (2.4), our problem of minimizing (2.3) constrained by (2.1) becomes maximizing L , subject to (2.6):

$$\begin{aligned} L &= \frac{1}{2} \sum \alpha_i y_i x_i \cdot \sum \alpha_j y_j x_j - \sum \alpha_i y_i x_i \cdot \sum \alpha_j y_j x_j - \sum \alpha_i y_i b + \sum \alpha_i \\ &= \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \end{aligned}$$

w and b can easily be found from (2.5) and $\alpha_i[y_i(w \cdot x_i + b) - 1] = 0$ (for any $i \mid \alpha_i \neq 0$), respectively.

Now, as we finally plug (2.5) back into our decision rule, it becomes clear that both training and prediction phases depend only on the dot product between the sample vectors: [23]

$$\text{decision rule} \begin{cases} \sum \alpha_i y_i x_i \cdot u + b \geq 0 \implies y_u = +1 \\ \sum \alpha_i y_i x_i \cdot u + b < 0 \implies y_u = -1 \end{cases}$$

Kernel functions

”The kernel function represent the dot product of both vectors projected onto the new space”. [23]

Some data sets are not linearly separable, as mentioned in a few sections above. They can, however, be projected to a different vector space, where they hopefully will be. To achieve this, a transformation ϕ is applied on both vectors. The dot product between those is then calculated in the new space and the value is used during training (the maximizing L) and prediction (the decision rule):

$$L = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j)$$

Remark 2.6.1 Practically, Let $k : (\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R} \mid k(u, v) = \phi(u) \cdot \phi(v)$,

$$L = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

Which entails, given a known function k , there is no need to explicitly know which are the projections of x_i and x_j onto the new vector space. This is commonly known as the **kernel trick**.

$$\text{decision rule} \begin{cases} \sum \alpha_i y_i k(x_i, u) + b \geq 0 \implies y_u = +1 \\ \sum \alpha_i y_i k(x_i, u) + b < 0 \implies y_u = -1 \end{cases}$$

The figure below illustrates a non-linearly separable data set defined in the \mathbb{R} being projected to the \mathbb{R}^2 .

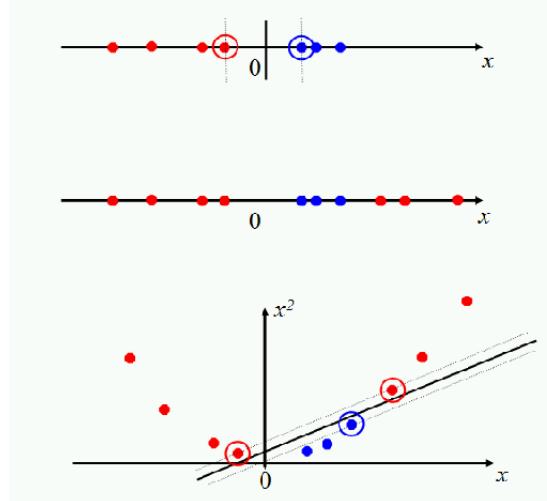


Figure 2.8: Projection of samples from the \mathbb{R} to the \mathbb{R}^2 , allowing SVM to find a hyperplane that perfectly separates both classes. [2]

Between many different kernels, two often used are the RBF: $\exp(-\frac{\|x-x'\|^2}{2\sigma^2})$ and the Polynomial: $(u^T v + c)^d$. [2]

2.6.3 Multi-class Classification

Not all problems are binary. In classification, this reflects data sets that have their samples separated into more than two classes. The Iris flower data set is an example of this, when predicting the samples' *species*.

There are many different approaches for multi-class classification. [24] For the scope of this work, consider only the following:

One-vs-All (OVA) n binary classifiers are built, where n is also the number of classes in the data set. For the classifier $c^j, j \in [1, n]$, all samples of the

j th class are taken as positive examples, at the same time that all the other samples are considered negative.

If

$$c^j(x) = \sum_{i=1}^m y_i \alpha_i^j x \cdot x_i + b$$

Then positive values for $c^j(x)$ indicate that the sample x belongs to the j th class. Additionally, greater $c^j(x)$ values imply on further distance from the hyperplane (i.e., $c^j(x)$ can also be interpreted as a **confidence value**) and the sample x should be assigned to the class which holds greatest confidence.

[25] Shortly, classification is given by

$$f(x) = \arg \max_j c^j(x)$$

All-vs-All (AVA) Also known as all-pairs or one-vs-one, a classifier c_{ij} is built for each pair of classes (i, j) , resulting in a total of $n(n - 1)$ classifiers. c_{ij} responds with positive values for samples of the i th class and negative values for samples of the j th class. Classification can be done by simply counting the class most frequently associated with x :

$$f(x) = \arg \max_i \sum_{j=1}^n c_{ij}(x)$$

Or equivalently, but only using $\frac{n(n-1)}{2}$ classifiers,

$$f(x) = \arg \max_i \sum_{j=1}^n \frac{j - i}{|j - i|} c_{\min(i,j) \max(i,j)}(x)$$

2.6.4 Evaluating learners

Machine Learning algorithms might be susceptible to data noise, incorrect configuration or even random factors, which would eventually decrease the generated model's accuracy. In order to evaluate this same accuracy, models are quite often tested after trained.

Considering that testing with the same data used for training will most likely produce unreliable results, a simple way to test a learner is to separate the labeled

data set into two chunks, where the first is used for training. The second chunk is then given to the learner, which attempts to predict the samples. Finally, the predictions made by the learner would be compared with the actual labels.

Confusion Matrix

When testing classification models, one way to visualize the wrong predictions made by the learner is a confusion matrix, where the item a_{ij} is the number of times that a sample of the class i was classified as being of the class j .

	a	b	c	d
a	12	3	2	0
b	7	12	2	4
c	0	4	54	8
d	6	0	1	23

Table 2.4: Example of confusion matrix for a data-set with four different classes.

Naturally, a diagonal matrix represents that all samples of the class i were classified as i , which is the best possible outcome (no errors).

Cross Validation

Sometimes (for example, when the data set does not have too many samples), partitioning of the data set into subsets might be benign to the learning process, as the model will be constructed only considering a small, random portion of the samples. This event is known as **underfitting**.

When k -fold cross-validating, [26] the data set can be partitioned into k folds. For each fold k_i , a model is trained with all folds, except for k_i . The model is then tested over k_i . Finally, the score reported by the cross-validation method is the average accuracy when testing over all folds.

Grid Search

Machine Learning algorithms may require specific parameters to run. For instance, SVM requires C and which *kernel* it should use. Additionally, kernels might re-

quire their own parameters. As these parameters strongly affect how the generated model will be, one cannot choose them arbitrarily.

Grid Search is a traditional method used to find the parameters that optimize the generalization of learning model over a specific data set. [27] Given a set of possible parameters, it will exhaustively search for the combination of those that generate the best outcome, which might be evaluated by a validation method such as cross-validation.

2.6.5 Examples of Learning

Coffee Selling Rate

The figure bellow illustrates the distribution of coffee sales per time of the day in a particular coffee-shop. [28] Clearly being a regression problem, a machine learning algorithm must create a model that appropriately generalizes the distribution observed. Such model will eventually be used to predict the selling rate in the following days.

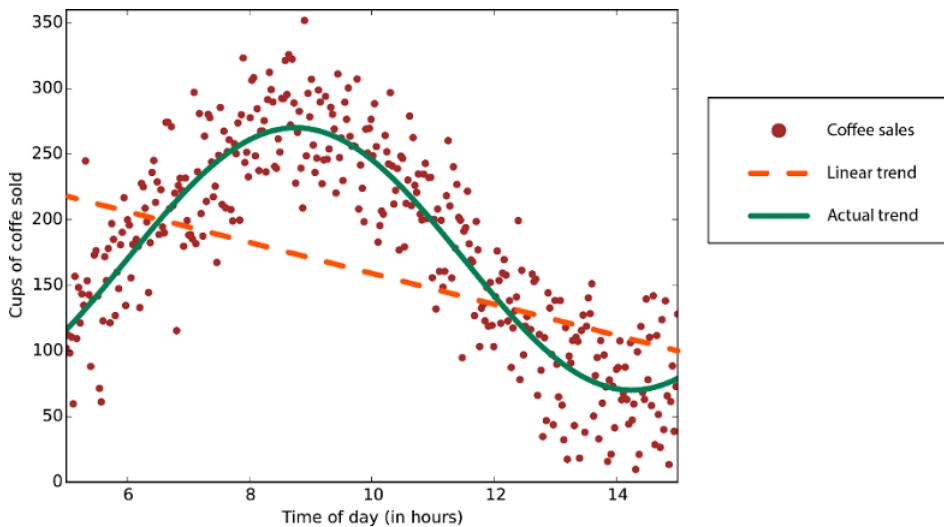


Figure 2.9: graphic representation of a data set generalization by a linear (orange) and a nonlinear model (green).

The orange line and the green arc represent two different models. The orange line, which represents the linear model, clearly does not generalize the data set

appropriately, once it induces an error much larger than necessary. [28] The non-linear model, i.e., the green arc, was capable of generalize the data inducing a smaller error.

Iris Flower

Consider the Iris flower data set as defined in 2.1. In order to predict the feature *Species* of a given sample, one could train a Support Vector Machine classifier.

Through GridSearch, it was found that the SVM algorithm with $C = 100$, $\gamma = .01$ and rbf kernel is capable of finding a model yielding .99% accuracy. The samples misclassified during the test phase were represented by the confusion matrix bellow.

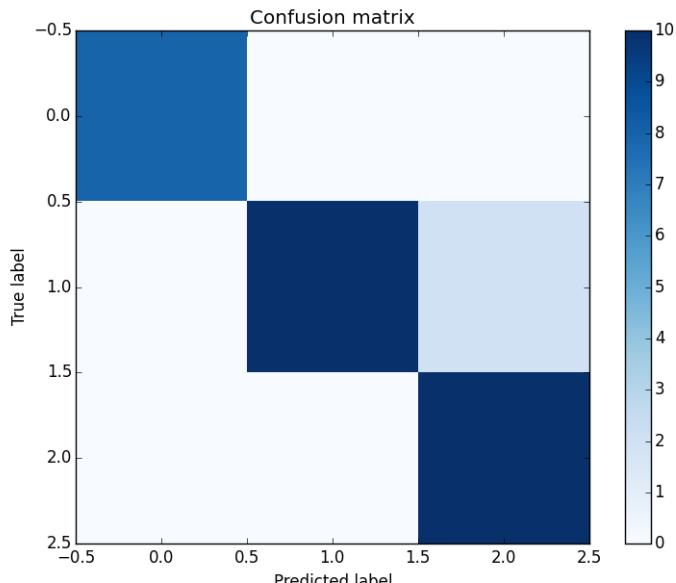


Figure 2.10: Confusion matrix of a SVM with $C = 100$, $\gamma = .01$ and rbf kernel when predicting samples from the Iris flower data set.

Chapter 3

Linear Dimensionality Reduction

As presented in the previous sections, data sets with many features may present a series of issues: difficult visualization, high performance requirements, noise etc. In this section, it will be discussed methods related with linear dimensionality reduction, i.e., the shrinking of data sets by transformation and/or removal of features, while minimizing information loss.

Consider the data set K . K has its samples expressed by two similarly scaled dimensions. It is clear, however, that the samples follow a very particular distribution:

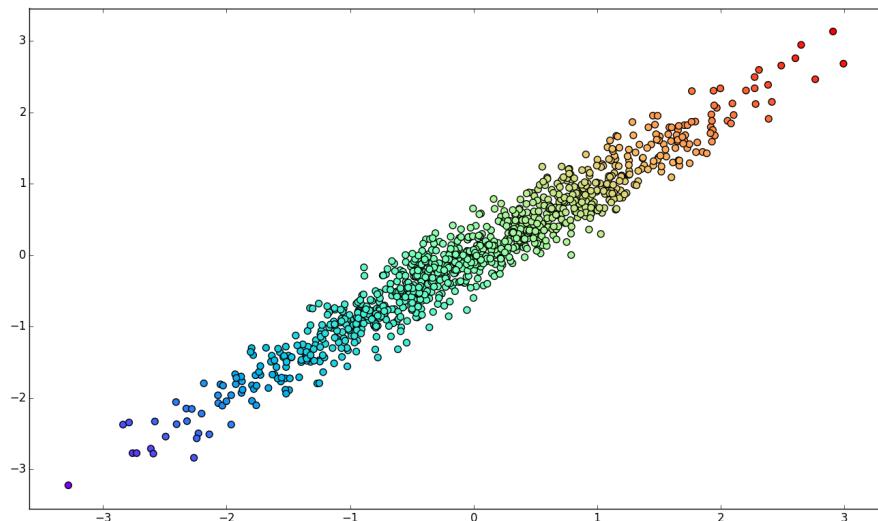


Figure 3.1: The data set $K \in \mathbb{R}^2$.

Additionally, something interesting can be observed when analyzing the covariance matrix of K : as it is not a diagonal matrix, the variance of x from its mean somehow correlates with the variance of y . [29]

	x	y
x	1.26682132	1.29158697
y	1.29158697	1.40358478

Table 3.1: Covariance between the components of K .

3.1 Principal Component Analysis

As in K , some data sets follow certain distributions that are majorly contained in a few orthogonal components, where a component is the result of a linear combination of the original features.

Principal Component Analysis (PCA) is a statistical technique that attempts to transform a n -dimensional data set X into a m -dimensional data set Y , where, hopefully, $k \ll n$. Furthermore, the dimensions of Y will necessarily be orthogonal components aligned with the direction in which the variance of samples in X is maximum, commonly referred to as **principal components**. [30] In figure 3.2, the orange and purple vectors are the principal components of the data set K .

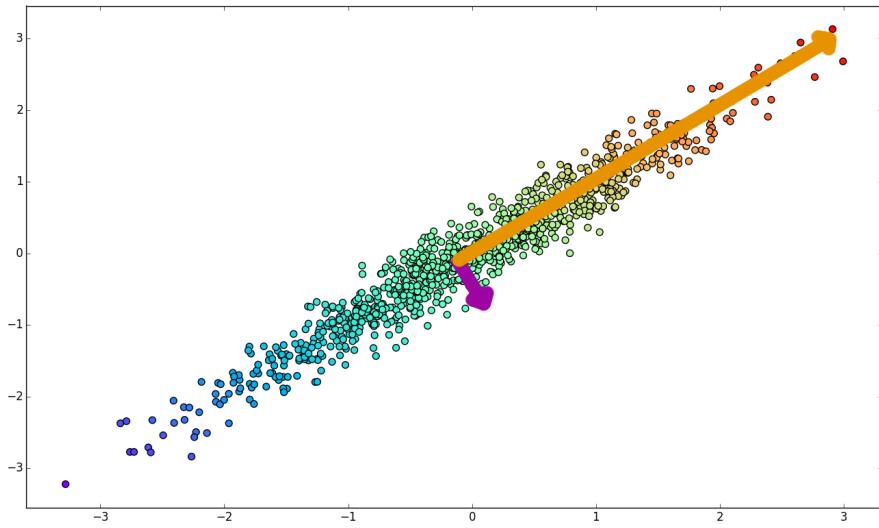


Figure 3.2: The principal components of K .

3.1.1 Study of the PCA Algorithm

Let D be a dataset with n samples and f features and $X = HD$, where H is the centering matrix. Our goal is to find which are the principal components of the covariance matrix Σ_X .

$$\Sigma_X = \frac{1}{n} X^T X \quad (3.1)$$

Using the **Singular Value Decomposition** method described in section 2.3.3, we known that

$$X = U\Sigma V^T \quad (3.2)$$

Needless to say, Σ is the diagonal matrix of singular values, not to be mistaken by the covariance matrix Σ_X .

From 3.1 and 3.2:

$$\begin{aligned} \Sigma_X &= \frac{1}{n} X^T X \\ &= \frac{1}{n} (U\Sigma V^T)^T U\Sigma V^T \\ &= \frac{1}{n} V\Sigma^2 V^T \end{aligned}$$

Which entails that V is the orthonormal matrix with Σ_X 's eigenvectors as columns, whereas Σ contains the correspondent eigenvalues σ_{ii}^2 associated with $v_i \in V$. $v_i \in V$ is, in fact, a principal component of X and its associated eigenvalue σ_i module gives X spectral radius. As we are interested in the dimensions that give most variance, keep only the $m \in \mathbb{R}$ most significant eigenvalues and their correspondent eigenvectors.

Finally, it also worth remarking once again that the principal components are linear combinations of the original features (the canonical base) and V is the change-of-basis matrix from the generated base to the canonical. Naturally, V^{-1} is a change-of-basis matrix from the original space to the one that is generated by the principal components. Formally, if x is a sample (row vector) from the X data set, its project y is defined as:

$$\begin{aligned} Vy^T &= x^T \\ y^T &= V^{-1}x^T \end{aligned}$$

In the other hand, V is orthogonal, hence V^{-1} exists and it is equal to V^T :

$$\begin{aligned} y^T &= V^T x^T \\ y &= (V^T x^T)^T \\ &= xV \end{aligned}$$

3.1.2 Formalization of the PCA Algorithm

Let D be a data set with n samples and f features and $m \in \mathbb{R}$ the number of dimensions desired for the reduced data set. [31] [32]

1. Find $X = HD$, where H is the centering matrix.
2. Calculate the covariance matrix Σ_X .
3. Use singular value decomposition to find the eigenvalues $\lambda = \{\lambda_i\}$ and eigenvectors $V = \{v_i\}$ of Σ_X .
4. Sort the eigenvalues by their absolute value in descending order and select the first m ones and their respective eigenvectors.

3.2 Multidimensional Scaling

Alternatively to PCA, Multidimensional Scaling (or simply MDS) can be used to reduce the dimensionality of a data set. The method has, however, an extensive application domain and often appears in the literature in different contexts. An example of this is the problem of, given a set of objects O and a dissimilarity measurement $\delta_{rs}, \forall(r, s) \in O \times O$, finding a suitable representation in the \mathbb{R}^n for the objects in O . [10]

For this project, we study the **classic MDS**. That is, when the dissimilarities considered are the euclidean distances between coordinates in the \mathbb{R}^n .

3.2.1 Study of the MDS

If $\delta = [\delta_{rs}]_{n \times n}$ is the dissimilarity matrix, where δ_{rs} represents the euclidean distances between two samples $x_r, x_s \in \mathbb{R}^m$ from the data set $[X]_{n \times m}$ induced by the L_2 -norm. In other words,

$$\begin{aligned} \delta_{rs} &= \sqrt{\sum_i (x_{ri} - x_{si})^2} \\ &\iff \\ \delta_{rs}^2 &= \sum_i (x_{ri} - x_{si}^2) \\ &= (x_r - x_s) \cdot (x_r - x_s) \\ &= x_r \cdot x_r + x_s \cdot x_s - 2x_r \cdot x_s \end{aligned} \tag{3.3}$$

Now consider the matrix $B = XX^T$, where $b_{rs} = x_r \cdot x_s$. B can be decomposed as $U\Sigma U^T = U\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}U^T = U\Sigma^{\frac{1}{2}}(U\Sigma^{\frac{1}{2}})^T = XX^T \iff X = U\Sigma^{\frac{1}{2}}$. If B can be derived from 3.3, the problem is reduced to simply decompose B and using its eigenvalues and eigenvectors (similarly to what was done in PCA) to construct the data set Y . [10]

Firstly, we will assume that Y is centered in the origin (i.e., Y has its features' means equal to zero):

$$\sigma_f = \sum_i y_{if} = 0, \forall f \in [0, m) \tag{3.4}$$

Now, 3.3 \implies

$$\begin{aligned}
\frac{1}{n} \sum_r \delta_{rs}^2 &= \frac{1}{n} \sum_r (x_r \cdot x_r + x_s \cdot x_s - 2x_r \cdot x_s) \\
&= \frac{1}{n} \sum_r x_r \cdot x_r + \sum_r x_s \cdot x_s - 2 \sum_r x_r \cdot x_s \\
&= \frac{1}{n} \sum_r x_r \cdot x_r + nx_s \cdot x_s - 2 \sum_r 0 \cdot x_s \\
&= \frac{1}{n} \sum_r x_r \cdot x_r + x_s \cdot x_s \\
&\iff \\
x_s \cdot x_s &= \frac{1}{n} \left(\sum_r \delta_{rs}^2 - \sum_r x_r \cdot x_r \right)
\end{aligned} \tag{3.5}$$

Similarly, to 3.5:

$$x_r \cdot x_r = \frac{1}{n} \left(\sum_s \delta_{rs}^2 - \sum_s x_s \cdot x_s \right) \tag{3.6}$$

Putting 3.5 and 3.6 back in 3.3:

$$\begin{aligned}
\delta_{rs}^2 &= \frac{1}{n} \left(\sum_s \delta_{rs}^2 - \sum_s x_s \cdot x_s + \sum_r \delta_{rs}^2 - \sum_r x_r \cdot x_r \right) - 2x_r \cdot x_s \\
&\implies \\
x_r \cdot x_s &= -\frac{1}{2} \left(\delta_{rs}^2 - \frac{1}{n} \left[\sum_s \delta_{rs}^2 - \sum_s x_s \cdot x_s + \sum_r \delta_{rs}^2 - \sum_r x_r \cdot x_r \right] \right) \\
&= -\frac{1}{2} \left(\delta_{rs}^2 - \frac{1}{n} \left[\sum_s \delta_{rs}^2 + \sum_r \delta_{rs}^2 - 2 \sum_r x_r \cdot x_r \right] \right)
\end{aligned} \tag{3.7}$$

To eliminate the $x_r \cdot x_r$ term from 3.7:

$$\begin{aligned}
\frac{1}{n^2} \sum_s \sum_r \delta_{rs}^2 &= \frac{1}{n^2} \sum_s \sum_r (x_r \cdot x_r + x_s \cdot x_s - 2x_r \cdot x_s) \\
&= \frac{1}{n^2} \sum_s \left(\sum_r x_r \cdot x_r + \sum_r x_s \cdot x_s - 2 \sum_r x_r \cdot x_s \right) \\
&= \frac{1}{n^2} \sum_s \left(\sum_r x_r \cdot x_r + n x_s \cdot x_s \right) \\
&= \frac{1}{n^2} \left(n \sum_r x_r \cdot x_r + n \sum_s x_s \cdot x_s \right) \\
&= \frac{1}{n^2} 2n \sum_r x_r \cdot x_r \\
&= \frac{2}{n} \sum_r x_r \cdot x_r
\end{aligned} \tag{3.8}$$

Finally, applying 3.8 on 3.5:

$$B_{rs} = x_r \cdot x_s = -\frac{1}{2}(\delta_{rs}^2 - \frac{1}{n}[\sum_s \delta_{rs}^2 + \sum_r \delta_{rs}^2 - \frac{1}{n} \sum_s \sum_r \delta_{rs}^2]) \tag{3.9}$$

From 3.9, it becomes clear that B is, in fact, the double centering of the matrix $A = -\frac{1}{2}\delta^2$. I.e., $B = HAH$. Spectral decomposition can now be performed onto B , resulting in the matrices Σ and U .

Finally, in order to reduce the dimensionality of the embedding, we can sort the eigenvalues (and their respective eigenvectors, the columns of U) in decrease order and keep only the ones that offer greater variance.

Remark 3.2.1 As euclidean distances were used to build the dissimilarity matrix δ , B is indubitably positive semidefinite, hence $\Sigma_i \geq 0, \forall i \in [0, n]$. However, negative eigenvalues might appear if other dissimilarity measurement were to be used. In these cases, one might consider to simply ignore such components.

3.2.2 Formalization of the Multidimensional Scaling Method

Let X be a data set with n samples and f features and $m \in \mathbb{R}$ the number of dimensions desired for the reduced data set. [10]

1. Calculate the dissimilarity matrix $[\delta]_{rs}$, where $\delta_{rs} = \sqrt{\sum_i (x_{ri} - x_{si})^2}$

2. Calculate the matrix $A = -\frac{1}{2}\delta_{rs}^2$ and $B = HAH$, where H is the centering matrix.
3. Use spectral decomposition to find the matrices Σ and U .
4. Select the m greatest eigenvalues in Σ . From these, create the matrices $\Sigma' = [\sigma'_{m \times m}]$ and $U' = [u'_{n \times m}]$, where each column i contains the eigenvector associated with σ'_i .
5. Construct the m -dimensional embedding $Y = U'\Sigma'^{\frac{1}{2}}$

3.3 Evaluating Reductions

“Often it is the researcher’s past experience with MDS and his or her judgment that ultimately determine whether the fit of a particular MDS solution is acceptable or not.” [33]

Although the factors that determine what is a “good” reduction are strongly influenced by particularities of the problem in hand, some measures were developed to attempt to somehow formalize it. One in particular, which recurrently appears in literature, is known as the **Kruskal’s stress**.

Intuitively, Kruskal’s stress [33] considers reductions that preserve dissimilarities between samples a better fit than the ones which highly distort them. Formally, let $X_{n \times f}$ be a data set with n samples and f dimensions, $Y_{n \times p}$ its reduction to p dimensions, and the dissimilarity measurements δ_{ij} and $\hat{\delta}_{ij}$ defined for all samples i and j in X and Y , respectively:

$$Stress = \left[\frac{\sum_i \sum_j (\delta_{ij} - \hat{\delta}_{ij})^2}{\sum_i \sum_j \delta_{ij}^2} \right]^{\frac{1}{2}}$$

From the formula above, *Stress* is visibly contained in the interval $[0, 1]$, where 0 represents the best possible fit (all dissimilarities are the same), while 1 represents the worse.

3.4 Classification and Regression Over Linearly Reduced Data Sets

This section reports the performance of classifiers and regressors over determined data sets and their reduced forms. All experiments followed the format bellow:

1. The data set $X_{n \times m}$ was loaded.
2. Grid Search was executed over the original data set.
3. for $d \in D \subset \mathbb{N} \mid d \in D \implies d < m$:
 - (a) The d -dimensional embedding Y of X was calculated.
 - (b) Grid Search was executed over Y .

3.4.1 K Data Set

The figure bellow illustrates the results of PCA algorithm application over the artificial K data set. Notice that, for the second application, it correctly chose to discard the vertical dimension, as the samples offer less variability in this component.

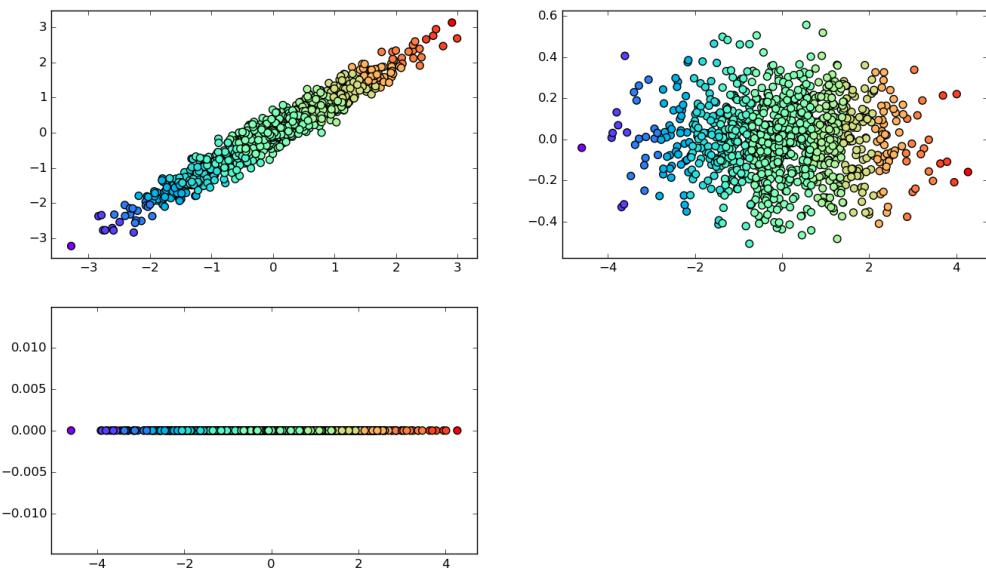


Figure 3.3: The PCA algorithm reducing K to 2 and 1 dimension, respectively.

	Original \mathbb{R}^k	\mathbb{R}^2	\mathbb{R}
Pred. accuracy	.98	.98	.99
GridSearch time	1.15 s	.84 s	1.26 s
Reduction time	-	0.995 ms	1.118 ms
Stress	-	0	.0399
Data size	15.62 KB	15.62 KB	7.81 KB

Table 3.2: Description of predictions and reduction performance for K .

From the table above, one can also observe how Kruskal's stress might not be an appropriate quality measurement, given a specific domain. Indeed, for data set K , stress increase resulted from the removal of one of the components did not imply on prediction accuracy decrease.

3.4.2 The Iris Flower Data Set

The Iris flower 2.1, containing 150 samples and 5 features.

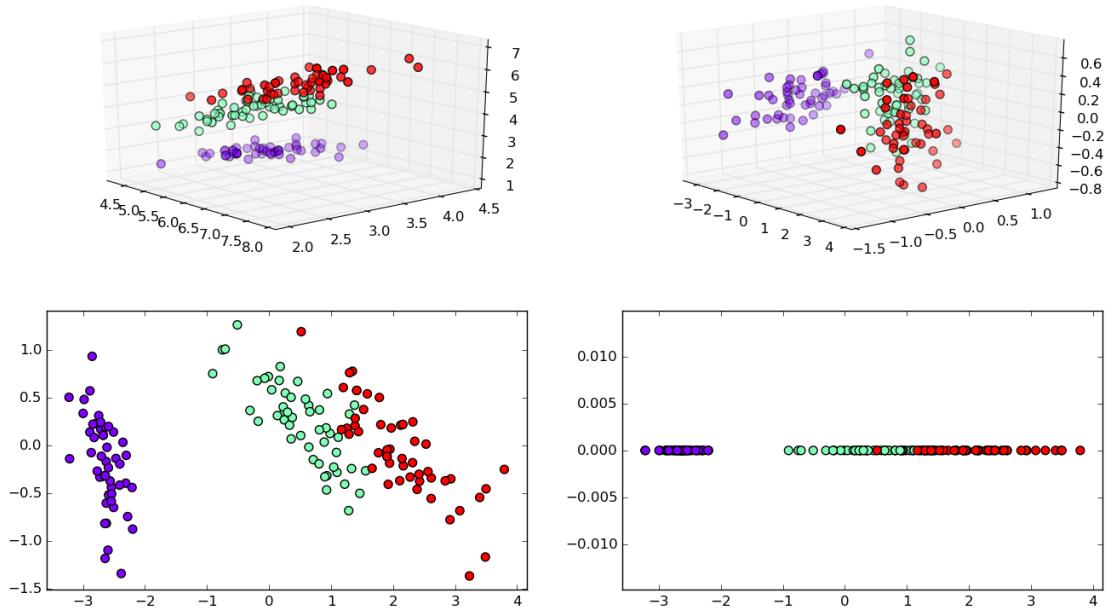


Figure 3.4: The Iris Flower data set and its reductions to 3, 2, and 1 dimensions, respectively.

Although the data set became nonlinearly separable, classes are still somewhat organized in different clusters.

	Original data	Reduced data (\mathbb{R}^2)	Reduced data (\mathbb{R})
Pred. accuracy	.99	.97	.94
GridSearch time	1.71 s	1.64 s	1.78 s
Reduction time	-	0.995 ms	1.118 ms
Stress	-	.0418	.1095
Data size	4.68 KB	2.34 KB	1.17 KB

Table 3.3: Description of predictions and reduction performance for Iris flower.

3.4.3 The Digits Data Set

Digits data set is composed by 1797 samples, 64 features and 10 classes. Each sample is a 8x8 image of a hand-written digit from 0 to 9.

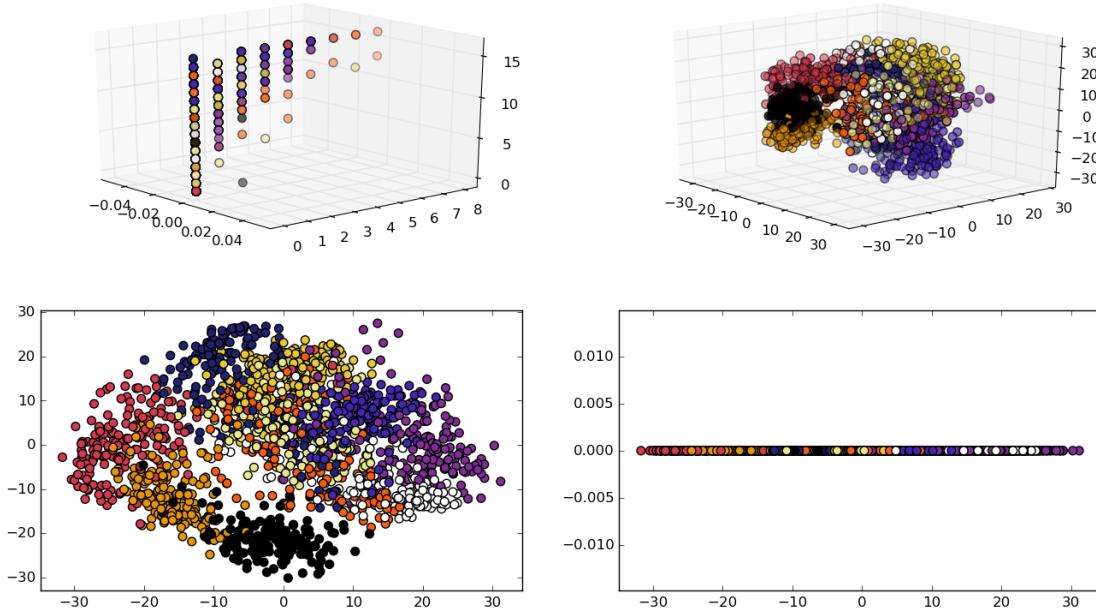


Figure 3.5: Digits data set reduced to 10, 3, 2 and 1 dimension, respectively.

	\mathbb{R}^{64}	\mathbb{R}^{10}	\mathbb{R}^3	\mathbb{R}^2	\mathbb{R}
Pred. accuracy	.98	.95	.74	.64	.39
GridSearch time	8.51 s	21.33 s	151.04 s	132.18 s	113.78 s
Reduction time	-	0.02 s	0.01 s	0.01 s	0.01 s
Stress	-	.1594	.4218	.5405	.7092
Data size	898.5 KB	140.39 KB	42.11 KB	28.07 KB	14.03 KB

Table 3.4: Description of predictions and reduction performance for Digits.

Notice that it was possible to eliminate 54 dimensions, consistently reducing the data set size, and only suffering 3% of prediction accuracy loss. The score drastically decreased, however, when more dimensions were removed. Furthermore, accuracy decrease was followed by stress increase.

Chapter 4

Nonlinear Dimensionality Reduction

Although PCA has presented promising results in the previous section, hence its great popularity in dimensionality reduction problems, there are many examples in which PCA will fail in its task. Consider the example bellow:

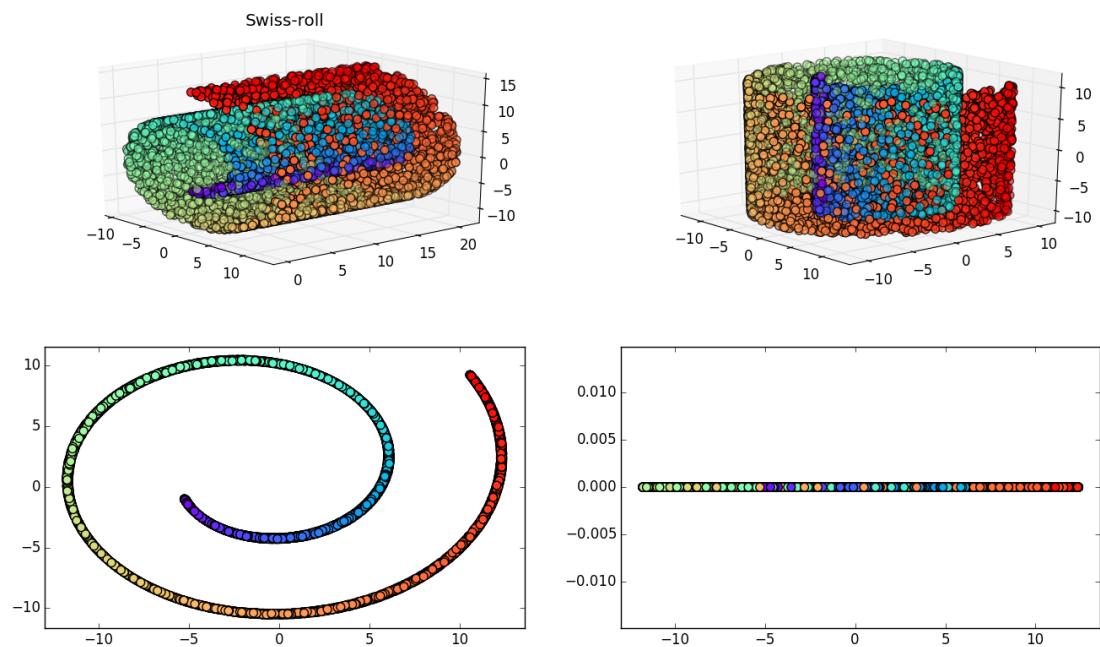


Figure 4.1: The **Swiss-roll manifold** and its reductions to 3, 2 and 1 dimensions, respectively, using the PCA algorithm.

Looking at figure 4.1 (specially in the last reduction, to a single dimension), it becomes clear that PCA has a big draw back: it assumes that the data lies on a liner subspace [7] and, therefore, applying linear transformations to it will rotate and scale it, without distort the original data structure. When this assumption does not hold, PCA will incorrectly extract the underlying structure, possibly mixing very dissimilar samples.

The table bellow describes a regression attempt over the data sets illustrated in figure 4.1, were the feature being predicted is the contiguous value represented by the vertexes' colors:

	Original data	Reduc. (\mathbb{R}^3)	Reduc. (\mathbb{R}^2)	Reduc. (\mathbb{R}^1)
Accuracy	1.	1.	.68	.54
GS time	10.98 s	10.48 s	5.20 s	2.59 s
Reduc. time	-	0.038 s	0.038 s	0.035 s
Stress	-	0	.2788	.5271
Data size	23.44 KB	23.44 KB	15.62 KB	7.81 KB

Table 4.1: Regression accuracy and reduction performance for the Swiss-roll data set.

Clearly, linear dimensionality reduction techniques are not adequate to reduce the Swiss-roll. In fact, it is not adequate to reduce any data set lying on a nonlinear manifold.

4.1 The ISOMAP Algorithm

Firstly suggested by Tenenbaum, de Silva and Langfor, **Isometric Feature Mapping** (or ISOMAP) assumes that the data lies near a smooth manifold. If the assumption is reasonable, it is be possible to explore concepts such as neighborhood and local linearity to map the manifold to a linear structure before reducing it with a linear algorithm.

In this section we will discuss the ISOMAP algorithm and its inner workings. We will then proceed to formalize it. Finally, empirical tests results achieved during the project will be shown.

4.1.1 Study of the ISOMAP Algorithm

As the original data set might be folded, twisted or curved, [34] we must first find a suitable linear representation for it.

Let S be our original data set, as illustrated in figure 4.2.

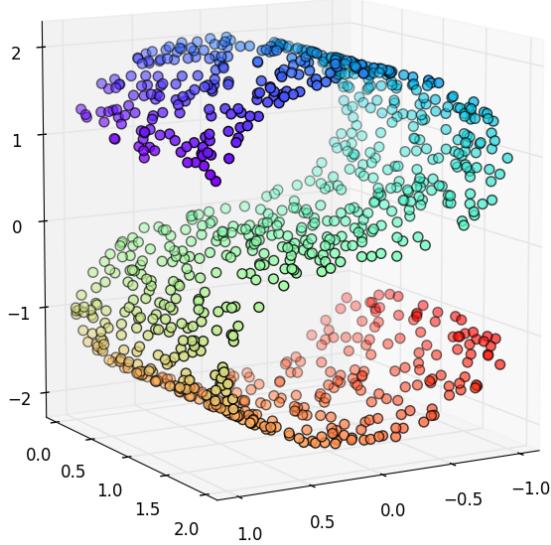


Figure 4.2: The data set S , consisting of 1000 samples and 3 features.

Additionally, consider the symmetric undirected weighted graph $G = (V, E)$ and $w: E \rightarrow \mathbb{R} \mid w(x, y) = \delta_{xy}$, where δ_{xy} is the euclidean distance between the samples x and y in S . That is,

$$\delta_{xy} = \sqrt{\sum_i (x_i - y_i)^2}, \forall (x, y) \in S \times S \mid x \neq y$$

Now that only distances were kept, an infinite number of n-dimensional embeddings can be found with **MDS**, as every solution can be transposed, rotated or reflected. This does not fix the non-linearity of the data, though, as the original distances strictly constraint the samples to their original pattern. In order to achieve this, **Nearest neighbor search** can be performed over G , resulting in the graph H . Nearest-neighbor search will preserve edges connecting closer samples, hence preserving local (linear) distances, but erase edges connecting samples which are far from each other (non necessarily linear). Obviously, the search parameters

(K or ϵ) must be carefully chosen to limit the connectivity of the vertices to a small (linear) neighborhood while maintaining the graph completely connected. Ideally, H will be a **mesh graph**.

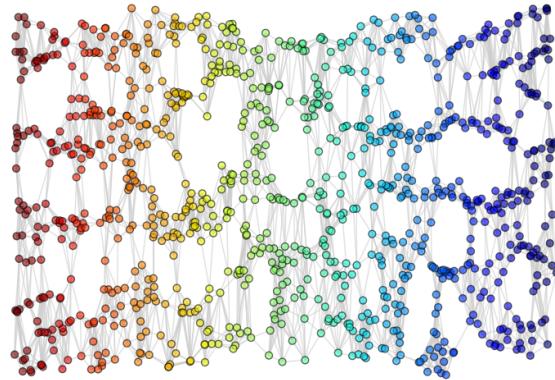
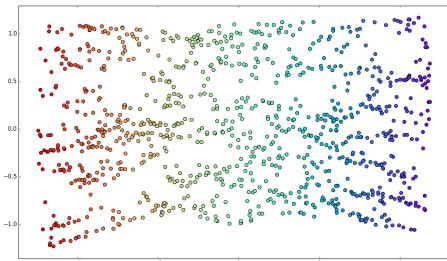


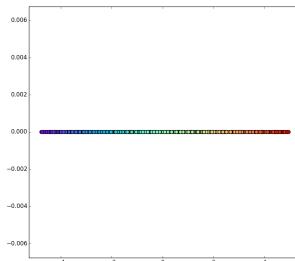
Figure 4.3: The graph H .

At this moment, not all distances in H are defined. This can be easily handled, though, by performing the **Floyd-Warshall** algorithm over H . Alternatively, M can be achieved by performing **Dijkstra's algorithm** for all nodes and joining all shortest-path trees found.

Finally, M is a euclidean graph which roughly lies on a linear subspace. Furthermore, the adjacency matrix associated to M contains not the distance induced by the L_2 norm, but the geodesic pairwise distances. [8] The **MDS** method can now be used to construct a representation in sub-spaces of the \mathbb{R}^n . S , specifically, can be reduced to the \mathbb{R}^2 or \mathbb{R} :



(a) S reduced to two dimensions.



(b) S reduced to one dimension.

4.1.2 Formalization of the ISOMAP Algorithm

Let X be the original data set and $p \in \mathbb{R}$ the number of dimensions desired for the reduced data set, [34]

1. Construct the weighted graph G from the distances pairwise $\delta_{xy}, \forall(x, y) \in S \times S, x \neq y$ and find the graph H by applying the **nearest-neighbor algorithm** on the graph G .
2. Compute the shortest path graph M between all pairs of nodes from graph H . This might be done by the **all-pairs Dijkstra's** or by the **Floyd-Warshall algorithm**.
3. Use M to construct the p -dimensional embedding using the **MDS** algorithm.

4.1.3 Computational Complexity

If n is the number of training samples, $f \in \mathbb{R}$ the number of features and $k \in \mathbb{R}$ the number of nearest neighbors:

1. The time complexity associated with building the neighborhood graph is $O(n^2)$, whereas the space necessary to represent the distances kept is n^2 .
2. Algorithms for finding shortest path graph can be executed in-place and do not increase space complexity. Regarding time complexity, however:
 - (a) Dijkstra's algorithm implementation using Fibonacci Heaps have time complexity $O(nk + n \log n)$. As the algorithm must be calculated for each node, this step has time complexity $O[n^2(k + \log n)]$.
 - (b) Alternatively, using the Floyd-Warshall algorithm, this step has time complexity equals to $O(n^3)$.
3. Taking $O(n^3)$ time steps to calculate the spectral decomposition, MDS clearly is the bottleneck of the entire algorithm. [7] [35] As for space complexity, $O(f^2 + nf)$ is required for storing the matrices Σ and U .

The time complexity of the ISOMAP algorithm (when using Dijkstra's) is, therefore, $O[n^2 + n^2(k + \log n) + n^3]$, and the space complexity is $O(n^2 + f^2 + nf)$.

Experiment 4.1.1 (Timing the ISOMAP Algorithm) Consider the data set **Digits** with 1797 samples and 64 features. The following table indicates the time duration for each step of the ISOMAP algorithm when reducing Digits to 3 dimensions:

Pairwise distances from data set	.44 s
K Nearest Neighbors Search	1.3 s
All Pairs Dijkstra's	51.44 s
MDS	118.84 s
Total Time	172.69 s

Table 4.2: Listing of time spent on each step of the ISOMAP algorithm.

As expected, most of the time (68.81%) was spent executing MDS.

In practice, ISOMAP's time complexity of $O[n^2 + n^2(k + \log n) + n^3]$ makes it unsuitable for data sets with high number of samples. This is illustrated in Shi and Gu's experiments: ISOMAP could not reduce data sets with more than 6000 samples in reasonable time. [36]

Experiment 4.1.2 Let **Spam** be a data set with 4601 samples and 57 features. The table bellow describes the time necessary for the implemented ISOMAP and sk-ISOMAP (the algorithm from the scikit-learn library) to reduce Spam to 3 dimensions:

Algorithm	Time
ISOMAP	29.9 m
sk-ISOMAP	10.76 s

Table 4.3: Timing the implemented ISOMAP and scikit-learn's implementation.

The implemented ISOMAP required almost 30 minutes to reduce Spam. Surprisingly, scikit-learn's implementation executed incredibly fast. A set of factors contribute for this result:

- **Ball Tree** is used for efficient neighbor search (requiring only $O(fn \log k \log n)$ time steps).

- Many components in scikit-learn have python signatures, but are actually implemented in C, considerably increasing performance.
- ISOMAP is implemented as a kernel for the **KernelPCA** method. Further investigation is done in section 4.1.4.

4.1.4 Extensions

ISOMAP as a variation of Kernel PCA

While “classic” PCA aims to find the principal components of the covariance matrix $\Sigma = \frac{1}{n}X^T X = \frac{1}{n}\sum_k x_{ik} \cdot x_{kj}, \forall(i, j) \in [0, |S|]^2$ of a centered data set X , it can be easily modified to attempt to do the same, but over the matrix $\Sigma' = \frac{1}{n}\sum_i \phi(x)_i \cdot \phi(y)_i$, where $\phi: \mathbb{R}^f \rightarrow \mathbb{R}^p$ is a kernel function that projects points in the original f -space to a different p -space. Such modification is named Kernel PCA.

In his paper, Ghodsi describes how MDS (and therefore, ISOMAP) is equivalent to Kernel PCA, given it is using the following kernel function: [8]

$$K_{ISOMAP} = -\frac{1}{2}H\delta^2H$$

Where $H\delta H$ is the double centered dissimilarity matrix, i.e, the double centered geodesic distances between samples found from the shortest-path graph.

An advantage of substituting the MDS algorithm by Kernel PCA is that the latter provides an embedding, i.e., a transformation matrix from the original space to the reduced one; whereas MDS only embeds the data set. [37] However, Silva and Tenenbaum have shown that MDS’s result can be used to reduce new samples (Landmark-MDS). [35] More will be discussed in 4.1.4.

The figure bellow displays the reduction of the Swiss-roll data set done by the original implementation of ISOMAP (using MDS) and scikit-learn’s implementation (using Kernel PCA). Clearly, they are identical.

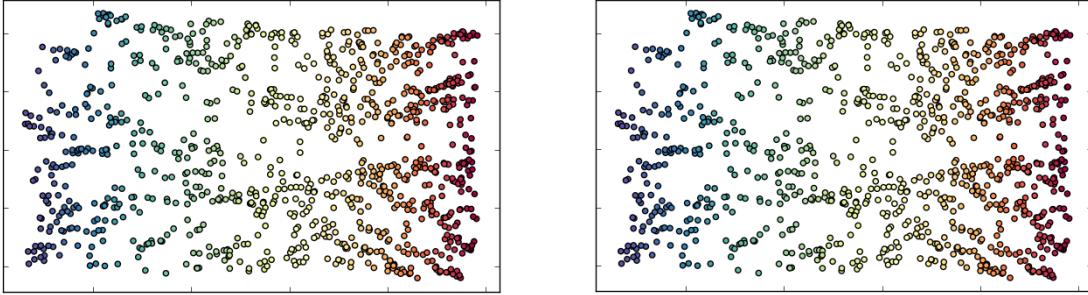


Figure 4.5: The embedded swiss-roll data set by the original and scikit-learn’s implementation.

Incremental ISOMAP

The original ISOMAP algorithm reduces the data set in a single execution. Even its implementation with Kernel PCA will not update the reduction model if new data points are given, only reduce these points using the eigenvectors previously found. Therefore, updating ISOMAP reduction model requires recalculating it at every new data “batch”, which is very costly.

In order to allow ISOMAP to continuously learn from a stream of data, Incremental ISOMAP was proposed. [38] Relying on the fact that new data points will unlikely affect a large subsets of vertices, it selectively updates the structures kept by ISOMAP. The update is performed as it follows.

Let δ_{ab} be the dissimilarity measure between two samples a and b , x_i be a sample in the nearest neighbor graph, x_j a neighbor of x_i such that their dissimilarity is higher among all x_i ’s neighbors. Finally, consider x_{n+1} a new sample to increment in the reduction model.

- The new sample x_{n+1} will replace x_j as a neighbor of x_i if $\delta_{x_i x_{n+1}} < \delta_{x_i x_j}$. A list of added or removed edges should be stored.
-

L-ISOMAP

ISOMAP’s time complexity depends mainly on the number of samples. Remember: for N samples, ISOMAP will require $O[n^2(\log n + k)]$ and $O[n^3]$ time

operations to compute the shortest-paths and eigenvectors/values, respectively.

Landmark ISOMAP aims to reduce the complexity of the algorithm by finding the dissimilarities between n landmarks (a small subset of samples) and all the original samples. Its complexity can then be reduced to $O[n^2(\log N + k) + n^2N]$. If $n \ll N$, this represents a expressive performance improvement. [35]

Let X be the initial data set of interest, l be the dimensionality to which X should be reduced and $\epsilon > 0 \in \mathbb{N}$ a number to “ensure stability”. [35] L-ISOMAP is defined as it follows.

1. Select n random samples from X , such that $n > l + 1 + \epsilon$.
2. Calculate the nearest-neighbors of each selected samples (all the original samples should be considered as neighbors).
3. Calculate the shortest-paths from the neighborhood graph obtained in the previous step and the dissimilarity matrix $[\delta]_{n \times N}$.
4. Perform Landmark MDS, an slightly modifed MDS:
 - (a) Apply classical MDS to the submatrix of landmarks $[\delta]_{n \times n}$.
 - (b) Embed the remaining points through:

$$y = -\frac{1}{n}L^{-1}(\Delta_x - \bar{\Delta}_n)$$

$$y = -\frac{1}{n} \begin{pmatrix} \frac{v_1^T}{\sqrt{\lambda_1}} \\ \frac{v_2^T}{\sqrt{\lambda_2}} \\ \vdots \\ \frac{v_n^T}{\sqrt{\lambda_n}} \end{pmatrix} (\Delta_x - \bar{\Delta}_n)$$

Where Δ_x is the square dissimilarity column vector from sample x to all landmarks and $\bar{\Delta}_n$ the vector containing the column mean of Δ_n (the square dissimilarity matrix between landmarks).

p-ISOMAP

The embedding found by ISOMAP is strongly depentend of the parameters k or ϵ . When these parameters are too small, the neighborhood graph extracted

might be disconnected, causing the dissimilarity between some vertices to be infinite. On the other hand, when the parameters are too big, vertices from different neighborhoods are prone to stay connected, preventing the manifold to be “unfolded”. Hence the importance of choosing appropriate parameters. However, in many cases, these “appropriate” parameters might not be evident. Hence the interest of developing a parameterless ISOMAP.

p-ISOMAP is an extension that automatically selects the parameters k or ϵ . [39] Assuming a result of ISOMAP for a particular parameter value is available, a new parameter will be considered and p-ISOMAP will iteratively update the solution by:

1. Adding/removing neighbors for every vertex in the neighborhood graph regarding the new parameter. This operation has time complexity $O(n \max_i |\Delta e_i|)$, where $|\Delta e_i|$ is the number of inserted/removed edges associated with sample x_i .
2. Update the shortest-paths described by the pairs of vertices in $F \subset Q$, where Q is the set of all pairs of vertices and F is the set of pairs associated with the updated edges in the previous step.
 - (a) If the parameter value increased in comparison to the previous one, edges might have been added. Updating the shortest-path tree consists in simply relaxing edges associated with any of the vertices in the pairs of F . The time complexity is loosely bounded by $O(|A|q|F|)$, where $|A|$ is the number of inserted edges and q is the maximum degree of vertices in G .
 - (b) If the parameter was decreased, then edges might have been removed and all pairs of vertices whose shortest-paths contain those edges are added to F . Identifying F can be done in $O(n^2)$ times steps to execute, given some optimization (see [39]). Now, Dijkstra’s can be selectively performed, requiring $O(n \max_i (|E'_i| \log |V_d(i)| + |E''_i|))$, where, given a vertex x_i , $V_d(i) = \{x_j | (x_i, x_j) \in F\}$, E'_i is the set of edges connecting whichever pair of vertices in $V_d(i)$ and E''_i is the set of edges connecting the vertices x_a and x_b such that x_a is in $V_d(i)$ and x_b is not.

- Finally, the last step is to update the eigenvectors and eigenvalues associated with the matrix B . They can be approximated using the Lanczos algorithm, an iterative refining method that takes an initial solution candidate. Officially, the time complexity of Lanczos algorithm is $O(n^2)$. [40] Yet, its timeframe for convergence also strongly depends on the solution candidate given. Assuming the previous inner matrix and the current one are similar, one might use the previous eigenvalues/vectors found as candidate. In this scenario, Lanczos's will likely finish much faster than other algorithms.

4.1.5 Evaluating Reductions

ISOMAP's aim to “unfold” the data set and, of course, it might distort dissimilarities between samples in the process. Hence drastically increasing Kruskal's stress. Clearly, this measure is unsuitable for evaluating ISOMAP. As proposed by L. Shi and J. Gu, Kruskal's stress can be easily updated to match ISOMAP nature. That is, to preserve dissimilarities within neighborhoods: [36]

Let $X_{n \times f}$ be a data set with n samples and f dimensions, $Y_{n \times p}$ its reduction to p dimensions, and δ_{ij} and $\hat{\delta}_{ij}$ the dissimilarity measurements defined for all samples i and j in both X and Y . Furthermore, given a sample i and $\beta \in \mathbb{N} \mid \beta \leq n$, take $\Omega = K \cup \Theta$, where K is the set containing the nearest neighbors of i and Θ is a set of random samples such that $|\Theta| \leq \beta \leq n$:

$$Stress_{ISOMAP} = \left[\frac{\sum_i \sum_{j \in \Omega} (\delta_{ij} - \hat{\delta}_{ij})^2}{\sum_i \sum_{j \in \Omega} \delta_{ij}^2} \right]^{\frac{1}{2}}$$

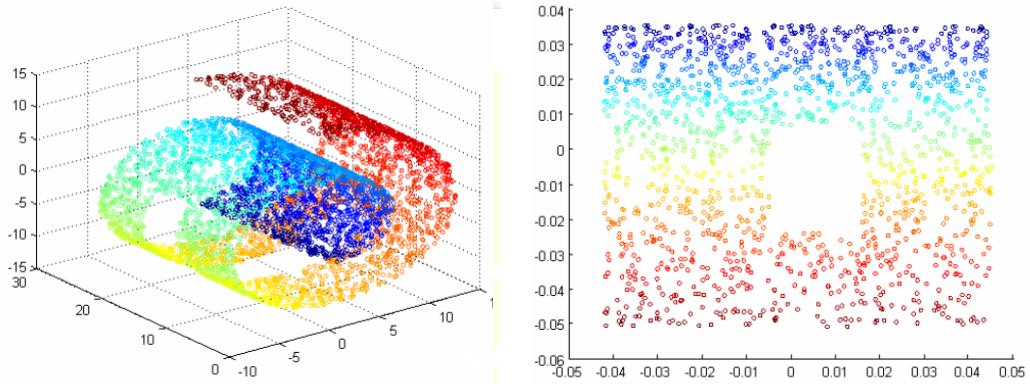
4.1.6 Applicability and Limitations

It is quite clear that ISOMAP outperforms PCA on the data set S . Unfortunately, this cannot be projected for a generic case considering the strict and artificial nature of S . In real-world data sets, the application of ISOMAP may lead to poor low-dimensional embeddings. [41] Below are listed some the issues that have great influence over ISOMAP's results:

Manifold Assumption It refers to the initial assumption that the data lies on a low-dimensional manifold. Although wildly exploited by many authors, it is difficult to assert whether such assumption holds or not in real-world data sets. [42] Furthermore, even if the data roughly lies on a manifold, discontinuities in the data pattern can characterize the manifold as non-smooth. In these situations, graphs with edges that disrespect locality would be extracted and, hence, poor low-dimensional representations are likely to be produced.

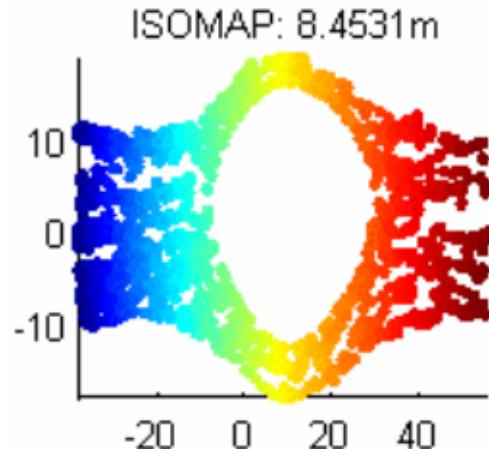
Manifold Convexity "ISOMAP relies on the data being geodesically convex."

[43] This issue becomes clear when dealing with data sets with "holes" in it that occupy sufficiently large areas, resulting in a neighborhood graph with great disconnected areas that require paths with great curvature to get around. Lerman demonstrated how reducing non-convex data sets with ISOMAP can easily yield distorted results: [44]



(a) The non-convex Swiss-roll.

(b) The expected reduction.



(c) The actual reduction.

Noise Noise in data is expressed by samples that somehow diverge from their neighborhood (outliers). As these samples become farthest from its original neighborhood, the chances of being linked to samples from other neighborhood increase, possibly decreasing the quality of the solution. A possible solution is to remove these samples during the pre-processing stage [41].

The image bellow illustrates an attempt to use ISOMAP to reduce the $M_{n=.4}$, which is the Swiss-roll data set subjected to a noise factor of .8:

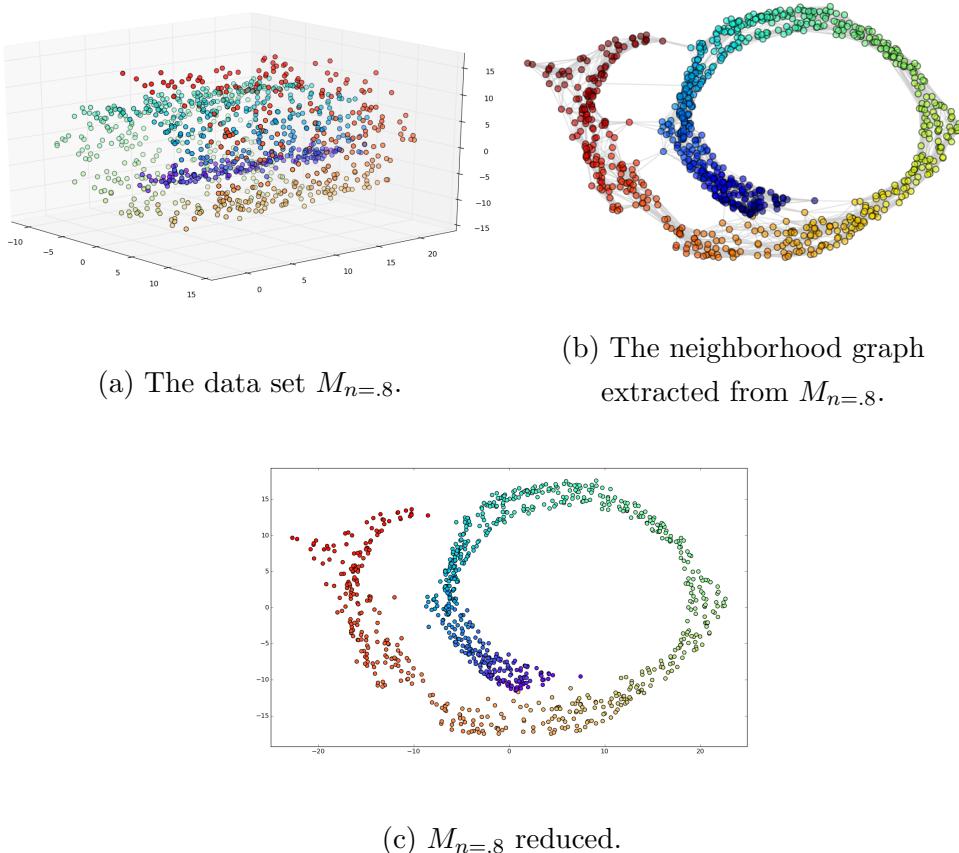


Figure 4.7: The ISOMAP applied on a noisy data set.

Observe how some outliers were sufficiently far from their original neighborhood in the graph 4.7b, to the point that Nearest-Neighbor Search maintained the edge between them and samples of completely different colors. MDS then attempted to maintain this dissimilarity, resulting in a deformed reduction (figure 4.7c).

4.2 Classification and Regression Over Data Sets Reduced with ISOMAP

Just as in 3.4, this section reports reduction, classification and regression experiments made in common data sets, following the same format as before.

4.2.1 The Swiss Roll Data Set

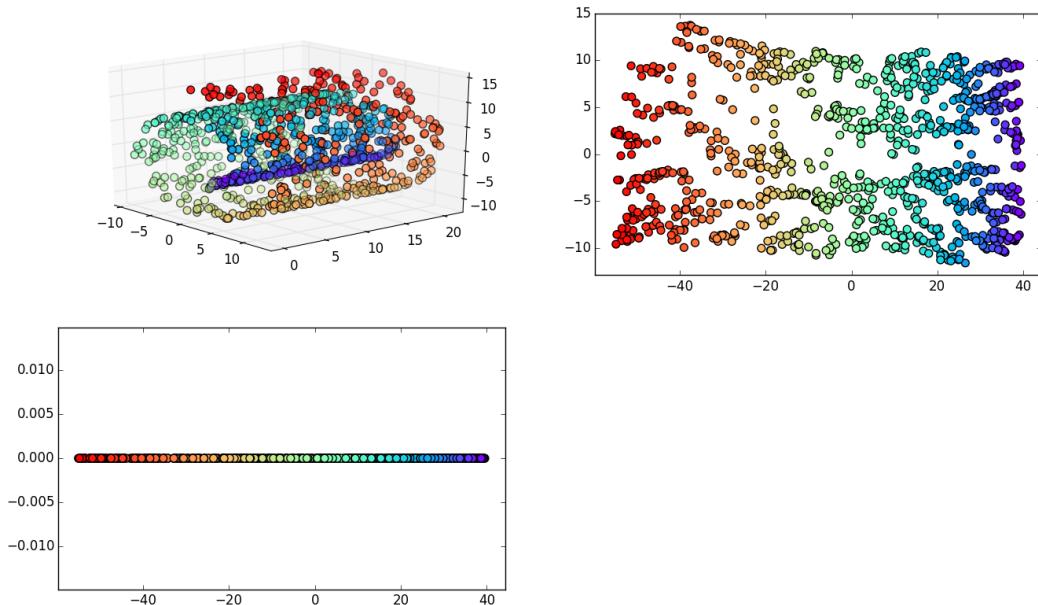


Figure 4.8: The Swiss Roll data set and its reductions to 2 and 1 dimensions, respectively.

	Orig. \mathbb{R}^3	\mathbb{R}^2	\mathbb{R}
Prediction accuracy	1	1	1
GridSearch time	15.53 s	415.72 s	387.98 s
Reduction time	-	0.51 s	0.48 s
Data size	23.44 KB	15.62 KB	7.81 KB

Table 4.4: Description of predictions and reduction performance for Swiss Roll.

4.2.2 The Iris Flower Data Set

Once again, the Iris flower data set, containing 150 samples and 5 features.

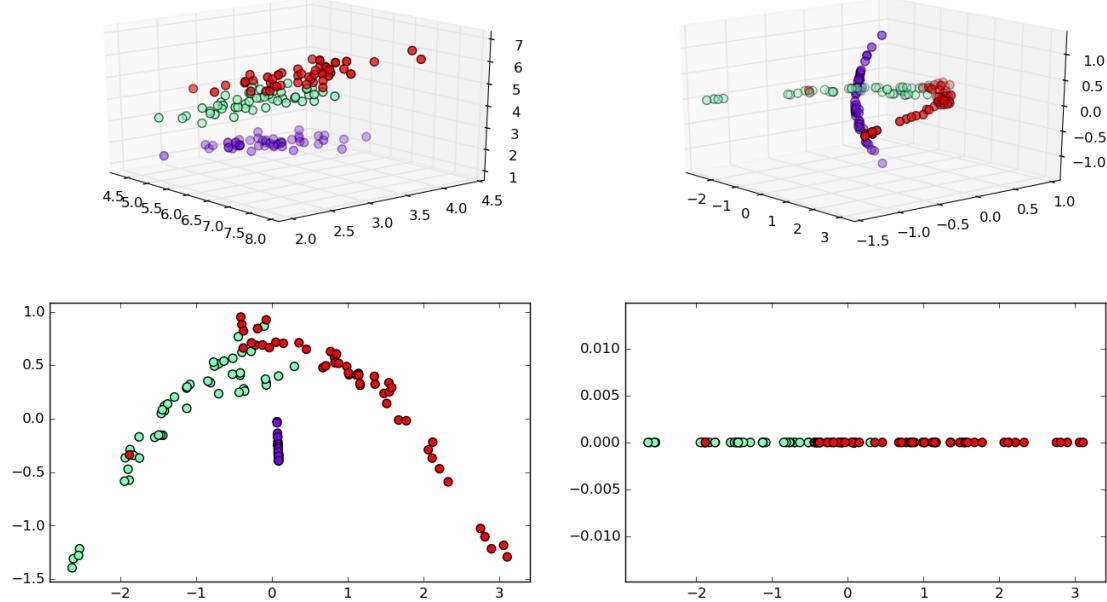


Figure 4.9: The Iris data set and its reductions to 3, 2 and 1 dimensions, respectively.

	Orig. \mathbb{R}^4	\mathbb{R}^3	\mathbb{R}^2	\mathbb{R}
Prediction accuracy	.99	.97 s	.97	.89
GridSearch time	.27 s	.25 s	.25 s	.27 s
Reduction time	-	.19 s	.19 s	.36 s
Stress	-	.474525	.569293	.655614
Data size	4.69 KB	3.52 KB	2.34 KB	1.17 KB

Table 4.5: Description of predictions and reduction performance for Iris.

4.2.3 The Glass Data Set

The Glass data set contains 214 samples and 11 features, where the first one is the identification number of a given sample and the last one is the class to which that sample belongs.

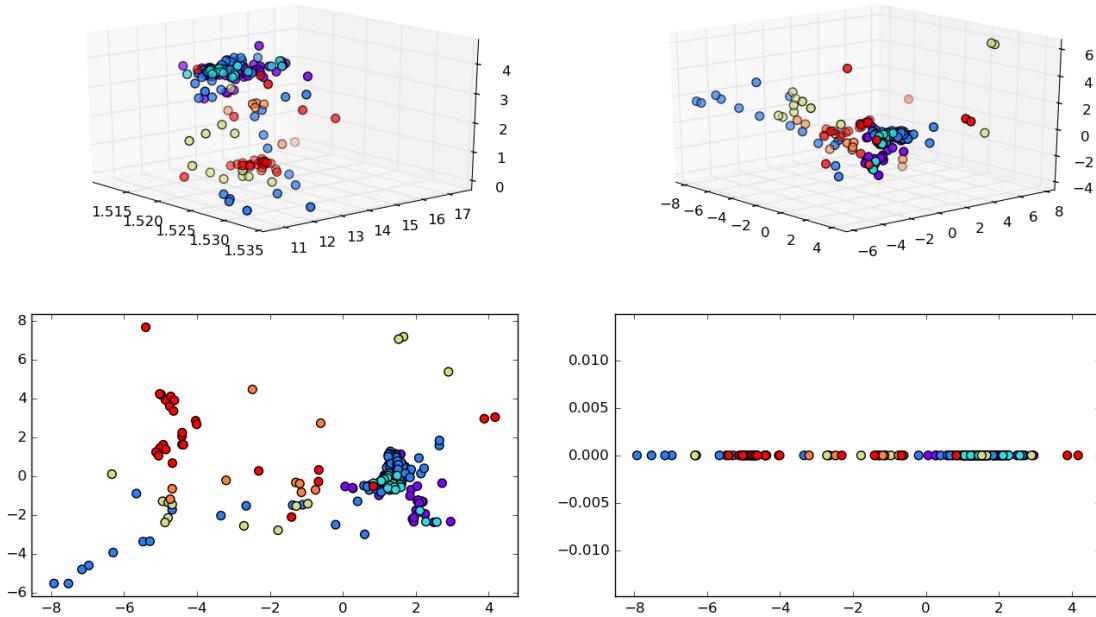


Figure 4.10: The Glass data set and its reductions to 3, 2 and 1 dimensions, respectively.

	Orig. \mathbb{R}^9	\mathbb{R}^8	\mathbb{R}^6	\mathbb{R}^4	\mathbb{R}^3
Prediction accuracy	.64	.63	.66	.65	.65
GridSearch time	.36 s	.35 s	2.77 s	2.36 s	1.26 s
Reduction time	-	.66 s	.65 s	.68 s	.64 s
Stress	-	.513717	.462840	.362030	.306092
Data size	15.05 KB	13.37 KB	10.03 KB	6.69 KB	5.02 KB

Table 4.6: Description of predictions and reduction performance for Glass.

4.2.4 The Digits Data Set

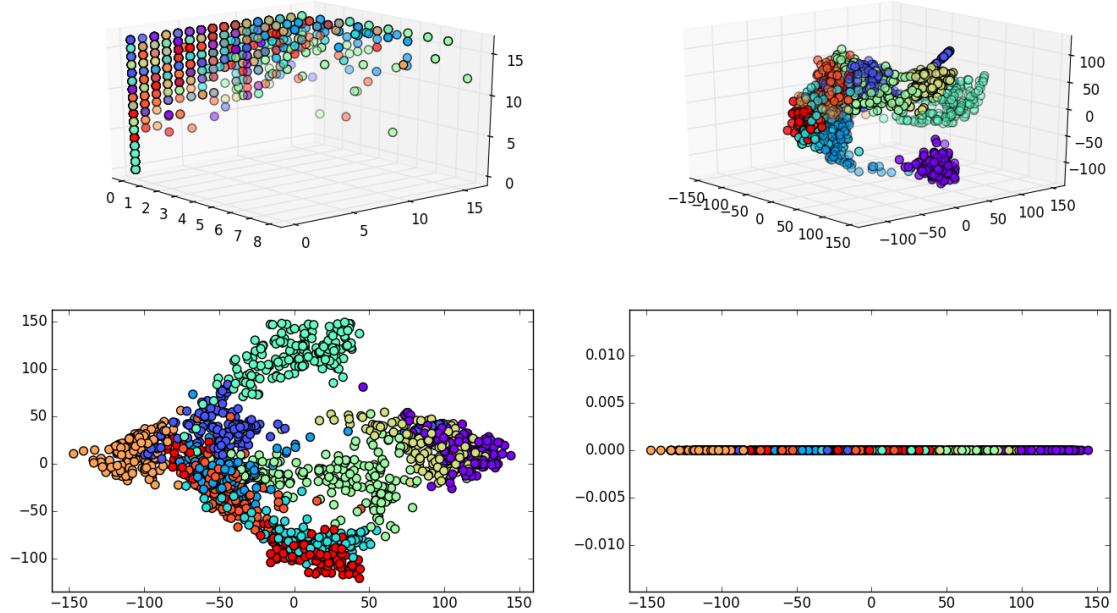


Figure 4.11: Digits data set and its reductions to 3, 2 and 1 dimension, respectively.

	Orig. \mathbb{R}^{64}	\mathbb{R}^{10}	\mathbb{R}^3	\mathbb{R}^2	\mathbb{R}
Pred. accuracy	.98	.96	.91	.69	.45
GridSearch time	12.22 s	98.08 s	287.46 s	630.47 s	949.88 s
Reduction time	-	5.43 s	2.72 s	2.71 s	2.04 s
Data size	898.50 KB	140.39 KB	42.12 KB	28.08 KB	14.04 KB

Table 4.7: Description of predictions and reduction performance for Digits.

Notice that the data set was reduced to only 3 dimensions, while maintaining 91% of accuracy (remember that dimensionality reduction with PCA would reduce accuracy to 74%). Accuracy loss is still observable when reducing it to two or one dimensions, although it is less intense than losses caused by linear reduction.

4.2.5 The Dermatology Data Set

The Dermatology data set, composed by 366 samples and 35 features. Here, the target is to correctly diagnose the erythemato-squamous diseases.

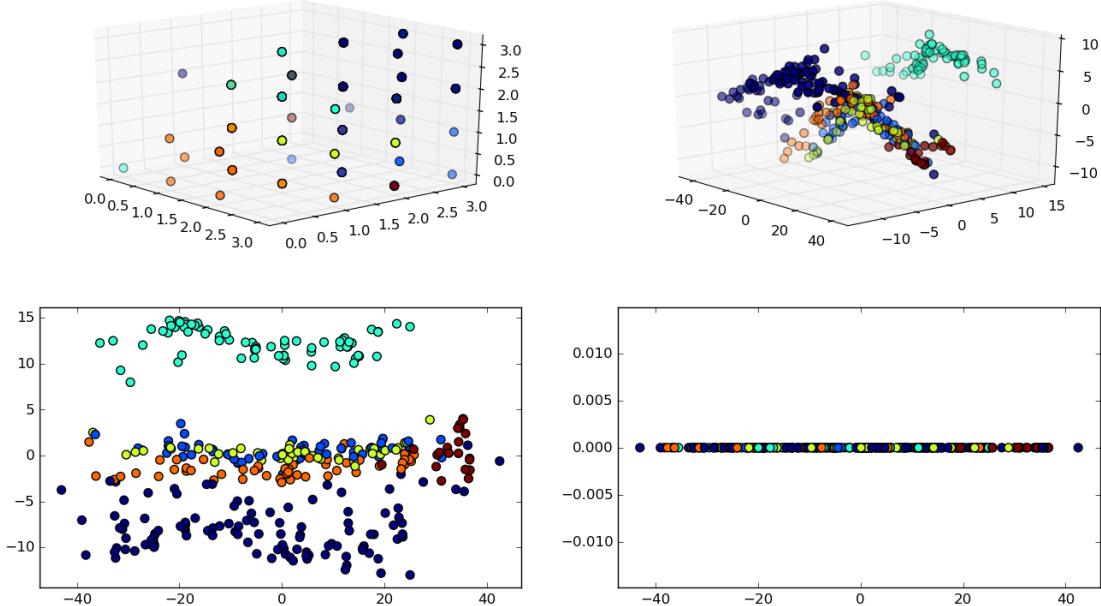


Figure 4.12: The Dermatology data set and its reductions to 3, 2 and 1 dimensions, respectively.

	Orig. \mathbb{R}^{34}	\mathbb{R}^{20}	\mathbb{R}^{10}	\mathbb{R}^3	\mathbb{R}^2
Prediction accuracy	.953552	.915301	.907104	.827869	.784153
GridSearch time	.67 s	.76 s	2.06 s	27 s	44.1 s
Reduction time	-	5.1 s	4.89 s	4.91 s	4.96 s
Stress	-	.771817	.597724	.465458	.416376
Data size	97.22 KB	57.19 KB	28.59 KB	8.59 KB	5.72 KB

Table 4.8: Description of predictions and reduction performance for Dermatology.

4.2.6 The Leukemia Data Set

Extracted from mldata, the Leukemia data set contains 72 samples and 7130 features. The first 7129 features are expression levels of the genes in a given patient. The 7130th feature $t \in \{-1, 1\}$ indicate which of two variants of leukemia is present in the sample (AML, 25 samples, or ALL, 47 samples). [45]

The Results in this experiment are quite impressive: although visualization has barely changed from 7129 to 3 dimensions, a 7129-dimensional space was reduced to a 10-dimensional one with only 10% of accuracy loss.

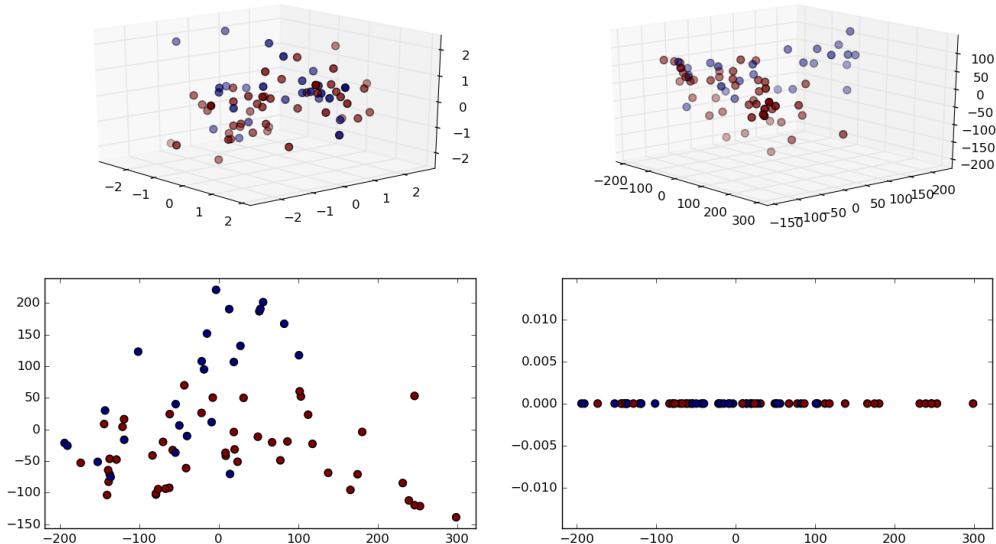


Figure 4.13: The Leukemia data set and its reduction to 3, 2 and 1 dimensions, respectively.

	Orig. \mathbb{R}^{7129}	\mathbb{R}^{30}	\mathbb{R}^{20}	\mathbb{R}^{10}
Pred. accuracy	.99	.88	.85	.88
GridSearch time	2.42 s	.26 s	.26 s	141.13 s
Reduction time	-	.06 s	.04 s	.04 s
Data size	4010.06 KB	16.88 KB	11.25 KB	5.62 KB

Table 4.9: Description of predictions and reduction performance for Leukemia data set.

4.2.7 The WDBC Data Set

The Wisconsin Diagnostic Breast Cancer (WDBC) data set, containing 596 samples and 32 features computed from a breast mass. WDBC has benign 357 samples and 212 malignant. During this experiment, the first feature was disregarded, as it represents the identification numbers of the samples.

Regarding the experiment, observe how prediction accuracy consistently increased when the data set was reduced.

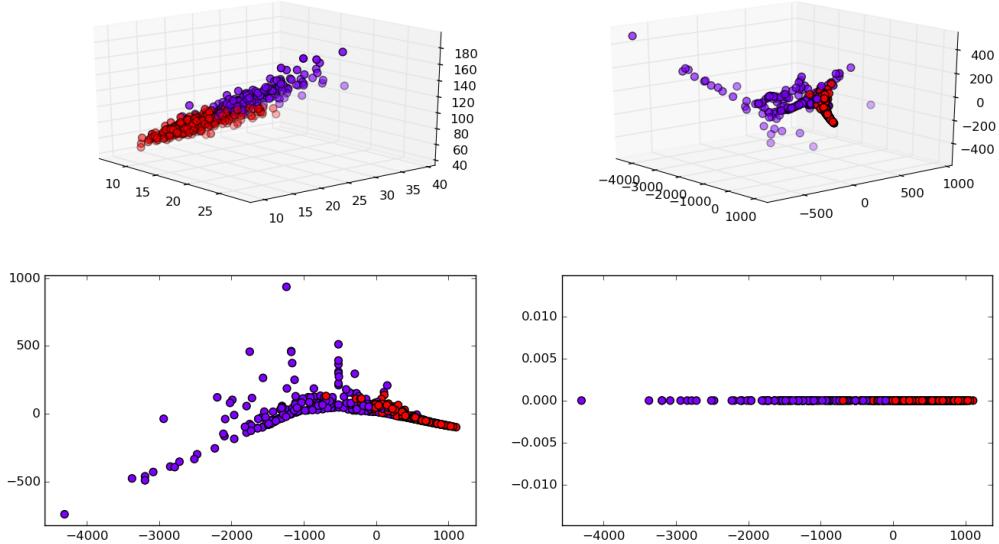


Figure 4.14: The WDBC data set and its reduction to 3, 2 and 1 dimensions, respectively.

	Orig. \mathbb{R}^{30}	\mathbb{R}^{20}	\mathbb{R}^{10}	\mathbb{R}^3	\mathbb{R}^2
Pred. accuracy	.63	.89	.8963	.8998	.8998
GridSearch time	0.58 s	0.46 s	0.36 s	0.36 s	0.36 s
Reduction time	-	3.31 s	3.25 s	3.26 s	3.31 s
Stress	-	.53	.465460	.378905	.352277
Data size	133.36 KB	88.91 KB	44.45 KB	13.34 KB	8.89 KB

Table 4.10: Description of predictions and reduction performance for Leukemia data set.

4.2.8 Diabetes Data Set

With 9 features observed from 768 samples collected from residents of Arizona, USA, the Diabetes data set seeks to diagnose whether a patient (sample) shows signs of diabetes (class 1) according to World Health Organization criteria or not (class 0). It is also important to mention that only 268 samples belong, in fact, to class 1.

In this experiment, Diabetes was compressed to one forth of its original size (48 KB to 12 KB), and no significant prediction accuracy was lost.

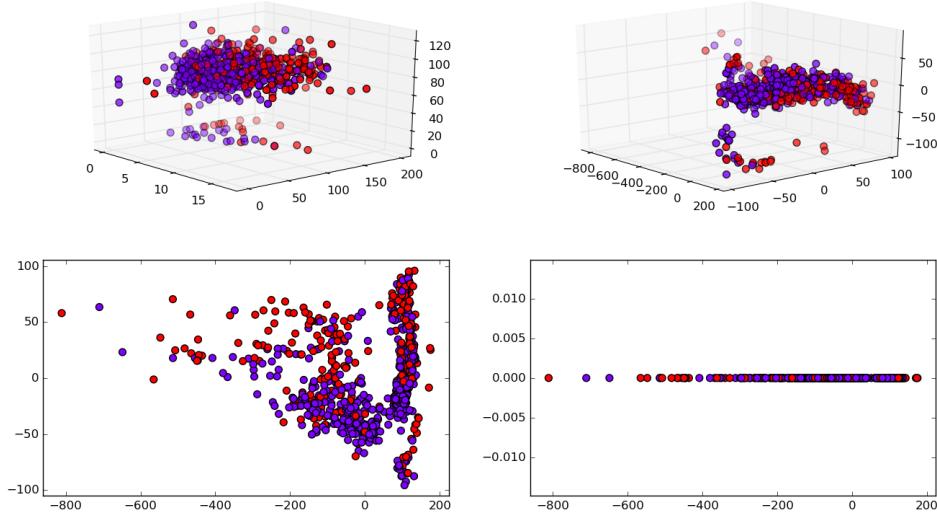


Figure 4.15: The Diabetes data set and its reduction to 3, 2 and 1 dimensions, respectively.

	Orig. (\mathbb{R}^8)	\mathbb{R}^6	\mathbb{R}^4	\mathbb{R}^2	\mathbb{R}
Pred. accuracy	.7239	.7174	.7252	.7148	.6588
GridSearch time	.79 s	.94 s	.58 s	.99 s	7.4 s
Reduction time	-	23.55 s	24.38 s	26.5 s	23.44 s
Stress	-	.2875	.2362	.3001	.4462
Data size	48 KB	36 KB	24 KB	12 KB	6 KB

Table 4.11: Description of predictions and reduction performance for Diabetes data set.

Final Considerations

In this report, the subject of dimensionality reduction was approached. First, focusing on linear methods and their capacity of extract features that maximize variance in a data set, we successfully reduced, visualized and learned well known data sets. Furthermore, we demonstrated how these methods would fail to reduce data sets that followed a nonlinear distribution.

In order to reduce nonlinear data sets, we introduced the ISOMAP algorithm, which takes advantage of properties commonly present in manifolds (e.g., linear locality, neighborhood) to map the data set of interest to an intermediate representation that only preserves dissimilarities of restrictive neighborhoods, “unfolding” the data set before applying a linear reduction method. We then proceeded to formally define ISOMAP’s implementation, to analyze its complexity and present some of its limitations, variations and applications. Finally, we demonstrated that ISOMAP can successfully reduce data sets that roughly lie on nonlinear manifolds, but it also strongly dependents on many conditions, such as the manifold assumption, manifold convexity and controlled data noise, severely affecting the number of problems to which it might be applied.

In conclusion, we have observed that although its limitations, ISOMAP represented an advance in manifold learning, being a highly regarded method in dimensionality reduction nowdays. In practice, its importance is clearly observed when considering the great number of machine learning libraries, languages and computational environments which implement it, as well as the great load of study done by many authors trying to apply, improve or extend ISOMAP.

Bibliography

- [1] scikit learn, “Svm: Maximum margin separating hyperplane.” [Online]. Available: http://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane.html.
- [2] P. R. C. D. Manning and H. Schutze, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.
- [3] J. Brownlee, “Practical machine learning problems.” [Online]. Available: <http://machinelearningmastery.com/practical-machine-learning-problems/>, 2013.
- [4] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*. A Bradford book, A Bradford Book, 2001.
- [5] J. Wang, *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer Berlin Heidelberg, 2012.
- [6] S. S. A. H. Renear and K. M. Wickett, “Definitions of dataset in the scientific and technical literature,” *ASIS&T 2010*, 2010.
- [7] L. Cayton, “Algorithms for manifold learning,” *Univ. of California at San Diego Tech. Rep*, pp. 1–17, 2005.
- [8] A. Ghodsi, “Dimensionality reduction a short tutorial,” *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada*, 2006.
- [9] S. Ross, *Introductory Statistics*. Elsevier Science, 2010.
- [10] T. Cox and M. A. A. Cox, *Multidimensional scaling*. Boca Raton, FL, USA: CRC Press, 2000.

- [11] W. Gander, *The Singular Value Decomposition*, December 2008.
- [12] J. M. Lee, *Manifolds and differential geometry*, vol. 107. American Mathematical Society Providence, 2009.
- [13] E. W. Weisstein, “Stereographic projection.” [Online]. Available: <http://mathworld.wolfram.com/StereographicProjection.html>.
- [14] C. Berge and E. Minieka, *Graphs and hypergraphs*, vol. 7. North-Holland publishing company Amsterdam, 1973.
- [15] W. Mayeda, *Graph Theory*. John Wiley & Sons, 1972.
- [16] T. Cormen, *Introduction to Algorithms*. Introduction to Algorithms, MIT Press, 2001.
- [17] M. J.Golin, “Lecture 15: The floyd-warshall algorithm.” [Online]. Available: <http://www.cse.ust.hk/faculty/golin/COMP271Sp03/Notes/MyL15.pdf>, 2003.
- [18] P. Langley, *Elements of Machine Learning*. Machine Learning Series, Morgan Kaufmann, 1996.
- [19] L. W. Hosch, “Machine learning.” [Online]. Available: <http://global.britannica.com/technology/machine-learning>.
- [20] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. New York, NY, USA: Apress, 2015.
- [21] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.
- [22] J. Weston, “Support vector machine (and statistical learning theory) tutorial.” [Online]. Available: http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf.
- [23] P. Winston, “Learning: Support vector machine.” [Online]. Available: https://www.youtube.com/watch?v=_PwhiWxHK8o, 2010.

- [24] R. Rifkin, “Multiclass classification.” [Online]. Available: <http://www.mit.edu/~9.520/spring09/Classes/multiclass.pdf>, 2008.
- [25] “Svm multi-class classification.” [Online]. Available: <https://www.sec.in.tum.de/assets/lehre/ws0910/ml/slideslecture9.pdf>, 2008.
- [26] scikit learn, “Cross-validation: evaluating estimator performance.” [Online]. Available: http://scikit-learn.org/stable/modules/cross_validation.html.
- [27] scikit learn.
- [28] B. Rohrer, “How to choose algorithms for microsoft azure machine learning.” [Online]. Available: <https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm>, 2015.
- [29] J. Shlens, “A tutorial on principal component analysis,” *arXiv preprint arXiv:1404.1100*, 2014.
- [30] G. Dunteman, *Principal components analysis*. No. 69, Newbury Road, CA, USA: Sage, 1989.
- [31] L. I. Smith, “A tutorial on principal components analysis,” *Cornell University, USA*, vol. 51, p. 52, 2002.
- [32] S. Raschka, “Implementing a principal component analysis (pca) in python step by step.” [Online]. Available: http://sebastianraschka.com/Articles/2014_pca_step_by_step.html, April 2014.
- [33] T. Naes and E. Risvik, *Multivariate Analysis of Data in Sensory Science. Data Handling in Science and Technology*, Elsevier Science, 1996.
- [34] V. S. T. B. Joshua and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [35] V. D. Silva and J. B. Tenenbaum, “Global versus local methods in nonlinear dimensionality reduction,” in *Advances in neural information processing systems*, pp. 705–712, 2002.

- [36] L. Shi and J. Gu, “A fast manifold learning algorithm,” *Information Technology Journal*, vol. 11, no. 3, pp. 380–383, 2012.
- [37] S. M. J. Ham, D. D. Lee and B. Schölkopf, “A kernel view of the dimensionality reduction of manifolds,” in *Proceedings of the twenty-first international conference on Machine learning*, p. 47, ACM, 2004.
- [38] M. H. C. Law and A. K. Jain, “Incremental nonlinear dimensionality reduction by manifold learning,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 377–391, 2006.
- [39] H. L. J. Choo, C. K. Reddy and H. Park, “p-isomap: An efficient parametric update for isomap for visual analytics.,” in *SDM*, pp. 502–513, SIAM, 2010.
- [40] J. L. Y. Chen and J. Wang, *Machine Learning and Statistical Modeling Approaches to Image Retrieval*. The Information Retrieval Series, Springer US, 2004.
- [41] E. P. L. Maaten and J. Herik, “Dimensionality reduction: A comparative review,” *Journal of Machine Learning Research*, vol. 10, no. 1-41, pp. 66–71, 2009.
- [42] T. Lin and H. Zha, “Riemannian manifold learning,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 5, pp. 796–809, 2008.
- [43] D. Donoho and C. Grimes, *When Does ISOMAP Recover the Natural Parameterization of Families of Articulated Images?* Technical report (Stanford University. Dept. of Statistics), Department of Statistics, Stanford University, 2002.
- [44] G. Lerman, “Manifold learning techniques: so which is the best?.” [Online]. Available: http://www.math.ucla.edu/~wittman/mani/mani_presentation.pdf, 2005.
- [45] F. Ducatelle, “Datasets for data mining.” [Online]. Available: <http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html>.