Lecture 12: Dimensionality reduction

Introduction to Machine Learning

Sophie Robert

L3 MIASHS — Semestre 2

2023-2024

Introduction

Definition

Dimensionality reduction

Dimensionality reduction is the transformation of the data from a **high**-dimensional space into a **lower**-dimension space, with as little information loss as possible.

Definition

Dimensionality reduction

Dimensionality reduction is the transformation of the data from a **high**-dimensional space into a **lower**-dimension space, with as little information loss as possible.

We want to:

- Reduce the number of features
- Retain as much information as possible

Motivation

Why do we need to reduce dimension ?

Question

Why do you think we need to reduce dimensions ?

Why do we need to reduce dimension ?

Question

Why do you think we need to reduce dimensions ?

- Computation time

Why do we need to reduce dimension ?

Question

Why do you think we need to reduce dimensions ?

- Computation time
- Easier data visualization

Why do we need to reduce dimension ?

Question

Why do you think we need to reduce dimensions ?

- Computation time
- Easier data visualization
- Possible unrelated features acting as noise

Why do we need to reduce dimension ?

Question

Why do you think we need to reduce dimensions ?

- Computation time
- Easier data visualization
- Possible unrelated features acting as noise
- Possible correlated features that do not bring any new information to solve the task

Why do we need to reduce dimension ?

Question

Why do you think we need to reduce dimensions ?

- Computation time
- Easier data visualization
- Possible unrelated features acting as noise
- Possible correlated features that do not bring any new information to solve the task
- **The curse of dimensionality**

Curse of dimensionality

Curse of dimensionality

The **curse of dimensionality*** refers to various phenomena that arise when analyzing data in high-dimensional space.

Curse of dimensionality

Curse of dimensionality

The **curse of dimensionality*** refers to various phenomena that arise when analyzing data in high-dimensional space.

The main problem is that the **dataset becomes sparse**: we do not have enough combinations of values to properly learn from it !

Methods for feature reduction

Methods for feature reductions

Question

What is in your opinion possible approaches to reduce the number of features ?

Methods for feature reductions

Question

What is in your opinion possible approaches to reduce the number of features ?

- **Removing** some features

Methods for feature reductions

Question

What is in your opinion possible approaches to reduce the number of features ?

- **Removing** some features
- **Projecting the features** into a lower dimensional space.

Feature selection

Feature selection

Feature selection consists in selecting the subset of features that bear the most information and removing the features that bear little information from the dataset.

Feature selection

Feature selection

Feature selection consists in selecting the subset of features that bear the most information and removing the features that bear little information from the dataset.

Question

Do you remember what features we removed from the Pokemon dataset ? Why and how ?

Feature selection

Feature selection

Feature selection consists in selecting the subset of features that bear the most information and removing the features that bear little information from the dataset.

Question

Do you remember what features we removed from the Pokemon dataset ? Why and how ?

We need some objective rules to define what subset of variables to keep.

Feature projection

Feature projection

Feature projection* (also called **feature extraction**) transforms the data from the high-dimensional space to a space of fewer dimensions.

Feature projection

Feature projection

Feature projection* (also called **feature extraction**) transforms the data from the high-dimensional space to a space of fewer dimensions.

We need to project the data into another space which is a combination of the features **with as little information loss as possible**.

Feature projection

Feature projection

Feature projection* (also called **feature extraction**) transforms the data from the high-dimensional space to a space of fewer dimensions.

We need to project the data into another space which is a combination of the features **with as little information loss as possible**.

The transformation can be:

- Linear (PCA, NMF ...)
- Non-linear (t-SNE, ...)

Questions

Questions ?