Felix Reinbott

Dimension reduction, clustering and more

An overview of some unsupervised learning techniques

Warnemünde, 26.09.2023

Institute for Mathematical Stochastics
Otto-von-Guericke-University Magdeburg

A few examples of typical problems from applications

Detecting insurance fraud

Assume we have requests for insurance offers based on user criteria like age, height, etc.
Some people do multiple requests by changing parameters slightly to obtain better offers.
Can we find the requests that belong to one person?

A few examples of typical problems from applications

Detecting insurance fraud

Assume we have requests for insurance offers based on user criteria like age, height, etc. Some people do multiple requests by changing parameters slightly to obtain better offers. Can we find the requests that belong to one person?

Image compression

Assume we have 256×256 pixel images of a certain handwritten number. Can we compress these images in a memory-efficient way? It might be plausible that these images lie in some low dimensional subspace of $\mathbb{R}^{256 \cdot 256}$.

A few examples of typical problems from applications

Detecting insurance fraud

Assume we have requests for insurance offers based on user criteria like age, height, etc. Some people do multiple requests by changing parameters slightly to obtain better offers. Can we find the requests that belong to one person?

Image compression

Assume we have 256×256 pixel images of a certain handwritten number. Can we compress these images in a memory-efficient way? It might be plausible that these images lie in some low dimensional subspace of $\mathbb{R}^{256 \cdot 256}$.

Detecting prototypical customers

A supermarket chain wants to identify $k \in \mathbb{N}$ archetypes of customers to optimize their market layout. Is it possible to find k archetypes of customers in a reasonable manner?

A general outline of unsupervised learning techniques

Assume that we have data $x_1, \ldots, x_n \in \mathbb{R}^d$, d "large", that are realizations of an i.i.d. sample from some probability measure $\mathbb{P}^X$ that is unknown.

A general outline of unsupervised learning techniques

Assume that we have data $x_1, \ldots, x_n \in \mathbb{R}^d$, d "large", that are realizations of an i.i.d. sample from some probability measure $\mathbb{P}^X$ that is unknown.

Usual task: Infer properties of interest about $\mathbb{P}^X$ from data

Usually we care about things like

- Does the data concentrate around some lower dimensional space or manifold?
- If $\mathbb{P}^X$ has a density f, does f have multiple local maxima?

A general outline of unsupervised learning techniques

Assume that we have data $x_1, \ldots, x_n \in \mathbb{R}^d$, d "large", that are realizations of an i.i.d. sample from some probability measure $\mathbb{P}^X$ that is unknown.

Usual task: Infer properties of interest about $\mathbb{P}^X$ from data

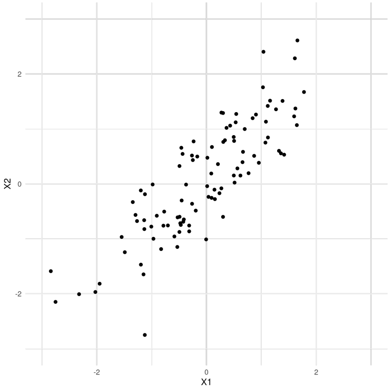
Usually we care about things like

- Does the data concentrate around some lower dimensional space or manifold?
- If $\mathbb{P}^X$ has a density f, does f have multiple local maxima?

Immediate problem compared to linear regression

We do not have a reference variable to check how good our model is.

Principal Component Analysis (PCA)



Sample shows most variance along line with positive slope.

Figure: An i.i.d. sample of 100 bivariate gaussians.

Principal Component Analysis (PCA)

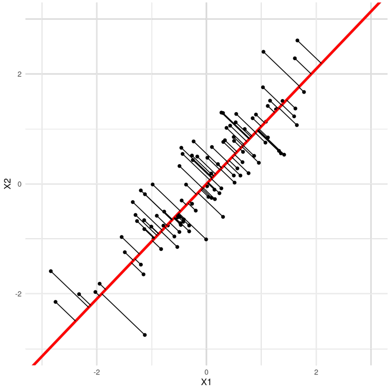


Figure: An i.i.d. sample of 100 bivariate gaussians.

Sample shows most variance along line with positive slope.

If we project on optimal line, we obtain new points which are one-dimensional and lie "close" to the original data.

Principal Component Analysis (PCA)

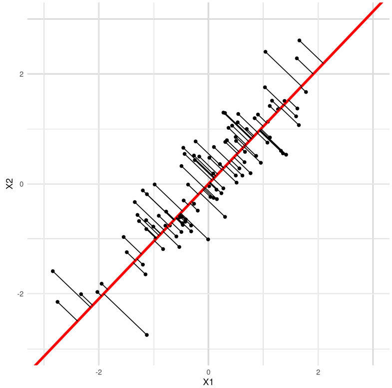


Figure: An i.i.d. sample of 100 bivariate gaussians.

Sample shows most variance along line with positive slope.

If we project on optimal line, we obtain new points which are one-dimensional and lie "close" to the original data.

→ Let's make that precise.

How to compute PCA

Goal: Find linear subspace that lies close to data

Assume that $X \in L^2$ is a d-dimensional random vector and we want to map X to some $p << d$ dimensional subspace that is best in the sense that

$$\mathbb{E}\left[\|X - PX\|_2^2\right]$$

is minimal and the orthogonal projection matrix P is a projection to the subspace.

How to compute PCA

Goal: Find linear subspace that lies close to data

Assume that $X \in L^2$ is a d-dimensional random vector and we want to map X to some $p << d$ dimensional subspace that is best in the sense that

$$\mathbb{E}\left[\|X - PX\|_2^2\right]$$

is minimal and the orthogonal projection matrix P is a projection to the subspace.

→ We can simplify that expression drastically!

A more convenient reformulation

Using standard linear algebra, it can be shown that minimizing the distance of X to some p dimensional subspace is the same as teratively solving

$$\max_{v_i \in \mathbb{R}^d : \|v_i\|_2 = 1} v_i^T \Sigma v_i, \quad \langle v_i, v_j \rangle = 0, i \neq j, i = 1, \ldots, p.$$

Solving this gives us that we choose v_i to be an eigenvector belonging to the i-th largest eigenvalue of Σ.

A more convenient reformulation

Using standard linear algebra, it can be shown that minimizing the distance of X to some p dimensional subspace is the same as teratively solving

$$\max_{v_i \in \mathbb{R}^d : \|v_i\|_2 = 1} v_i^T \Sigma v_i, \quad \langle v_i, v_j \rangle = 0, i \neq j, i = 1, \ldots, p.$$

Solving this gives us that we choose v_i to be an eigenvector belonging to the i-th largest eigenvalue of Σ.

Main takeaway: Projecting data on linear subspace is solving an eigenvalue problem!

→ We can easily apply this to data by estimating the covariance of the data!

An application: Hand written digits

asdf

Some extensions of PCA

asdf

A general overview of clustering

asdf

A simple clustering algorithm: Agglomerative clustering

asdf

Finding k prototypes with k-means clustering

asdf

Mixture models for soft-margin clustering

asdf

An application of the different clustering algorithms

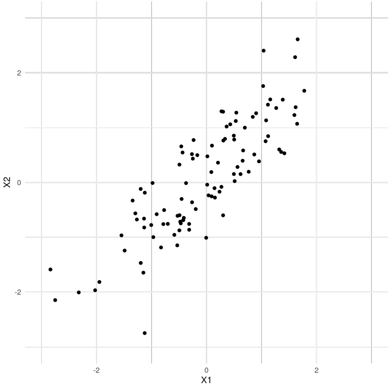


Figure: An i.i.d. sample of 100 bivariate gaussians.

Assume that we have data with existing variances.
We want to project down to a p-dimensional linear subspace.