Manifold Learning

High-Dimensional Data

Jairo Cugliari

Master Informatique

Parcours Data Mining

Université Lumière Lyon 2

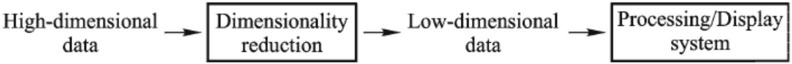
Université de Lyon

Olivetti faces: 400 total images, 64x64 size

Grayscale faces 8b, a few images of several different people

From objets to vectors (or points)

- Gray images of size $N = m \times n$ are seen as $N-$dimensional vectors (by row/col concatenation)
- Call $\mathcal{X} = \{x_1, \ldots, X_k\} \subset \mathbb{R}^n$ the vector set of images
- Geometrically, $\mathcal{X}$ is shown as a point cloud in a Euclidean space.



- Goals: sorting, recognition
- Key element : reduce N to a very low quantity (e.g. 2 or 3)

Curse of Dimensionality

In general, the sample size required to estimate a function of several variables to a given degree of accuracy grows exponentially with the increasing number of variables

Example

We want to cover the unit cube $[0, 1]^D$ with a $1/10$ grid.
We need 10^D points which grows exponentially with D!!!

A related fact : the empty space phenomenon

High-dimensional spaces are inherently sparse

Hypervolume of Cubes and Spheres in $\mathbb{R}^D$

Volume of the sphere with radius r and cube of size $2r$

$$V_{sph}^D(r) = \frac{\pi^{D/2} r^D}{\Gamma(D/2+1)} \qquad V_{cube}^D(r) = (2r)^D$$

Astonishingly, we get

$$\lim_{D \mapsto \infty} \frac{V_{sph}^D(r)}{V_{cube}^D} = 0$$

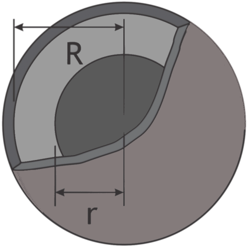
In high-dimensional spaces, the volume of the cube concentrates in its corners.

Hypervolume of a Thin Spherical Shell

Consider 2 concentric spheres with radii r and R, $r < R$.

(Relative) Hypervolume of the Thin

$$\frac{V^D_{sph}(R) - V^D_{sph}(r)}{V^D_{sph}(R)} = 1 - \left(\frac{r}{R}\right)^D$$



Which tends to 1 when $D \mapsto \infty$

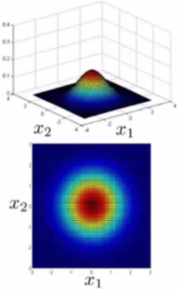
All the content of a $D-$dimensional sphere concentrates on its surface (which is only a $(D-1)$ dimensional manifold)

Tail probability of isotropic Gaussian distributions

Consider an isotropic Gaussian distribution in $\mathbb{R}^D$ which zero-mean and unit variance

Probability mass function

$$f(\mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{r^2}{2}\right), \quad r = \|\mathbf{y}\|$$



D	1	2	5	10	20
$Pr(r \geq 2)$	0.04550	0.13534	0.54942	0.94734	0.99995

(Semi) Diagonals of Cube

- Consider the $[-1, 1]^D \subset \mathbb{R}^D$ hypercube.
- Call $\mathbf{v}$ a (semi) diagonal from the center to a corner, so $\mathbf{v} = (\pm 1, \ldots, \pm 1)^T$

Angle between any $\mathbf{v}$ and a coordinate axis $\mathbf{e_i}$

$$\cos\theta = \left\langle \frac{\mathbf{v}}{\|\mathbf{v}\|}, \mathbf{e_i} \right\rangle = \frac{\pm 1}{\sqrt{D}}$$

The diagonals are nearly orthogonal to all coordinate axes (for large D) : visualization of high dimensional data by pairwise scatter plots may be misleading

Concentration of distances

Let $\mathbf{y} \in \mathbb{R}^D$ be a r.v. whose components are iid (and $E\|y\|^8 < \infty$)

Mean and Variance of the Euclidean Norm of y

$$\mu_{\|\mathbf{y}\|} \approx \sqrt{aD - b} \qquad \sigma^2_{\|\mathbf{y}\|} \approx b,$$

where a and b are known constants and the approximation terms are controlled (for large values of D).

Consequences

- Successive drawings of **y** yield almost the same norm
- Distance between any two vectors is approximately constant