Mathematical Foundations for Statistical Learning
-
Principal Components Analysis (PCA)

Atilla Kaan Alkan

Institut Polytechnique des Sciences Avancées, Ivry-sur-Seine

November 18, 2024

Table of Contents

Contents

Goals

- Provide a general introduction to statistical learning;
- Introduce some important tools and concepts for solving problems in statistical learning;
- Distinguish between different approaches to data modelling;
- Understand different tools behind statistical learning;
- Main background needed: basic notions in probabilities, statistics, linear algebra and programming;
- **Grading**: Tutorials (coefficient 1) and a final project (coefficient 3).

Contents

Unsupervised Learning

- So far, we have focused on supervised learning;
- In supervised learning, we are given examples (x_i, y_i), and we try to predict y for future x' s;
- In unsupervised learning, we are given only x_i' s, with no outcome y_i;
- Unsupervised learning is less well defined but consists of finding some structure in the x' s;
- Two most common types of unsupervised learning:
  - Finding lower-dimensional representations;
  - Finding clusters/groups.

Unsupervised Learning

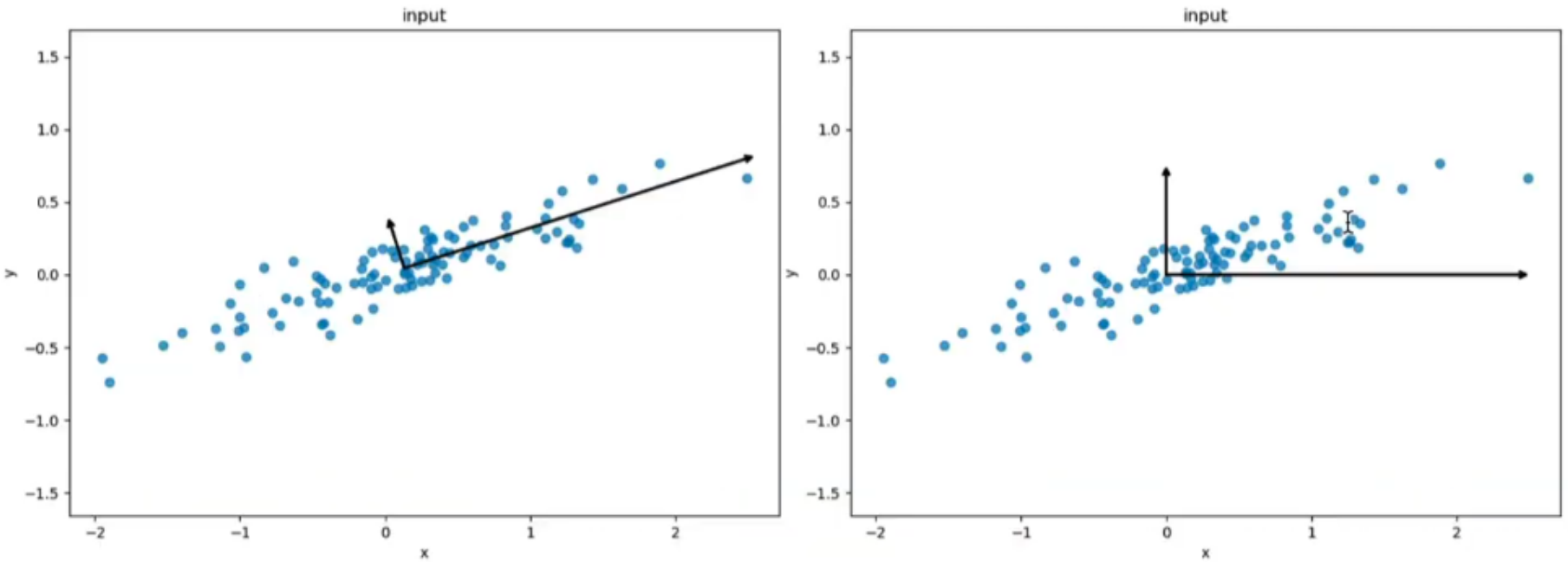
- In supervised learning, we can use the outcome y to evaluate performance reliably;
- This enables to:
  - Choose model settings;
  - estimate test performance.
- However, we do not have this luxury in unsupervised learning;
- A challenge is that often there is no standard way to evaluate the performance of an unsupervised method.

Contents

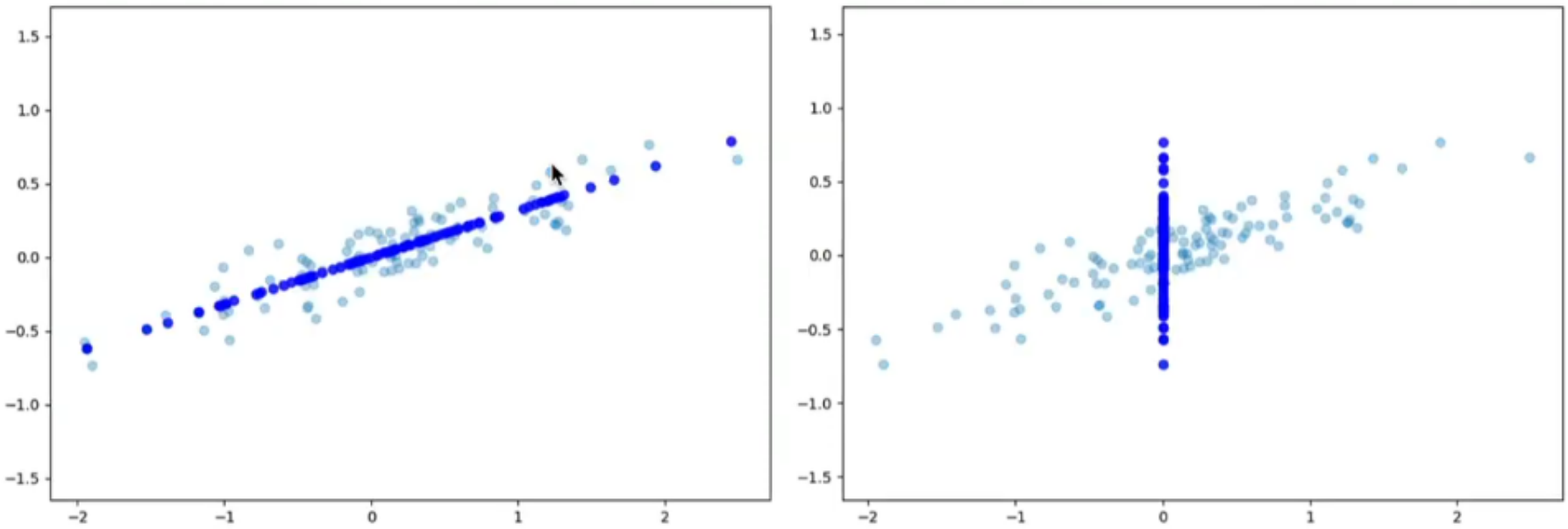
PCA: General Idea

- PCA is an **unsupervised** often used to reduce the dimensionality of the dataset by transforming a large set into a lower dimensional set that still contains most of the information of the large set.
- Find a transformation such that:
  - The transformed features are **linearly independent**;
  - **Dimensionality can be reduced** by taking only the dimensions with the **highest importance**;
  - Those newly found dimensions should **minimize the projection error**;
  - The projected points should have maximum spread, i.e., **maximum variance**.

PCA: General Idea



PCA: General Idea



Variance

- How much variation or spread the data has.

$$Var(X) = \frac{1}{n}\sum(X_i - \bar{X})^2$$

Covariance Matrix

- Indicates the level to which two variables.

$$Cov(X, Y) = \frac{1}{n} \sum (X_i - \bar{X})(Y_i - \bar{Y})^T$$

$$Cov(X, X) = \frac{1}{n} \sum (X_i - \bar{X})(X_i - \bar{X})^T$$

Eigenvectors and Eigenvalues

- Calculate the eigenvectors of $Cov(X, X)$;
- The eigenvectors point in the direction of the maximum variance, and the corresponding eigenvalues indicate the importance of its corresponding eigenvector;

$$A\tilde{v} = \lambda\tilde{v}$$

Steps of the PCA algorithm

1. Subtract the mean from X;
2. Calculate $Cov(X, X)$;
3. Calculate eigenvectors according to their eigenvalues in decreasing order;
4. Choose first k eigenvectors and that will be the new k dimensions;
5. Transform the original n-dimensional data points into k dimensions; (=projections with dot product).

Implementation in Python

- See tutorial session for implementation from scratch!