TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

10.1 What is Dimensionality Reduction?

Machine Learning 1: Foundations

Marius Kloft *(TUK)*

Recap

Last week, we started to look into **unsupervised learning**:

- Clustering

This week:

- **Dimensionality Reduction**

Contents of this Class

Dimensionality Reduction

Example: Genome of Europeans

Novembre et al. (*Nature*, 2008) performed an experiment:

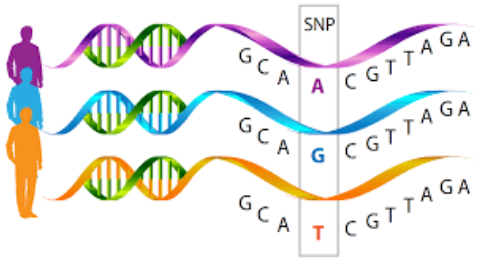
They collected blood samples

- from persons all over Europe

Extracted SNPs from blood

SNPs are point mutations

- capture most of the genetic variation



There exist millions of SNPs

- thus inputs $x_i \in \mathbb{R}^d$ with d in the millions

Finally they performed a dimensionality reduction...

The result was striking!

What is Dimensionality Reduction?

Definition

In **dimensionality reduction**, we want to represent the data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$

- in a lower-dimensional space $\mathbb{R}^k$ with $k < d$
- with as little loss of relevant information as possible!

This means:

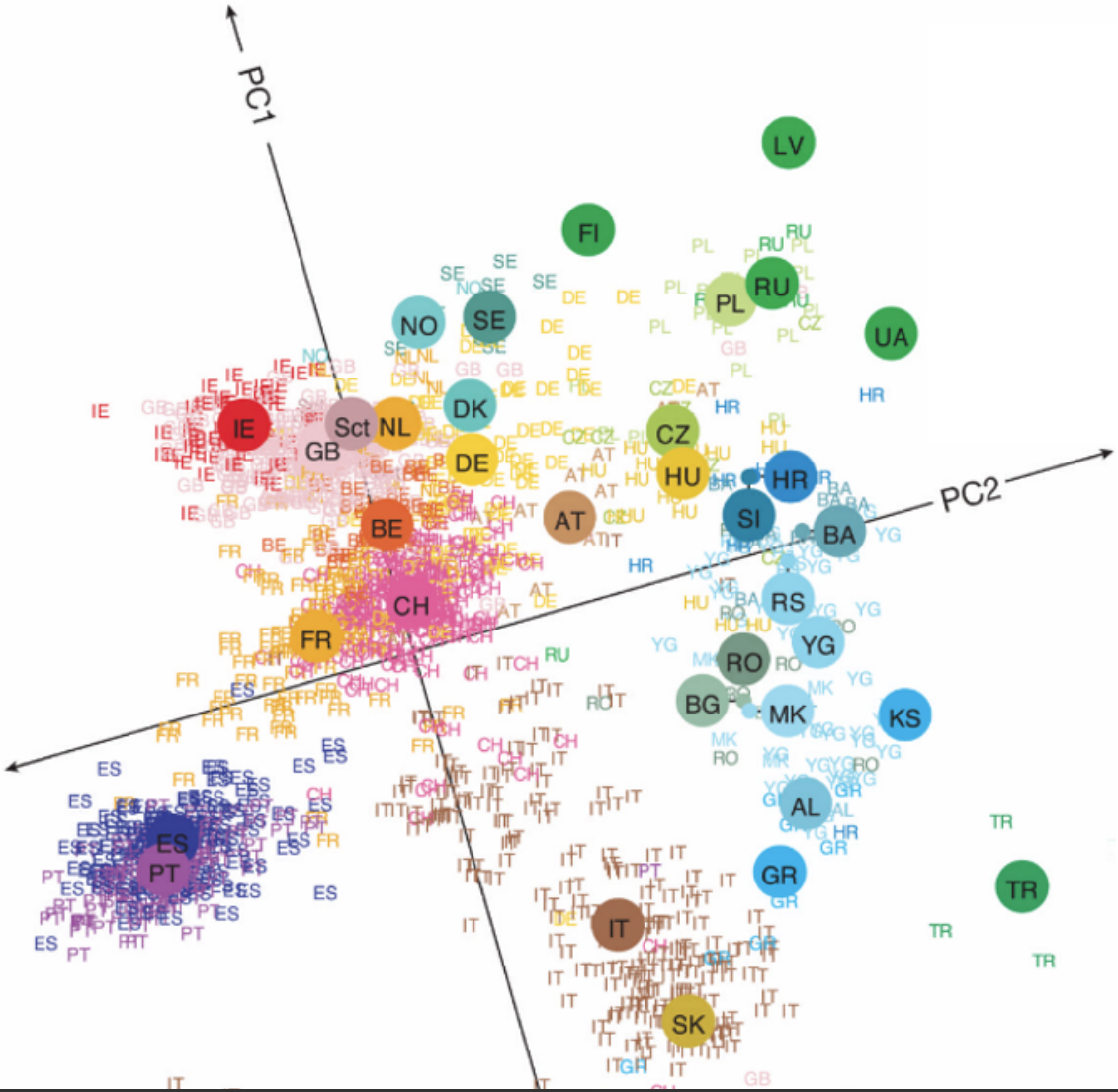
- We want to remove unnecessary (redundant) dimensions

Why Dimensionality Reduction?

Why Dimensionality Reduction?

- Data can be visualized (in two or three dimensions)
- Less storage, faster to compute with
- Less dimensions $\Rightarrow$ lower risk of overfitting

Genomes of Europeans Reduced to Two Dimensions:



Explanation of Result

Key finding:

- The genome of Europeans reflects the geometry of Europe.

Implication:

- Say, for instance, we want to find patterns in the DNA that increase the disease risk
- We perform a study on a heterogeneous pool of subjects (Germans, Spaniards, Britons, Dutch, ...)
- Say, just by chance, we happened to have more diseased Germans in our pool than Spaniards

DNA patterns that occur in Germans but not in Spaniards will seem to increase the disease risk!

The visualization from the previous slide was generated using a tool called **PCA**.