Data Science
High-dimensional representation spaces

Curse and blessing of dimensionality

Stéphane Marchand-Maillet

Department of Computer Science

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Master en Sciences Informatiques - Autumn semester

Table of contents

What is the Data Science course about?

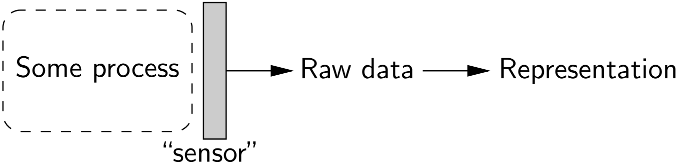
- ⋆ Study the modeling of phenomenons into digital (numerical, quantitative) data
- ⋆ Understand the geometrical and statistical properties of this data
- ⋆ Understand the geometrical and statistical properties of the spaces this data lives into
- ⋆ Analyze the data and develop tools for this analysis
- ⋆ Understand the assumptions made in the design of these tools
- ⋆ Work out the theory (in depth)

⚠ Reading: [5] (chap 1), [4] (chap 1.4)

Relationship to Machine Learning

- Data Science and Machine Learning are synonyms
- Data Science is the study of representation spaces within which Machine Learning acts
- ...

Typical data flow



Examples

Sensor		raw data		representation
⋆ Camera	⟶	image pixels	→	matrix
⋆ Population	⟶	poll results	→	matrix
⋆ Text documents	⟶	word occurrences	→	matrix
⋆ Social Network	⟶	relationships	→	matrix
⋆ Environment	⟶	measures	→	matrix

Representation space

We obtain data $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ where $\boldsymbol{x}_i \in \Omega \subseteq \mathbb{R}^D \quad \forall i \in [\![N]\!]$.

Statistical view

Data $\mathcal{X}$ is a N-sized sample of the underlying D-variate pdf $f_{\mathcal{X}} : \Omega \to \mathbb{R}^+$.
We declare N i.i.d random variables $\{X_i\}_{i \in [\![N]\!]}$, where $X_i \sim f_{\mathcal{X}}$.
Every data $\boldsymbol{x}_i$ is a sample of X_i

Algebraic view

Given $\mathcal{X}$, we form matrix $\boldsymbol{X} = \begin{pmatrix} | & & | \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_N \\ | & & | \end{pmatrix} \in \mathbb{R}^{D \times N}$ whose columns

$\boldsymbol{X}_{.j} = \boldsymbol{x}_j$ are the data vectors

Properties of the representation space

- ⋆ The representation space Ω is generally a subset of $\mathbb{R}^D$ (for some D)
- ⋆ It is a vector space and can be made an inner product space. In that case the inner product $\langle\cdot,\cdot\rangle$ induces a norm $\|\cdot\|$, which itself induces a distance function $d(\cdot,\cdot)$
- ⋆ This is therefore the favorable case where (Ω, d) is a metric space where learning can be performed

Questions:

- ⋆ What happens when D grows? when N grows?
- ⋆ How does this impact data modeling? How to cope?
- ⋆ How does that impact learning (if any)?

⇒ so-called “Curse of Dimensionality”

Note: Ω is also the base for a measurable space over which random variables may be defined

Data sparsity

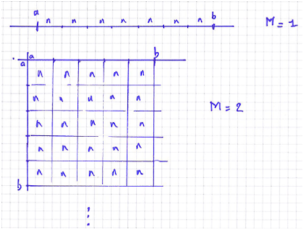
Given a population over hypercube $\Omega = [a, b]^D$ for some a, b, D. $\mathcal{X}$ is a sample of this population.
To obtain an empirical estimate (histogram) of the pdf, we quantize each dimension uniformly into b bins. To obtain a decent estimation quality we hope for an average of n samples per bin.

Q: How many samples do we need: what should N be ?

- ⋆ $D = 1 \rightarrow N \simeq n.b$
- ⋆ $D = 2 \rightarrow N \simeq n.b^2$

  $\vdots$

- ⋆ $D \rightarrow N \simeq n.b^D$



φ Example: $b = 10$, $n = 10$, $D = 6 \Rightarrow N \simeq 10^7$ samples

Data sparsity

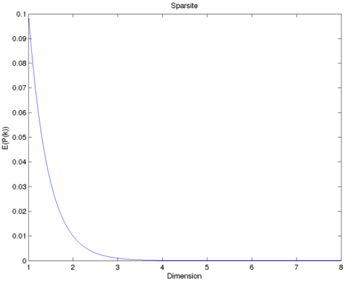
Say we now fix the number of samples to N

Q: What quality of estimation can be expected?

- ⋆ $D = 1 \rightarrow n = \frac{N}{b} \Rightarrow \mathbf{E}[\mathbb{P}(\boldsymbol{x}_i \in \mathrm{bin}_j)] \simeq \frac{1}{N}\frac{N}{b} = \frac{1}{b}$

  $\vdots$

- ⋆ Given $D \rightarrow n = \frac{N}{b^D} \Rightarrow \mathbf{E}[\mathbb{P}(\boldsymbol{x}_i \in \mathrm{bin}_j)] \simeq \frac{1}{b^D}$



$\Rightarrow$ Most bins are likely to be empty

Empty sphere

Given $\mathcal{B}_D(r)$ the hypersphere of radius r centered at the origin and included into the hypercube $\mathcal{C}_D(r) = [-r, r]^D$, draw N uniform samples from within the cube (i.e from $\mathcal{U}(\mathcal{C}_D(r))$)

Q: What is the proportion of samples falling into $\mathcal{B}_D(r)$?

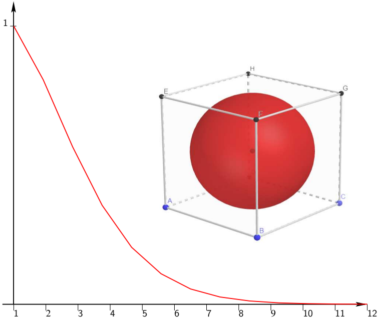
$$\text{vol}(\mathcal{B}_D(r)) = \frac{2r^D\pi^{D/2}}{D\Gamma(D/2)}$$

$$\text{vol}(\mathcal{C}_D(r)) = (2r)^D$$

$\Rightarrow$ Ratio $\frac{\text{vol}(\mathcal{B}_D(r))}{\text{vol}(\mathcal{C}_D(r))} = \frac{\pi^{D/2}}{D2^{D-1}\Gamma(D/2)}$

$\Rightarrow$

$$\lim_{D\to\infty} \frac{\text{vol}(\mathcal{B}_D(r))}{\text{vol}(\mathcal{C}_D(r))} = 0$$



$\Rightarrow$ The sphere is likely to be empty

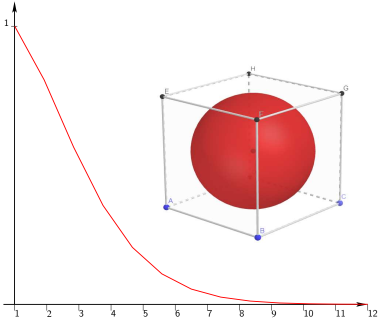
Note: $\Gamma(n)$ behaves like $n! \sim n^n$

Empty sphere

Given $\mathcal{B}_D(r)$ the hypersphere of radius r centered at the origin and included into the hypercube $\mathcal{C}_D(r) = [-r, r]^D$, draw N uniform samples from within the cube (i.e from $\boldsymbol{\mathcal{U}}(\mathcal{C}_D(r))$)

Q: What is the proportion of samples falling into $\mathcal{B}_D(r)$?

- ⋆ All the samples go in the part of the hypercube not in the sphere (the corners)
- ⋆ The volume of the cube is concentrated in it corners
- ⋆ All samples “escape” from the center
- ⋆ The relative volume of the hypersphere goes to zero



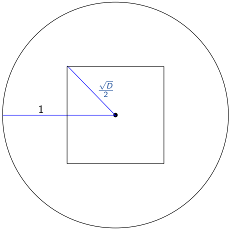
⇒ The sphere is likely to be empty

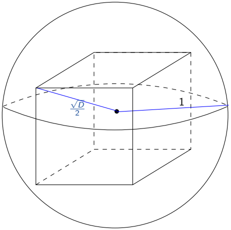
Corners of the hypercube

Given the reference unit hypersphere $\mathcal{B}_D(r)$ $(r = 1)$, consider the centered unit hypercube $\mathcal{C}_D(1) = [-\frac{1}{2}, \frac{1}{2}]^D$ with $\text{vol}(\mathcal{C}_D(1)) = 1$

- ⋆ $\mathcal{C}_D(1)$ has $n = 2^D$ corners
- ⋆ Every corner $\boldsymbol{x}_i$ has D coordinates $|\boldsymbol{x}_i(j)| = \frac{1}{2}$
- ⋆ Then $\|\boldsymbol{x}_i\|_2 = \frac{\sqrt{D}}{2}$
- ⋆

D	1	2	4	5	100
$\frac{\sqrt{D}}{2}$	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	1	$\frac{\sqrt{5}}{2}$	5





Distribution of volume

Q: Given a shape of measure r, what is the distribution of its volume?

Hypercube:

Ratio of volumes between the hypercubes of side length r and $(1-\varepsilon)r$ for $\varepsilon \in]0,1[$:
$\frac{\text{vol}(\mathcal{C}_D((1-\varepsilon)r))}{\text{vol}(\mathcal{C}_D(r))} = \frac{((1-\varepsilon)r)^D}{r^D} = (1-\varepsilon)^D$

$$\lim_{D\to\infty} \frac{\text{vol}(\mathcal{C}_D((1-\varepsilon)r))}{\text{vol}(\mathcal{C}_D(r))} = 0$$

Hypersphere:

Ratio of volumes between the hypersphere of radius r and $(1-\varepsilon)r$ for $\varepsilon \in]0,1[$:
$\frac{\text{vol}(\mathcal{B}_D((1-\varepsilon)r))}{\text{vol}(\mathcal{B}_D(r))} = \frac{2((1-\varepsilon)r)^D\pi^{D/2}}{D\Gamma(D/2)} \frac{D\Gamma(D/2)}{2r^D\pi^{D/2}} = (1-\varepsilon)^D$

$$\lim_{D\to\infty} \frac{\text{vol}(\mathcal{B}_D((1-\varepsilon)r))}{\text{vol}(\mathcal{B}_D(r))} = 0$$

Distribution of volume

Q: Given a shape of measure r, what is the distribution of its volume?

Any volume $\mathcal{V}$:

- ⋆ Consider decomposing the volume into arbitrarily small (hyper)cubic voxels
- ⋆ Shrink the voxel side length by factor $(1-\varepsilon)$

⇒ The voxel volume is shrunk by $(1-\varepsilon)^D$

⇒ The complete volume is shrunk by $(1-\varepsilon)^D$

⇒ The ratio of volumes tends to 0 as D increases

$$\lim_{D\to\infty} \frac{\mathrm{vol}(\mathcal{V}_D(r)) - \mathrm{vol}(\mathcal{V}_D((1-\varepsilon)r))}{\mathrm{vol}(\mathcal{V}_D(r))} = \lim_{D\to\infty} 1-(1-\varepsilon)^D = 1$$

⇒ The volume is concentrated near the surface of the shape

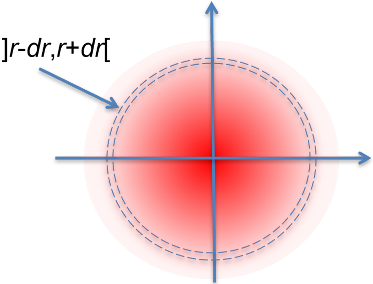
Gaussian egg

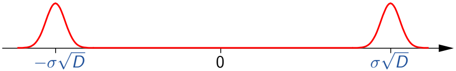
Q: Given a Normal distribution $\mathcal{N}(\mu, \Sigma)$, what is is density along radius r?

Coordinates are D iid r.v. $X_i \sim \mathcal{N}(0,1)$

Integrating along r is taking the norm. R is the r.v. attached to the radius:
$R^2 = \|X\|_2^2 = \sum_{i=1}^{D} X_i^2 \sim \chi^2(D)$

$$\Rightarrow \mathbf{E}[R^2] = D$$





We can considered the distribution centered and scaled (i.e $\mathcal{N}(0, \mathbf{Id}_D)$)

$\Rightarrow$ Most of the density is concentrated at radius $\sigma\sqrt{D}$

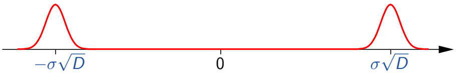
Gaussian egg

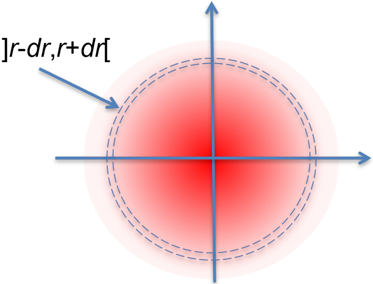
Q: Given a Normal distribution $\mathcal{N}(\mu, \Sigma)$, what is is density along radius r?

Gaussian Annulus theorem

Given $X \sim \mathcal{N}(0, \mathbf{Id}_D)$, for any $\beta \leq \sqrt{D}$,

$$\mathbb{P}\left[\left|\|X\|_2 - \sqrt{D}\right| \geq \beta\right] \leq 3e^{-c\beta^2} \qquad c > 0$$

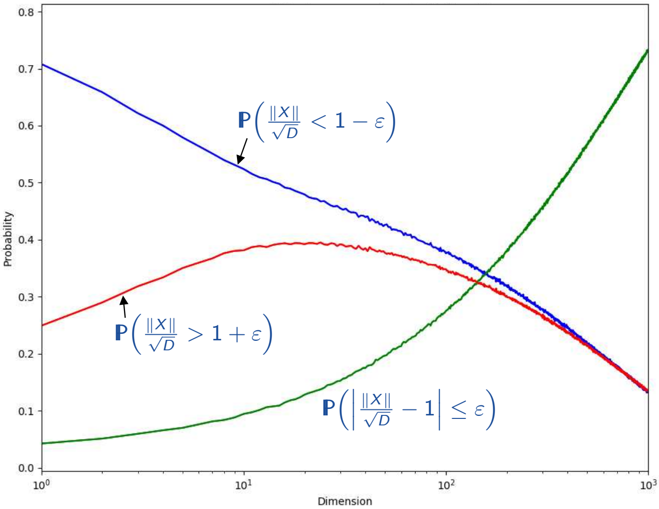




We can considered the distribution centered and scaled (i.e $\mathcal{N}(0, \mathbf{Id}_D)$)

$\Rightarrow$ Most of the density is concentrated at radius $\sigma\sqrt{D}$

Gaussian egg

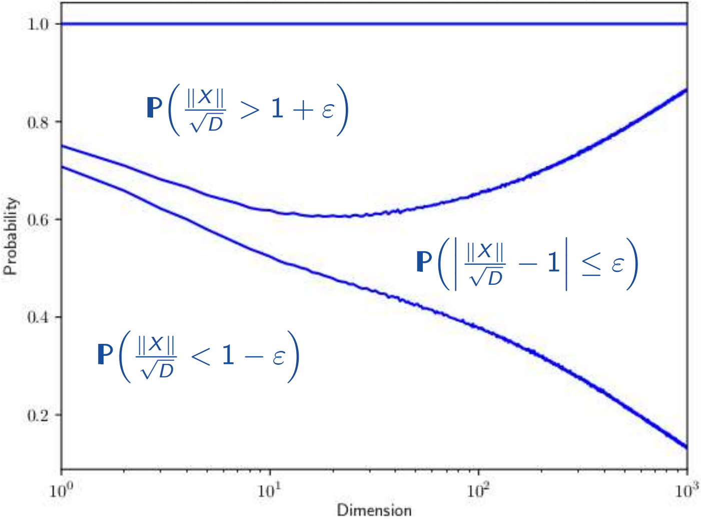


$\mathbb{P}\left(\frac{\|X\|}{\sqrt{D}} < 1 - \varepsilon\right)$

$\mathbb{P}\left(\frac{\|X\|}{\sqrt{D}} > 1 + \varepsilon\right)$

$\mathbb{P}\left(\left|\frac{\|X\|}{\sqrt{D}} - 1\right| \leq \varepsilon\right)$

Gaussian egg

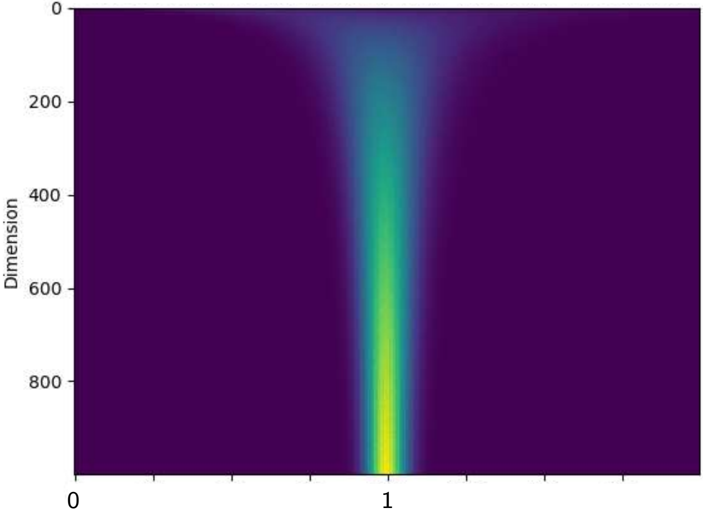


$\mathbf{P}\left(\frac{\|X\|}{\sqrt{D}} > 1+\varepsilon\right)$

$\mathbf{P}\left(\left|\frac{\|X\|}{\sqrt{D}} - 1\right| \leq \varepsilon\right)$

$\mathbf{P}\left(\frac{\|X\|}{\sqrt{D}} < 1-\varepsilon\right)$

Gaussian egg



Concentration of distances

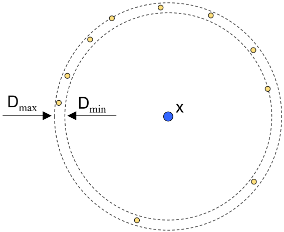
Given a data $\mathcal{X} \subset \Omega$, select one data x and its k nearest neighbors $\mathcal{V}^k_{\mathcal{X}}(\boldsymbol{x})$. D_{min} and D_{max} are the random variables attached to the distance from x to its closest and farthest k nearest neighbor then :

Theorem ([3])

If $\lim_{D\to\infty} \text{Var}\left(\frac{d(\boldsymbol{x},\boldsymbol{y})^p}{\mathbb{E}\, d(\boldsymbol{x},\boldsymbol{y})^p}\right) = 0$ *for* $0 < p < \infty$, *then for any* $\varepsilon > 0$

$$\lim_{D\to\infty} \mathbb{P}\left[\frac{D_{max} - D_{min}}{D_{min}} \leq \varepsilon\right] = 1$$

- ⋆ All neighbors are seen at distance D_{min} from $\boldsymbol{x}$
- ⋆ Discrimination power decreases exponentially
- ⋆ kNN and εNN structures are meaningless



Concentration of angles

Q: Draw vectors from a D-dimensional space, what is their expected angle?

To form 2 vectors, draw 4 points by declaring 4 iid random variables W, X, Y, Z over $\Omega \subset \mathbb{R}^D$ and compute

$$\mathbb{E} \cos(\underbrace{\angle(X-Y, Z-W)}_{\theta}) = \mathbb{E} \frac{\langle X-Y, Z-W \rangle}{\|X-Y\| \|Z-W\|}$$

Since W, X, Y, Z are iid, $X - Y$ and $Z - W$ are also iid and centered: $\mathbb{E}(X-Y) = \mathbb{E}(Z-W) = 0$ so the numerator is $\mathrm{cov}(X-Y, Z-W) = 0$, the denominator is a normalizer $\Rightarrow \mathbb{E}\cos(\theta) = 0$

Said otherwise

$\langle X-Y, Z-W \rangle = \langle X, Z \rangle - \langle X, W \rangle - \langle Y, Z \rangle + \langle Y, W \rangle \overset{\mathbb{E}}{=} 0$ since W, X, Y, Z are iid. The denominator is a normalizer $\Rightarrow \mathbb{E}\cos(\theta) = 0$

$\mathbb{E}\cos(\theta) = 0 \quad \Rightarrow \quad \theta \sim \frac{\pi}{2}$ $\underbrace{\text{~~“Every random triangle is a right triangle”~~}}_{\text{“Every 2 random vectors are likely to be quasi-orthogonal”}}$

Concentration of angles

Q: What is the expected angle in a random triangle?

To create a triangle, draw 3 points by declaring 3 iid random variables X, Y, Z and compute for any angle

$$\mathbb{E}\cos(\underbrace{\angle(X-Y, Z-Y)}_{\theta}) = \mathbb{E}\,\frac{\langle X-Y, Z-Y\rangle}{\|X-Y\|\|Z-Y\|}$$

Note: $X-Y$ and $Z-Y$ are no longer independent!

$$\frac{\langle X-Y,Z-Y\rangle}{\|X-Y\|\|Z-Y\|} \overset{\mathbb{E}}{=} \frac{\langle X-Y,Z-Y\rangle}{\langle X-Y,X-Y\rangle} = \frac{\langle X,Z\rangle-\langle X,Y\rangle-\langle Y,Z\rangle+\langle Y,Y\rangle}{\langle X,X\rangle+\langle Y,Y\rangle-2\langle X,Y\rangle} \overset{\mathbb{E}}{=} \frac{\langle Y,Y\rangle-\langle X,Y\rangle}{2\langle Y,Y\rangle-2\langle X,Y\rangle} = \frac{1}{2}$$

$\mathbb{E}\cos(\theta) = \frac{1}{2} \quad \Rightarrow \quad \theta \sim \frac{\pi}{3}$ "Every random triangle is a equilateral triangle"

Note: The above calculation takes a shortcut in the first equality but is correct

Impact for the sphere

Given unit ball $\mathcal{B}_D$, if X is a random point on its surface, given Y, call $\theta = \angle(X, Y)$:

- ⋆ if Y is sampled from Ω, then $\mathbb{E}\cos(\theta) = \frac{1}{2} \quad \Rightarrow \theta \sim \frac{\pi}{3}$
- ⋆ if Y is sampled from the surface of $\mathcal{B}_D$, then $\mathbb{E}\cos(\theta) = 0 \quad \Rightarrow \theta \sim \frac{\pi}{2}$

  $\Rightarrow$ "the volume of the sphere is contained at its equator"

  (for every "northern" direction X)

$\rightarrow$ Alternative proof for volume at surface of the sphere

φ Exercise: empirically verify these results

Note: To sample data over the unit hypersphere, sample a centrally symmetric distribution (e.g Gaussian) and normalize

Data hubness

Q: How many times a sample appears as k-nearest neighbor of another data?

Given $\mathcal{X}$, define

$$r_{ik}(\boldsymbol{x}_j) := \begin{cases} 1 & \text{if } x \in \mathcal{V}^k_{\mathcal{X}}(\boldsymbol{x}_i) \\ 0 & \text{otherwise} \end{cases}$$

then

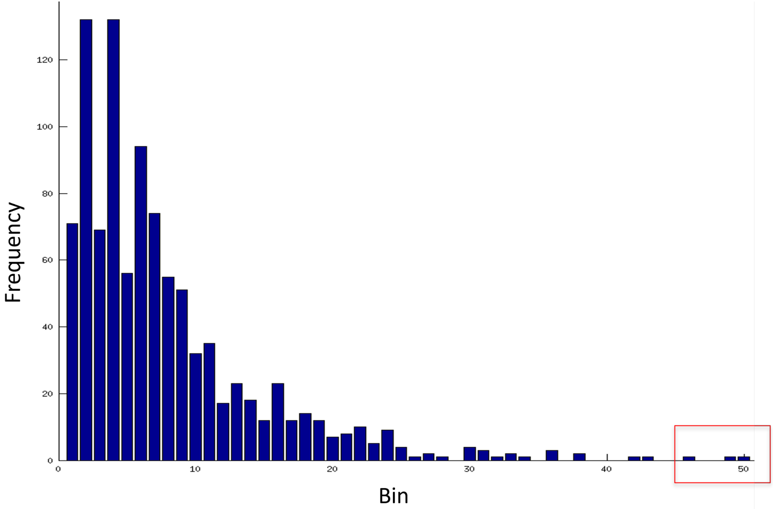
$$n_k(\boldsymbol{x}_j) := \sum_i r_{ik}(\boldsymbol{x}_j)$$

$\Rightarrow$ The distribution of values of n_k is skewed to the left [10]. A small number of samples appear in the neighborhood of many other samples.

$\Rightarrow$ These "hubs" are "close" to all other samples

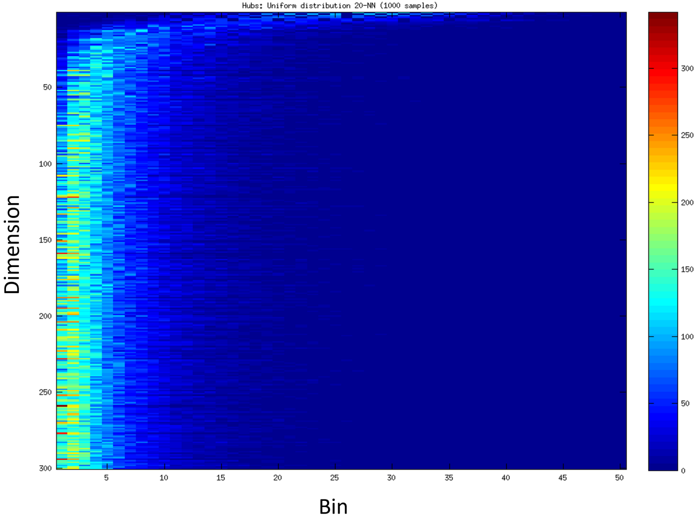
Data hubness (empirical)

20-NN with $D = 100$ and 1000 Uniform samples over 50 bins



Data hubness (empirical)

20-NN with $D = 100$ and 1000 Uniform samples over 50 bins



Concentration inequalities

- Markov's inequality tells us that for a non-negative r.v. X

$$\mathbb{P}(X > \varepsilon) \leq \frac{\mathbb{E}\,X}{\varepsilon}$$

(easily proven via the CDF of X)

- Chebyshev's inequality refines for a r.v of finite expected value μ and finite non-zero variance σ^2 and $c > 0$

$$\mathbb{P}(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

- From there Hoeffding's inequality [6] focuses on collections of independent variables. If $\{X_i\}_{i \in [\![D]\!]}$ are independent r.v with $\mathbb{P}(a \leq X_i \leq b) = 1$ and $\mathbb{E}\,X_i = \mu$, then for any $t > 0$

$$\mathbb{P}(|\overline{X} - \mu| > \varepsilon) \leq 2e^{-2D\varepsilon^2/(b-a)^2} \quad \text{where} \quad \overline{X} = \frac{1}{D}\sum_{i=1}^{D} X_i$$

Note: Example of a Probably Approximately Correct (PAC) analysis: probability of an event approximately occuring

Concentration inequalities

$$\mathbb{P}(|\overline{X} - \mu| > \varepsilon) \leq 2e^{-2D\varepsilon^2/(b-a)^2} \quad \text{where} \quad \overline{X} = \frac{1}{D}\sum_{i=1}^{D} X_i$$

- ⋆ Hoeffding's inequality tells us that when aggregating independent variables, the aggregate converges to their common expectation
- ⋆ The structure $\left(\sum_{i=1}^{D} \cdot\right)$ exactly corresponds to what is done for computing Minkowsky distances (inc Euclidean distance) $d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{D}(\boldsymbol{x}(i) - \boldsymbol{y}(i))^p\right)^{1/p}$. Hoeffding's inequality formally explains why distance values concentrate
- ⋆ Although it is counter-intuitive that all neighbors are almost equidistant, it is rather intuitive that for high values of D small and large values of $d_p(\boldsymbol{x}, \boldsymbol{y})$ become unlikely: for $d_p(\boldsymbol{x}, \boldsymbol{y})$ to be small (large), all D coordinates $\boldsymbol{x}(i)$ and $\boldsymbol{y}(i)$ should be (dis)similar. When D augments, this becomes unlikely

Weak Law of Large Numbers

Reminder: Given a set of D iid r.v $\{X_i\}_{i \in [\![D]\!]}$ with $\mathbb{E}\, X_i = \mu$ then for any $\varepsilon > 0$

$$\lim_{D \to \infty} \mathbb{P}(|\overline{X} - \mu| < \varepsilon) = 1 \quad \text{where} \quad \overline{X} = \frac{1}{D} \sum_{i=1}^{D} X_i$$

(can be proven from Chebyshev's inequality)

- ⋆ The WLLN tells us that the sample average $\overline{X}$ converges (in probability) towards the rue expectation μ as the number of samples augments
- ⋆ Fundamentally consistent with the frequentist approach for probabilities
- ⋆ It is the fundamental basis for estimation theory (and density estimation)
- ⋆ The shape and rate of convergence is given by the Central Limit Theorem

$$Z_D = \frac{\sqrt{D}}{\sigma}(\overline{X} - \mu) \sim \mathcal{N}(0, 1) \quad \text{if } \mathrm{Var}(X_i) = \sigma^2$$

Random projections

- ⋆ Reminder: two random vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are likely to be orthogonal
- ⋆ Quasi-orthogonality: $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \varepsilon$ (small). Exponentially many quasi-orthogonal directions in high dimensional spaces
- ⋆ Random projections exploit that fact to create random (quasi-orthonormal) basis for the decomposition and coding of signal
- ⋆ One appropriate application is locally sensitive hashing (LSH) for Approximate Nearest Neighbor (ANN) search [7, 9]
- ⋆ Caveat: rate of convergence slower than most concentrations

Lemma (Johnson-Lindenstrauss)

Given $\mathcal{X} \subset \Omega$ there exists a linear mapping $\boldsymbol{p} : \mathbb{R}^D \to \mathbb{R}^d$ such that with $0 < \varepsilon < 1$ and for all $i, j \in [\![N]\!]$

$$(1-\varepsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq \|\boldsymbol{p}(\boldsymbol{x}_i) - \boldsymbol{p}(\boldsymbol{x}_j)\| \leq (1+\varepsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$$

and $d = O\left(\frac{\log N}{\varepsilon^2}\right)$ ← *Does not depend on D !*

Do we care?

Sampling

- ⋆ Data sparsity tell us that we need an exponential number of samples to reach every position of a high-dimensional space
- ⋆ Concentration inequalities (inc WLLN and CLT) are indicators of the precision for density estimation
- ⋆ In practical processes like rejection sampling, rejection rate gets exponentially close to 100% (ref empty sphere)

Do we care?

Indexing

- ⋆ Indexing exploits the structure of the data to anticipate the parts of the data space Ω to visit at query time
- ⋆ If all neighbors are seemingly at equal distance from the center, the structure disappears
- ⋆ More formally, space partitioning techniques (e.g *kd*-trees) attempt to obtain a space partition so that every data is uniquely identified within an intersection of cells (paths in the search tree). With concentration of distances, the size of the cell in the space partition reaches the radius of the sphere to cover the complete space [11]

⇒ The search amounts to an exhaustive search ⇒ complexity is $O(N)$

+ Hubness makes some data an answer to almost every query

Do we care?

Learning

- ⋆ Machine learning is in large part about discovering function approximators [8]
- ⋆ Given $\phi : \mathcal{X} \rightarrow \mathcal{Y}$ the (true) map from data $\mathcal{X}$ to labels $\mathcal{Y}$, ML seeks an approximation $\phi_{\boldsymbol{\theta}} \in \mathbb{F}$ of ϕ minimizing an error (risk) over (training/test) samples
- ⋆ If $\mathbb{F}$ is a class of c-uniform Lipschitz functions over Ω then

$$\sup_{\phi_{\boldsymbol{\theta}} \in \mathbb{F}} \|\phi_{\boldsymbol{\theta}} - \phi\|_\infty \geq c \frac{\sqrt{D} N^{-1/D}}{2} \sqrt{\frac{2}{\pi e}} \left(1 + O\left(\frac{\log D}{D} \right) \right)$$

⇒ If we allow an error $c\varepsilon$ for $\varepsilon > 0$, the number of required training examples N is

$$N \geq \frac{\varepsilon^{-D} D^{D/2}}{(2\pi e)^{D/2}}$$

Do we care?

Learning

$\Rightarrow$ If we allow an error $c\varepsilon$ for $\varepsilon > 0$, the number of required training examples N is

$$N \geq \frac{\varepsilon^{-D} D^{D/2}}{(2\pi e)^{D/2}}$$

Examples:

- $\star$ $c = 1$, $\varepsilon = 0.1$, $D = 3 \Rightarrow N \geq 74$
- $\star$ $c = 1$, $\varepsilon = 0.1$, $D = 10 \Rightarrow N \geq 687 \cdot 10^6$
- $\star$ $c = 1$, $\varepsilon = 0.1$, $D = 15 \Rightarrow N \geq 377 \cdot 10^{12}$
- $\star$ $c = 1$, $\varepsilon = 0.01$, $D = 10 \Rightarrow N \geq 6.9 \cdot 10^{18}$

ML data has generically hundreds of features ($D \sim O(10^2)$) and we seek a typical error $\varepsilon \simeq 10^{-4} \Rightarrow N \geq \frac{10^{4\cdot 100} 100^{50}}{17.07^{50}} \simeq 10^{400}$

Forget about high-dimensional data science?

Q: Why do data science process work?

- ⋆ Most of the above analysis assumes some form of uniform (or isotropic – Normal, ...) distribution of the data
  However, real data does not populate the feature space uniformly (if the features are informative)
- ⋆ Most of the above analysis assumes independent features
  However, features show correlations

⇒ In practice the data populates subspaces of Ω of lower (local intrinsic) dimensionality

⇒ Interest in:

- ◦ Decorrelating the features (e.g via linear latent space)
- ◦ Discovering data subspaces (clustering, non-linear dimension reduction)

Summary

Geometry in high dimensional spaces ($D \simeq 10$) is different from the 3D geometry we are used to

- ⊕ Concentration inequalities are at the base for statistical estimation

- ⋆ Data becomes very sparse unless to use exponentially many data
- ⋆ Even with dense data, distance and angle values concentrate
  - ⇒ From a query, all neighbors are roughly equidistant
  - ⇒ From a direction all the rest looks orthogonal
- ⋆ The density concentrates over tight radii ⇒ sampling becomes biased
- ⋆ Hubs make some data part of all classes/query answer

⇒ Since this is by construction, the only solution is to somehow operate in low-dimensional spaces

Example questions [mostly require formal – mathematical – answers]

- What is the dimension?
- Cite a concrete instance of the Curse of Dimensionality
- Cite a concrete instance of the Blessing of Dimensionality
- Why does dimension increase data sparsity?
- Why volume concentrates on the surface of volumes?
- What is a concentration phenomenon? How can it be beneficial?
- Why do angles concentrate? How can we exploit this?
- What is the typical rate of concentration?
- Cite one concentration inequality. Can you prove it?
- Detail one concrete situation where concentration is beneficial
- Detail one concrete situation where concentration is adverse
- What is a PAC analysis?

It is strongly adviced to run the simulations contained in this chapter

References I

[1] Fabrizio Angiulli. On the behavior of intrinsically high-dimensional spaces: Distances, direct and reverse nearest neighbors, and hubness. *Journal of Machine Learning Research (JMLR)*, 18:1–60, 2018.

[2] Keith M. Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31, pages 1–58. Mathematical Sciences Research Institute, 1997.

[3] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Proceedings of the 7th International Conference on Database Theory*, ICDT '99, pages 217–235, London, UK, UK, 1999. Springer-Verlag.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. (available online).

[5] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020. (available online).

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

References II

[7] Yihe Dong, Piotr Indyk, Ilya P. Razenshteyn, and Tal Wagner. Learning space partitions for nearest neighbor search. *CoRR*, abs/1901.08544, 2019.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[9] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proc. of ACM Symposium on Theory of computing (STOC '98)*, pages 604–613. ACM, 1998.

[10] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, December 2010.

[11] Uri Shaft and Raghu Ramakrishnan. Theory of nearest neighbors indexability. *ACM Transactions on Database Systems*, 31(3):814–838, September 2006.

[12] Ramon van Handel. Probability in high dimension. APC 550 Lecture Notes – Princeton University (available online), 2016.

[13] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. (available online).

References III

[14] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.