Machine Learning

Talking sense about data

Emanuele Rodolà
rodola@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

2nd semester a.y. 2023/2024 · March 4, 2024

Data awareness

Machine learning involves dealing with data.

What do you do when you have a problem involving data?

Data awareness

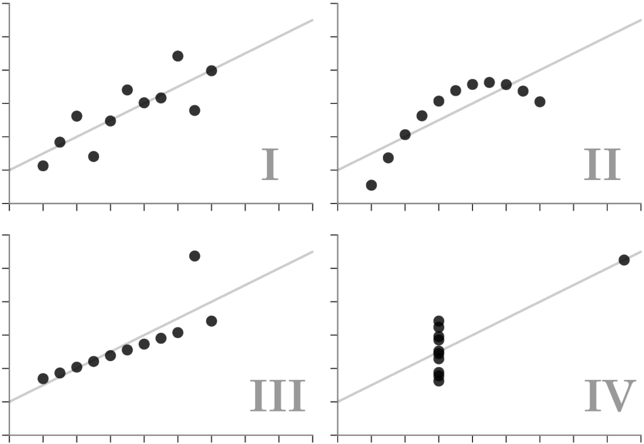
Machine learning involves dealing with data.

What do you do when you have a problem involving data?

First thing: **look at the data!**

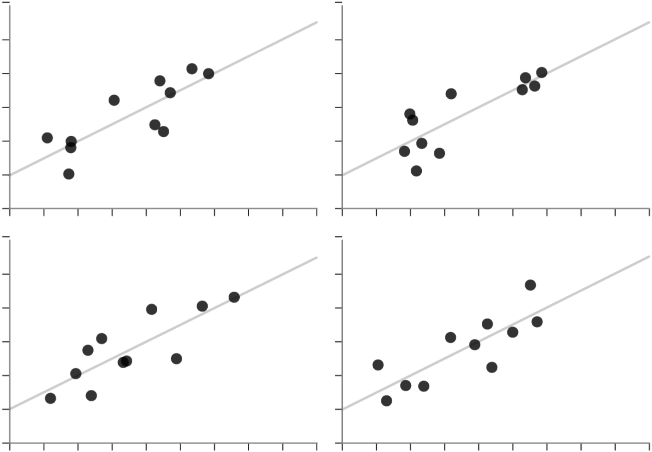
Anscombe’s Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



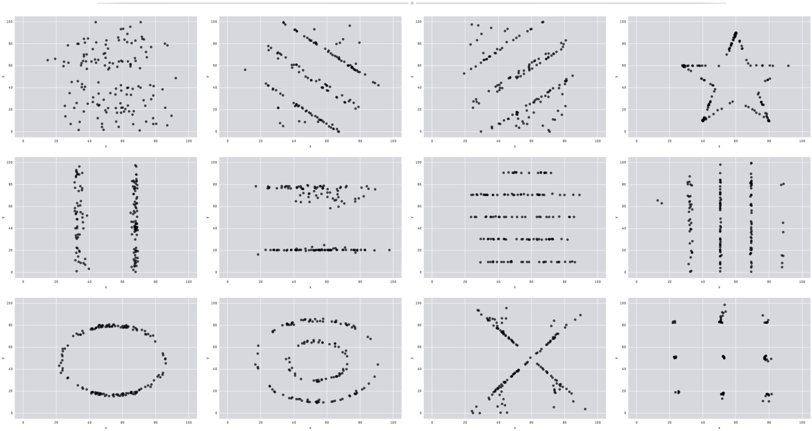
✗ Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



The datasaurus dozen

All these datasets have the same summary stats to 2 decimal places:



Matejka and Fitzmaurice, SIGCHI 2017

The datasaurus dozen

All these datasets have the same summary stats to 2 decimal places:

Matejka and Fitzmaurice, SIGCHI 2017

Data awareness

Machine learning involves dealing with data.

What do you do when you have a problem involving data?

First thing: **look at the data!**

Never trust summary statistics alone;
when possible, visualize your data

Data awareness

Machine learning involves dealing with data.

What do you do when you have a problem involving data?

First thing: **look at the data!**

Never trust summary statistics alone;
when possible, visualize your data

It will not always be easy to visualize.

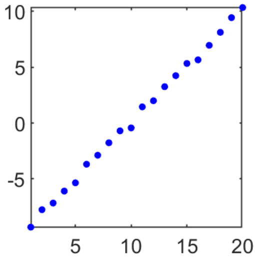
Difficult cases: high-dimensional data, no physical access to data, implicit access to data (e.g. latent spaces).

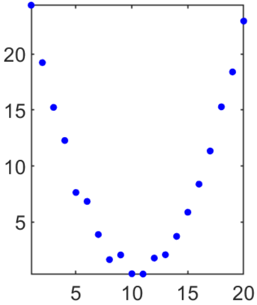
Models for describing the data

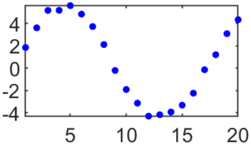
Learning is about describing data, or more specifically, describing the process, or model, that yields a given output from a given input.

Models for describing the data

Learning is about describing data, or more specifically, describing the process, or model, that yields a given output from a given input.

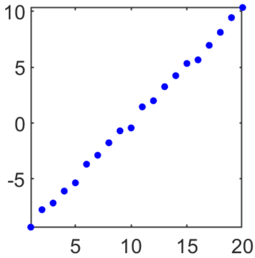




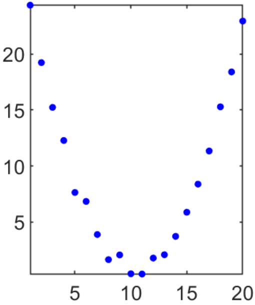


Models for describing the data

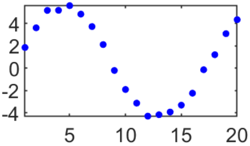
Learning is about describing data, or more specifically, describing the process, or model, that yields a given output from a given input.



$$y = ax + b$$



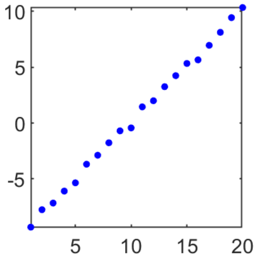
$$y = ax^2 + bx + c$$

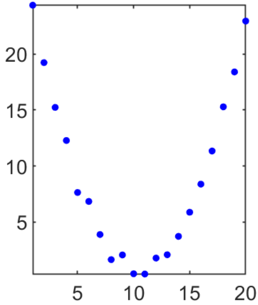


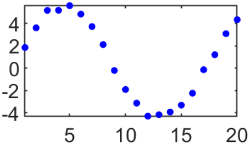
$$y = ax^3 + bx^2 + cx + d$$

Models for describing the data

Learning is about describing data, or more specifically, describing the process, or model, that yields a given output from a given input.







$y = ax + b$ $\quad$ $y = ax^2 + bx + c$ $\quad$ $y = a\sin(x) + bx + c$

Our model might use prior knowledge on the data.

For example, in the third plot, we might know *a priori* that the data actually comes from a periodic process.

Main idea: Look at the world, identify what knowledge we have about it, and use this knowledge to construct our model.

Modeling prior knowledge

Main idea: Look at the world, identify what knowledge we have about it, and use this knowledge to construct our model.

Some forms of prior knowledge:

- Data distribution

Main idea: Look at the world, identify what knowledge we have about it, and use this knowledge to construct our model.

Some forms of prior knowledge:

- Data distribution
- Energy function

Modeling prior knowledge

Main idea: Look at the world, identify what knowledge we have about it, and use this knowledge to construct our model.

Some forms of prior knowledge:

- Data distribution
- Energy function
- Constraints

Modeling prior knowledge

Main idea: Look at the world, identify what knowledge we have about it, and use this knowledge to construct our model.

Some forms of prior knowledge:

- Data distribution
- Energy function
- Constraints
- Invariances

Modeling prior knowledge

Main idea: Look at the world, identify what knowledge we have about it, and use this knowledge to construct our model.

Some forms of prior knowledge:

- Data distribution
- Energy function
- Constraints
- Invariances
- Input-output examples (data prior)

Modeling prior knowledge

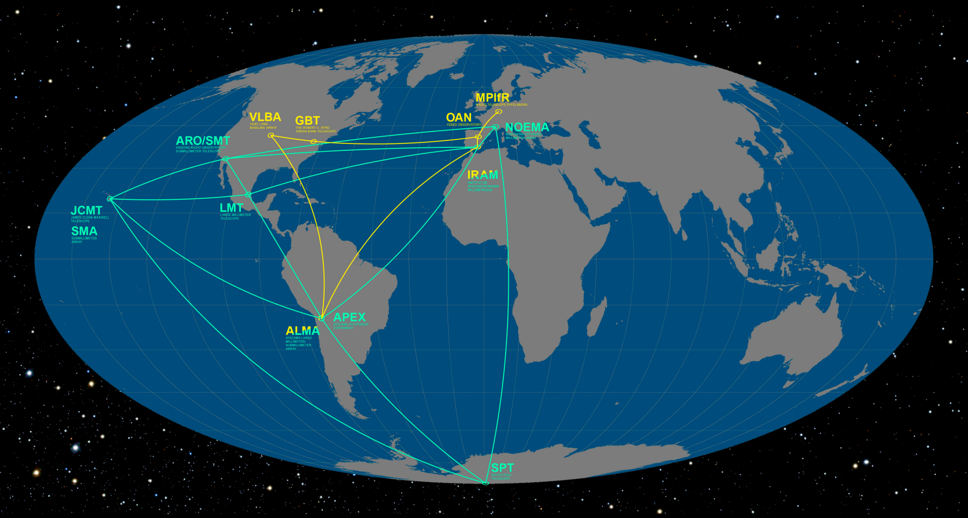
Main idea: Look at the world, identify what knowledge we have about it, and use this knowledge to construct our model.

Some forms of prior knowledge:

- Data distribution
- Energy function
- Constraints
- Invariances
- Input-output examples (data prior)

All these encode, to different extents, some expected behavior.

Event Horizon Telescope



Reliability of the prior: imaging the black hole

Problem: reconstruct an image from a sparse set of spectral measurements (VLBI imaging).

Bouman et al, "Computational imaging for VLBI image reconstruction", CVPR 2016

Reliability of the prior: imaging the black hole

Problem: reconstruct an image from a sparse set of spectral measurements (VLBI imaging).

It is an ill-posed inverse problem:

- Infinite number of possible images explain the data

Bouman et al, "Computational imaging for VLBI image reconstruction", CVPR 2016

Reliability of the prior: imaging the black hole

Problem: reconstruct an image from a sparse set of spectral measurements (VLBI imaging).

It is an ill-posed inverse problem:

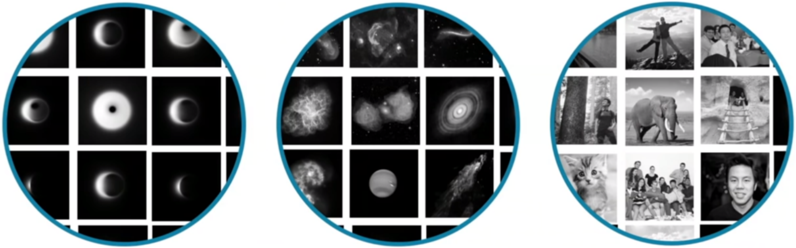
- Infinite number of possible images explain the data
- Optimization heavily relies on priors.
  Find an explanation that respects prior assumptions about the "visual" universe while still satisfying the observed data.

Bouman et al, "Computational imaging for VLBI image reconstruction", CVPR 2016

Reliability of the prior: imaging the black hole

Problem: reconstruct an image from a sparse set of spectral measurements (VLBI imaging).

Which of these datasets would you use?



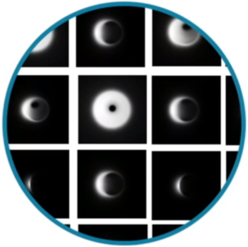
black holes astronomy everyday

Bouman et al, "Computational imaging for VLBI image reconstruction", CVPR 2016

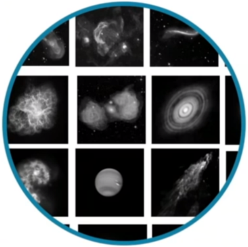
Reliability of the prior: imaging the black hole

Problem: reconstruct an image from a sparse set of spectral measurements (VLBI imaging).

Which of these datasets would you use?



black holes
unreliable



astronomy

everyday

Black holes are dangerous! They will yield what one expects to obtain.

Bouman et al, "Computational imaging for VLBI image reconstruction", CVPR 2016

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

Bias in the training dataset is still an open research problem!

Some possible causes:

- **Skewed sample**: a tiny initial bias grows over time, since future observations confirm prediction.

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

Bias in the training dataset is still an open research problem!

Some possible causes:

- **Skewed sample**: a tiny initial bias grows over time, since future observations confirm prediction. Example: Police intercept crime more densely in areas they watch.

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

Bias in the training dataset is still an open research problem!

Some possible causes:

- **Skewed sample**: a tiny initial bias grows over time, since future observations confirm prediction. Example: Police intercept crime more densely in areas they watch.
- **Tainted examples**: data produced by a human decision can be biased, and the bias is replicated by the system.

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

Bias in the training dataset is still an open research problem!

Some possible causes:

- **Skewed sample**: a tiny initial bias grows over time, since future observations confirm prediction. Example: Police intercept crime more densely in areas they watch.
- **Tainted examples**: data produced by a human decision can be biased, and the bias is replicated by the system.
- **Sample size disparity**: training data for a minority group is much less than the majority group.

Reliability of the prior: fairness

AI is objective only in the sense of learning what humans teach.

The data provided by human can be highly biased.

Bias in the training dataset is still an open research problem!

Some possible causes:

- **Skewed sample**: a tiny initial bias grows over time, since future observations confirm prediction. Example: Police intercept crime more densely in areas they watch.
- **Tainted examples**: data produced by a human decision can be biased, and the bias is replicated by the system.
- **Sample size disparity**: training data for a minority group is much less than the majority group.

Assessing data and prior reliability is crucial for any learning-based system.

Explaining the data

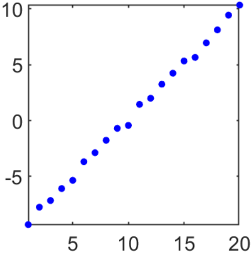
Learning is about discovering a **map** from input to output.

"Finding a model explaining the data" means determining the map.

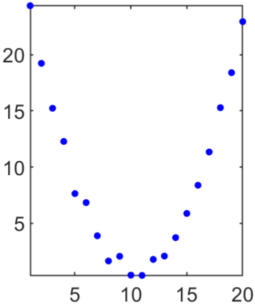
Explaining the data

Learning is about discovering a **map** from **input** to **output**.

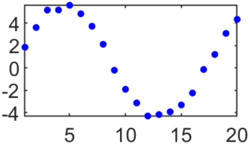
"Finding a model explaining the data" means determining the map.



$y = ax + b$



$y = ax^2 + bx + c$



$y = a\sin(x) + bx + c$

Explaining the data

Learning is about discovering a **map** from input to output.

"Finding a model explaining the data" means determining the map.

Key assumption: the data has an underlying structure.

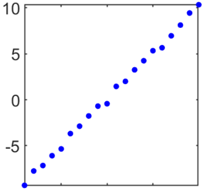
Explaining the data

Learning is about discovering a **map** from input to output.

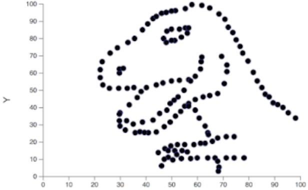
"Finding a model explaining the data" means determining the map.

Key assumption: the data has an underlying structure.

This structure is almost never captured by a simple expression.



$y = ax + b$



$?$

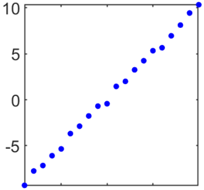
Explaining the data

Learning is about discovering a **map** from input to output.

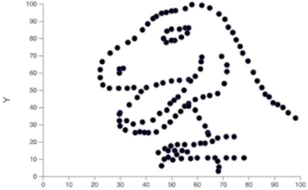
"Finding a model explaining the data" means determining the map.

Key assumption: the data has an underlying structure.

This structure is almost never captured by a simple expression.



$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_x \\ a_y \end{pmatrix} t + \begin{pmatrix} b_x \\ b_y \end{pmatrix}$$



(not a function in 1D)

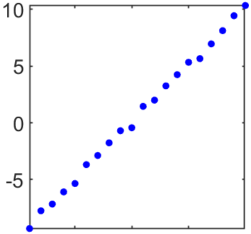
Clearly, data is not always one-dimensional.

Choosing a representation

The same data can be described in different ways.

What is the "right" way?

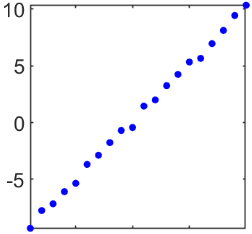
as a function



$$y = ax + b$$

$=$

as a parametric curve



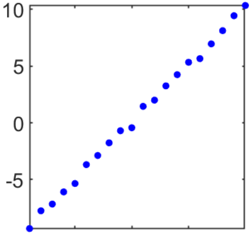
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_x \\ a_y \end{pmatrix} t + \begin{pmatrix} b_x \\ b_y \end{pmatrix}$$

Choosing a representation

The same data can be described in different ways.

What is the "right" way?

as a function

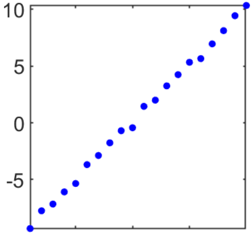


$$y = a x + b$$

2 weights

$=$

as a parametric curve



$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_x \\ a_y \end{pmatrix} t + \begin{pmatrix} b_x \\ b_y \end{pmatrix}$$

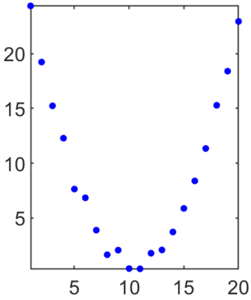
4 weights

Choosing a representation

The same data can be described in different ways.

What is the “right” way?

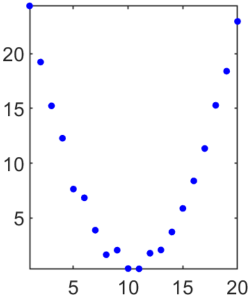
as a **function**



$$y = ax^2 + bx + c$$

=

as a **parametric curve**



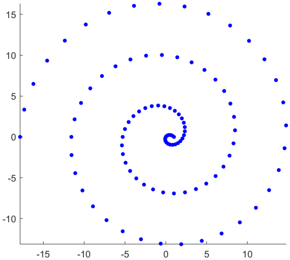
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_x \\ a_y \end{pmatrix} t^2 + \begin{pmatrix} b_x \\ b_y \end{pmatrix} t + \begin{pmatrix} c_x \\ c_y \end{pmatrix}$$

Choosing a representation

The same data can be described in different ways.

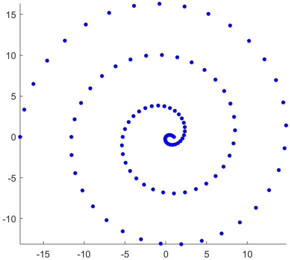
What is the “right” way?

as a **function**



$=$

as a **parametric curve**

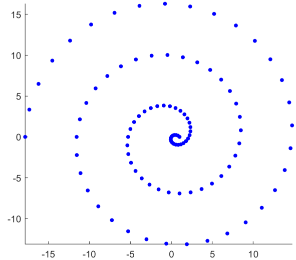


Choosing a representation

The same data can be described in different ways.

What is the "right" way?

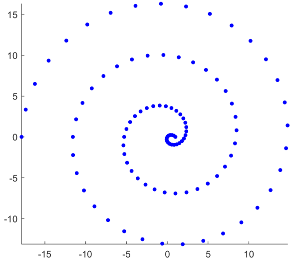
as a **function**



y is not a function of x

$=$

as a **parametric curve**



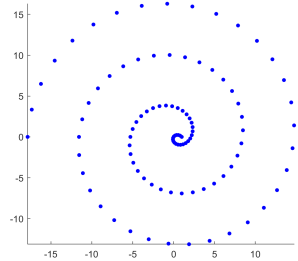
$$\begin{pmatrix} x \\ y \end{pmatrix} = a \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} (a - t)$$

Choosing a representation

The same data can be described in different ways.

What is the “right” way?

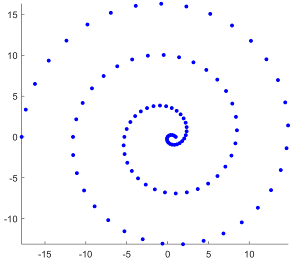
as a function



$r = a\theta$
(polar coordinates)

$=$

as a parametric curve



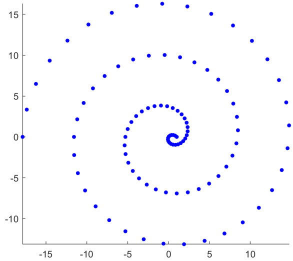
$$\begin{pmatrix} x \\ y \end{pmatrix} = a \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} (a - t)$$

Choosing a representation

The same data can be described in different ways.

What is the “right” way?

as a **function**

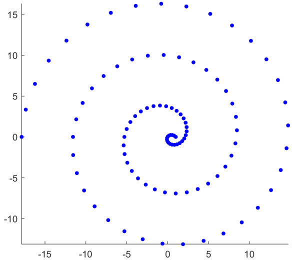


$r = a\theta$

linear!

$$=$$

as a **parametric curve**



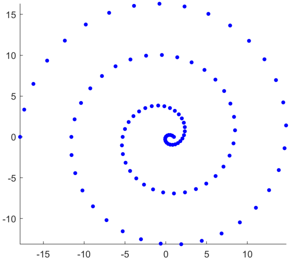
$$\begin{pmatrix} x \\ y \end{pmatrix} = a \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} (a - t)$$

Choosing a representation

The same data can be described in different ways.

What is the "right" way?

as a **function**

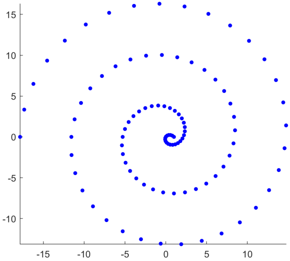


$r = a\theta$

linear!

$=$

as a **parametric curve**



$$\begin{pmatrix} x \\ y \end{pmatrix} = a \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} (a - t)$$

Trade-off between #weights and simplicity

The curse of dimensionality

Of course, data can have more than 1 or 2 dimensions.

The curse of dimensionality

Of course, data can have more than 1 or 2 dimensions.

For example, a $w \times h$ image has wh dimensions.

$\in \mathbb{R}^{w \times h} \cong \mathbb{R}^{wh}$

Example: $\sim$1 megapixel photo (grayscale) has $\sim 10^6$ dimensions.

The curse of dimensionality

Of course, data can have more than 1 or 2 dimensions.

For example, a $w \times h$ image has wh dimensions.

The value of each coordinate is given by the gray value at that pixel.
Then, the entire image is one point in a wh-dimensional space.

$\in \mathbb{R}^{w \times h} \cong \mathbb{R}^{wh}$

Example: ~1 megapixel photo (grayscale) has $\sim 10^6$ dimensions.

Are all those dimensions significant?

The curse of dimensionality

For simplicity, consider 1×1 images, i.e., consisting of one single pixel.

Each image is a point in one dimension.

The curse of dimensionality

For simplicity, consider 1×1 images, i.e., consisting of one single pixel.

Each image is a point in one dimension.

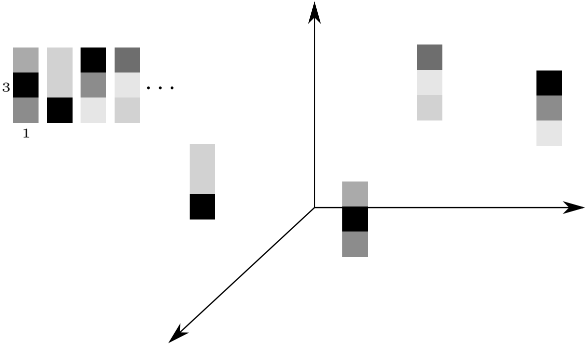
Similarly, with 2 pixels we get:

The curse of dimensionality

For simplicity, consider 1×1 images, i.e., consisting of one single pixel.

Each image is a point in one dimension.

Similarly, with 3 pixels we get:

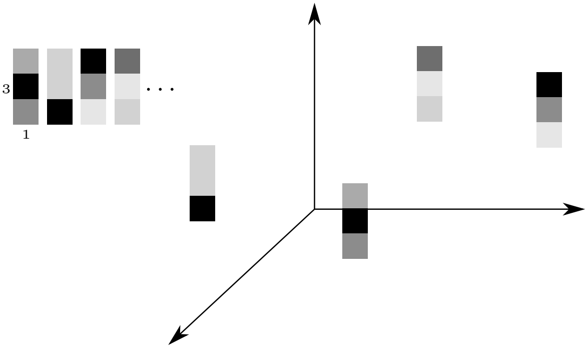


The curse of dimensionality

For simplicity, consider 1×1 images, i.e., consisting of one single pixel.

Each image is a point in one dimension.

Similarly, with 3 pixels we get:



Each new dimension increases sparsity of the point cloud.

The curse of dimensionality

A dataset of natural images will be extremely sparse in $\mathbb{R}^{w \times h}$, since each region of space is observed very infrequently.

New samples are less likely to fall close to the previous ones.

The curse of dimensionality

A dataset of natural images will be extremely sparse in $\mathbb{R}^{w \times h}$, since each region of space is observed very infrequently.

New samples are less likely to fall close to the previous ones.

To discover a pattern, we need exponentially many observations as we have dimensions!

The curse of dimensionality

A dataset of natural images will be extremely sparse in $\mathbb{R}^{w \times h}$, since each region of space is observed very infrequently.

New samples are less likely to fall close to the previous ones.

To discover a pattern, we need exponentially many observations as we have dimensions!

If n data points cover well the space of 1-dimensional images, then n^d data points are required for d-dimensional images.

More data points make interesting structures emerge

The curse of dimensionality

A dataset of natural images will be extremely sparse in $\mathbb{R}^{w \times h}$, since each region of space is observed very infrequently.

New samples are less likely to fall close to the previous ones.

To cover the space entirely, we need exponentially many observations as we have dimensions!

If n data points cover well the space of 1-dimensional images, then n^d data points are required for d-dimensional images.

Two options:

1. **Increase** the dataset
2. **Decrease** the dimensions

Favor simplicity

Let's play a game:

$$2, 4, 8, \ldots$$

Rules:

- **Task:** Discover the rule I used to produce the sequence
- Give me a number: I'll tell you if it's next in sequence or not
- **Once you're sure**, tell me the rule

Favor simplicity

Let's play a game:

$$2, 4, 8, \ldots$$

Rules:

- **Task:** Discover the rule I used to produce the sequence
- Give me a number: I'll tell you if it's next in sequence or not
- **Once you're sure**, tell me the rule

Occam's razor: Among competing hypotheses, select the one with the fewest assumptions.

Favor simplicity

Let's play a game:

$$2, 4, 8, \ldots$$

Rules:

- **Task:** Discover the rule I used to produce the sequence
- Give me a number: I'll tell you if it's next in sequence or not
- **Once you're sure**, tell me the rule

Occam's razor: Among competing hypotheses, select the one with the fewest assumptions.

Also: when feasible, add more data!

Suggested reading

Blog post on the datasaurus:
https://www.autodeskresearch.com/publications/samestats

TED talk on the idea behind imaging the black hole:
https://www.youtube.com/watch?v=BIvezCVcsYs

VLBI reconstruction dataset:
http://vlbiimaging.csail.mit.edu/

Paper on the black hole imaging technique:
https://arxiv.org/pdf/1512.01413.pdf

Tutorial video and slides on ML fairness:
https://nips.cc/Conferences/2017/Schedule?showEvent=8734