Manifold Learning

Introduction

Jairo Cugliari

Master Informatique

Parcours Data Mining

Université Lumière Lyon 2

Université de Lyon

Organisation pédagogique

Organisation

- Chaque séance : 1h CM + 2h Lab

Responsable

- Jairo Cugliari (`Jairo.Cugliari@univ-lyon2.fr`)

MCC

- Projet / Cours / Dossier en groupes (50%)
- Examen individuel (avec ordinateur 50%)

Bibliographie

Book

- C. Bishop, *Pattern Recognition and Machine Learning)*, 2007.
- T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2009.

E-learning

- `statweb.stanford.edu/~tibs/ElemStatLearn/`
- Andrew Ng’s course at Standford

Supervised learning

- Data
  - Response variables : $Y = (Y_1, \ldots, Y_m)^T$
  - Explicative variables : $X = (X_1, \ldots, X_p)^T$
- $\{(y_i, x_i), i = 1, \ldots, n\}$ with $x_i = (x_{i1}, \ldots, x_{ip})^T$
- Loss function : $L(y, \hat{y})$ (learn from errors)
- From a probabilist point of view

$$P(X, Y) = P(Y|X)P(X)$$

  where the target is $P(Y|X)$.

Available solutions depend on the size of p.

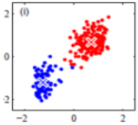
Unsupervised learning

- Data
  - Lack of response variable
  - $X = (X_1, \ldots, X_p)^T$
- n observations $\{x_i, i = 1, \ldots, n\}$ in dimension p
- How can we learn out of our errors ?
- From a probabilistic point of view the interest is on $P(X)$.

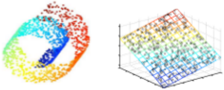
The solutions depend on the size p.

Goals of unsupervised learning

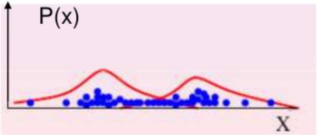
- Clustering: discover "clumps" of points



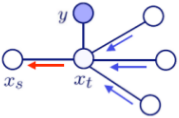
- Embedding: discover low-dimensional manifold or surface near which the data lives.



- Density Estimation. Find a function f such f(X) approximates the probability density of X, p(X), as well as possible.



- Finding good explanations (hidden causes)of the data;

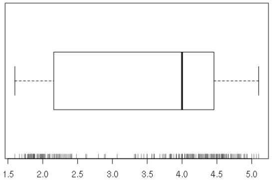


Example 1 : Density estimation

Old Faithful Geyser Data : waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

- Data : 272 obs × 2 vars
- Methods to analyze this data : summaries, plots, smth cleverer?

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85

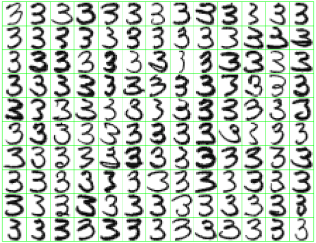


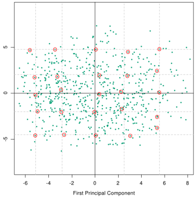
Example 2 : Principal Components Analysis (PCA)

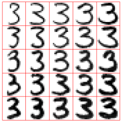
MNIST Handwritten Digits

- 658 Handwritten digits from ZIP US postal mail described by a 256 features
- Each image (16×16 8–bit gray scale) is 1 digit.
- After an SVD on the centred data matrix, we retain the projected points over the first two principal components.

Only linear relationships are taken into consideration.

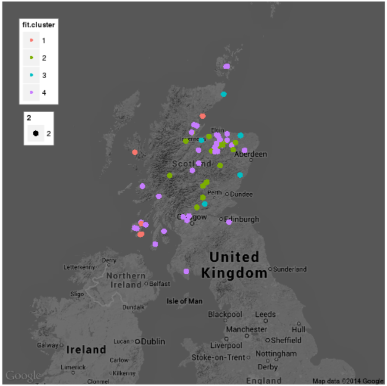






Example 3 : Clustering

- Data : 86 distilleries, 14 variables (taste scores), gps location of distilleries



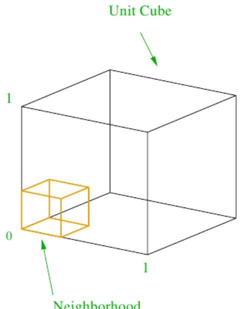
The Irish Whiskey Still

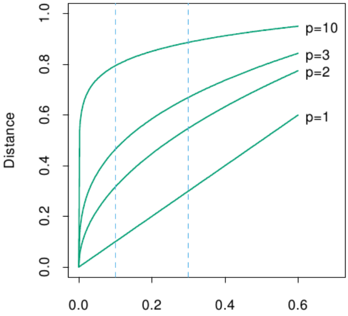
David Wilkie (1840)

Curse of dimensionality (Bellman, 1961)

Empty space phenomenon :

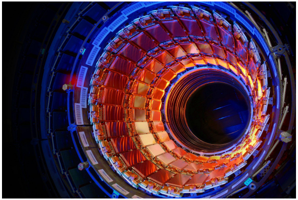
- When the dimension increases, the volume of the space increases so fast that the available data become sparse.
- Data needed to support a reliable result often grows exponentially with p.

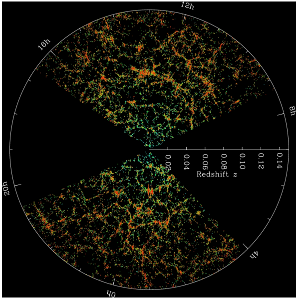




Data is ubiquitous I

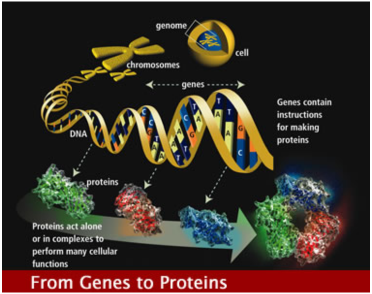
- LHC : 150M sensor sampling at 40M/sec. In one second over 600M collisions but 100 are of interest (<0.001%).
- Astronomy : photometric observations of 500M objects and spectra of 1M objects . (Sloan Digital Sky Survey)
- Remote Imagery : Satellite / Hyperspectral, e.g. resolve the earth surface to 1 meter accuracy; automatically discover natural resources





Data is ubiquitous II

- Web-based services : clickstreams are being tracked and sold
- Biotech Data : human genome, protein function and cell function (genomics to proteomics to ...?)
- Financial Data : high frequency financial data



The Manifold Hypothesis

We saw that :

- Data is ubiquitous and comes with lots of descriptors
- In high dimensions we suffer from the curse of dimensionality

We need then something to circumvent the curse. A popular choice is to assume that **data are aligned on a low dimensional** manifold embedeed on large host spaces.

Lectures

- Lecture 1 : Low-Dimensional Data : density estimation
- Lecture 2 : High-Dimensional Data (The manifold hypothesis)
- Lecture 3 : Dimensionality Estimation
- Lecture 4 : Metric Preservation (Metrics, MDS, Sammon's nonlinear mapping)
- Lecture 5 : Distance Preservation (Geodesic distance, ISOMAP)
- Lecture 6 : Fixed-grid Topology Preservation (SOM)
- Lecture 7 : Data-driven Topology Preservation (LLE)