Background of Representation Learning

Ke Chen

Department of Computer Science, The University of Manchester

Ke.Chen@manchester.ac.uk

Outline

Overview

- Importance of data representation
- Representation acquisition methodology

Curse of Dimensionality

- Illustrative example and implication
- Surprising fact on high-dimensional space

Manifold and Latent Factor

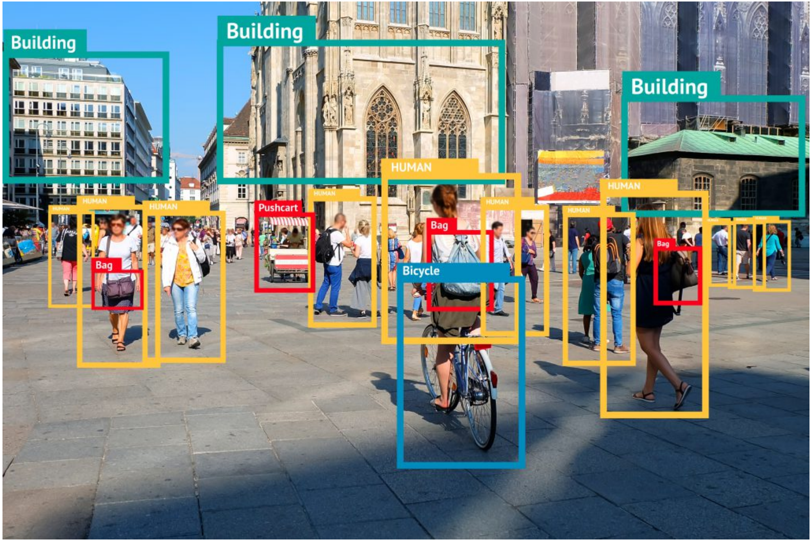
- Hidden “meaningful” low-dimensional structural space
- Hidden factors that explain/decide data generating distribution

General Aspect

- Supervised versus unsupervised representation learning
- Dimension reduction

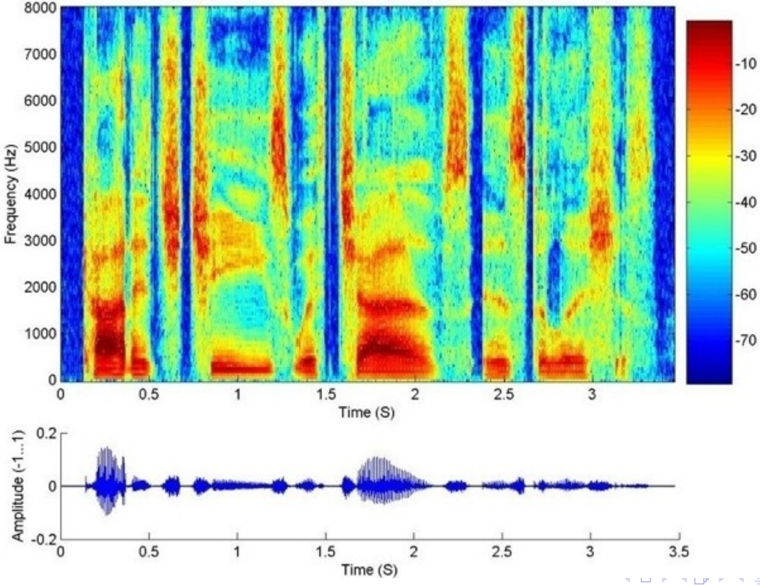
- **Importance of Data Representation**
  - Data are simply phenomena on surface and information conveyers.
  - Data representation or features provide an effective way to encode the information underlying data.
  - According to neuroscience/cognitive science, human intelligence acquisition is a hierarchical information processing process.
  - In the hierarchical process, the information at different levels is extracted and encoded in different representational forms.
  - Effective data representation is essential in intelligent system development and attributed to the success of machine learning.

- **Importance of Data Representation: Computer Vision**

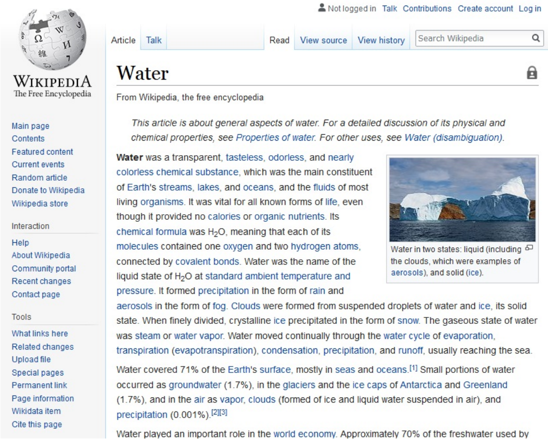


Overview

- **Importance of Data Representation: Speech Recognition**



- **Importance of Data Representation: Natural Language Processing**



Not logged in Talk Contributions Create account Log in

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Article Talk Read View source View history Search Wikipedia

Water

From Wikipedia, the free encyclopedia

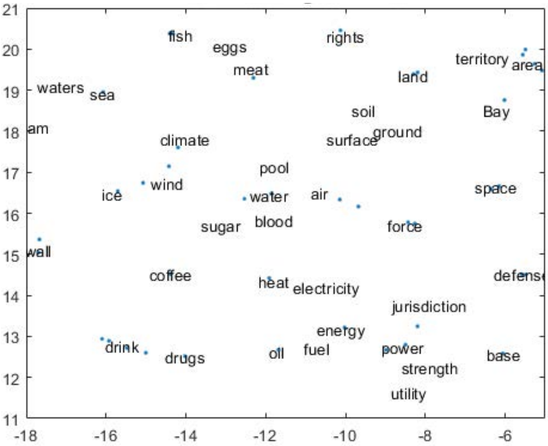
This article is about general aspects of water. For a detailed discussion of its physical and chemical properties, see Properties of water. For other uses, see Water (disambiguation).

Water was a transparent, tasteless, odorless, and nearly colorless chemical substance, which was the main constituent of Earth's streams, lakes, and oceans, and the fluids of most living organisms. It was vital for all known forms of life, even though it provided no calories or organic nutrients. Its chemical formula was H_2O, meaning that each of its molecules contained one oxygen and two hydrogen atoms, connected by covalent bonds. Water was the name of the liquid state of H_2O at standard ambient temperature and pressure. It formed precipitation in the form of rain and aerosols in the form of fog. Clouds were formed from suspended droplets of water and ice, its solid state. When finely divided, crystalline ice precipitated in the form of snow. The gaseous state of water was steam or water vapor. Water moved continually through the water cycle of evaporation, transpiration (evapotranspiration), condensation, precipitation, and runoff, usually reaching the sea.

Water in two states: liquid (including the clouds, which were examples of aerosols), and solid (ice).

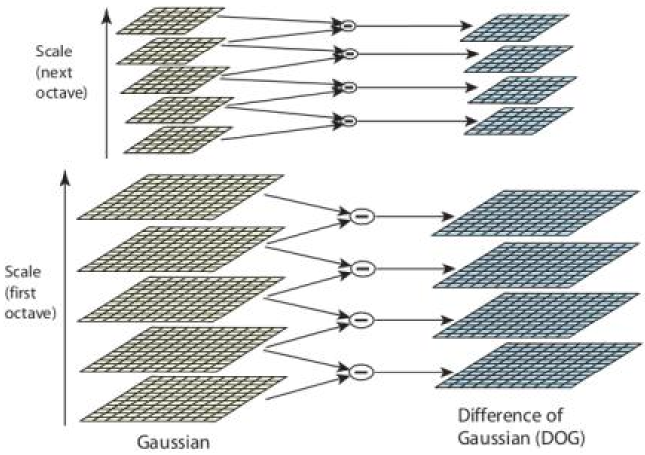
Water covered 71% of the Earth's surface, mostly in seas and oceans.[1] Small portions of water occurred as groundwater (1.7%), in the glaciers and the ice caps of Antarctica and Greenland (1.7%), and in the air as vapor, clouds (formed of ice and liquid water suspended in air), and precipitation (0.001%).[2][3]

Water played an important role in the world economy. Approximately 70% of the freshwater used by

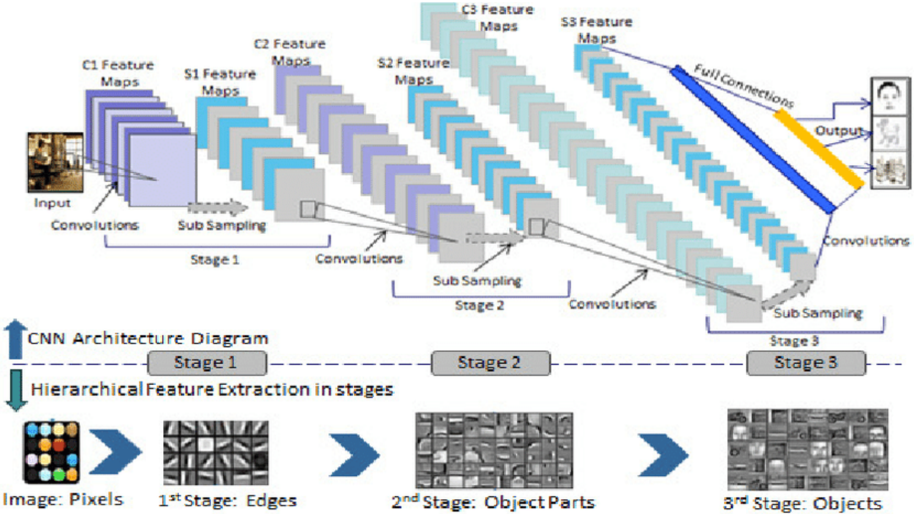


- **Representation Acquisition Methodology: Handcrafted Feature Engineering**

Scale-Invariant Feature Transformation (SIFT)

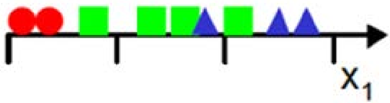


- **Representation Acquisition Methodology:** **Representation Learned from Data**



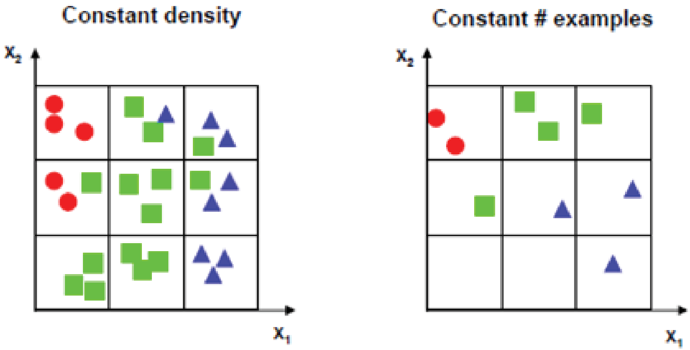
- **Introduction**
  - Curse of dimensionality coined by Richard E. Bellman when considering problems in dynamic programming
  - Various phenomena arising from analysing/organising data in high-dimensional spaces rather than low-dimensional settings, e.g., 3-D real-world space of our everyday experience
  - When increasing the dimensionality, its volume of the space increasing so fast that the available data become sparse
  - Such a cursed phenomenon is a big challenge in many domains requiring statistical significance, e.g., machine learning and statistics.

- **Illustrative Example: 3-class Classification**



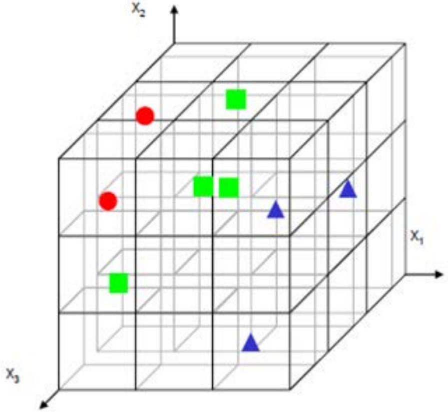
  - 3 class labels indicated by circle, square and triangle
  - A naive grid-based classification method as follows:
    - divide the feature space along each axis into uniform bins
    - calculate the ratio of examples for each class at each bin
    - for a test instance, find its bin/assign the label of the dominated class in the bin
  - For 1-D case: $3^1 = 3$ bins, $3 \times 3^1 = 9$ training examples
  - Due to a single feature used, the performance may be low due to too much overlapping among 3 classes
  - A straightforward idea: incorporate one more feature in order to improve the separability

- **Illustrative Example: 3-class Classification**



- In 2-D case, preserving the granularity of each axis leads to $3^2 = 9$ bins.
- Two choices for training examples in 2-D case:
  1. maintain the density of examples/bin, $3 \times 3^2 = 27$ examples
  2. keep the number of examples in 1-D case, only 9 examples but very sparse

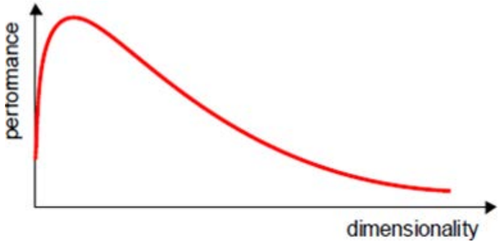
- **Illustrative Example: 3-class Classification**



In 3-D, the problem is exacerbated!

- The number of bins grows to $3^3 = 27$
- maintain the density of examples/bin, $3 \times 3^3 = 81$
- keep the number of examples in 1-D case, only 9 examples but too sparse to work

- **Generic Case**
  - Exponential growth in number of examples required to maintain a given sampling density; for a density of k examples/bin and d dimensions where each dimension is divided into n bins, the total number of examples needed is kn^d.
  - In our illustrative example, we have $k = 3$ and $n = 3$, so total $3 \times 3^d = 3^{d+1}$ examples are required to maintain the same sampling density.
- **Implication**
  For a given dataset, there is a maximum number of features above which the performance of a learning system is not improved but degraded in practice.

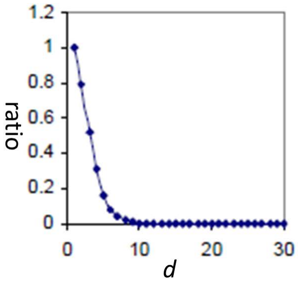


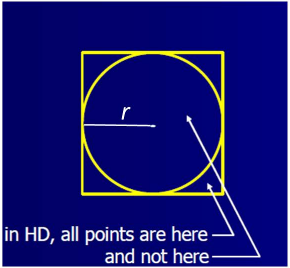
- **Surprising Fact on High-dimensional Space**

For hyper-cube (sc) and its embedded hyper-sphere (sc), observe the ratio between their volumes, $V_{hs}(d)$ and $V_{hc}(d)$, in different dimensions, d.

$$\text{ratio}(2) = \frac{V_{hs}(2)}{V_{hc}(2)} = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4}; \quad \text{ratio}(3) = \frac{V_{hs}(3)}{V_{hc}(3)} = \frac{\frac{4}{3}\pi r^3}{(2r)^3} = \frac{\pi}{6}.$$

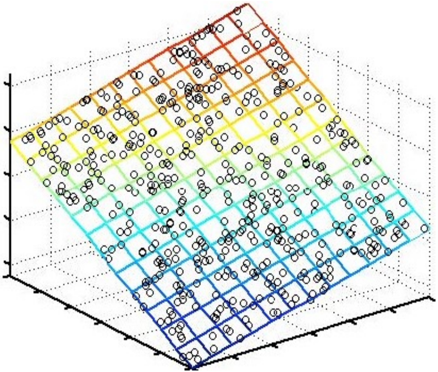
$$\text{ratio}(d) = \frac{V_{hs}(d)}{V_{hc}(d)} = \frac{\frac{2r^d\pi^{d/2}}{d\Gamma(d/2)}}{(2r)^d} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)}, \quad \text{for } d > 3.$$

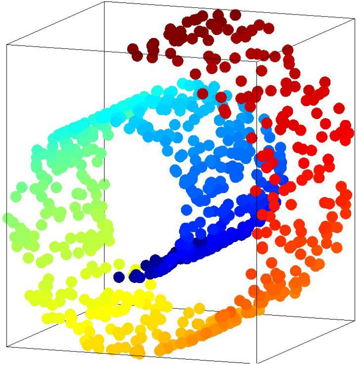




- **Manifold**

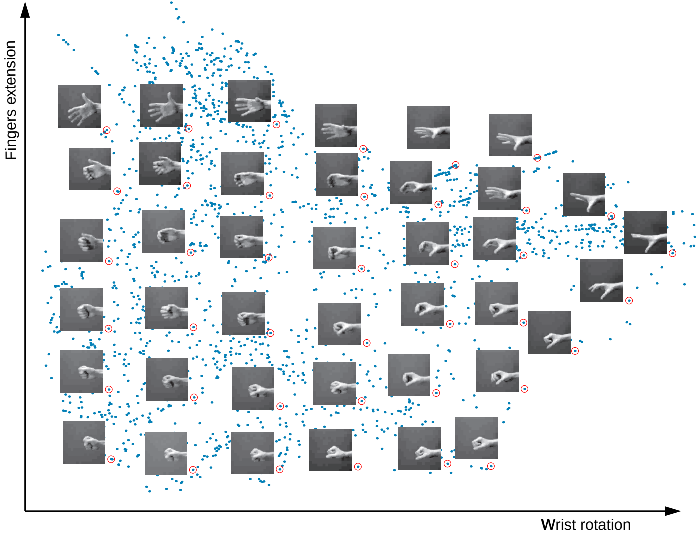
  "Meaningful" low-dimensional topological structure of data in high-dimensional space





- **Latent Factor**

  Few hidden factors able to explain/decide data generating distribution

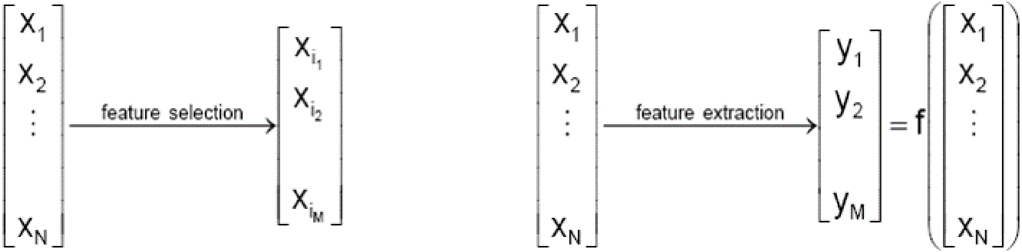


General Aspect

- **Supervised Representation Learning**
  - Supervised learning builds up a mapping from input to targets based on a dataset, then the mapping will be used to predict their targets of test input.
  - Learning done automatically by supervised end-to-end deep neural networks
- **Unsupervised Representation Learning**
  - Unsupervised learning discovers, recovers and models the intrinsic properties from different perspectives.
  - Learning low-dimensional data representation via dimension reduction
  - Learning high-level information summary of data via clustering analysis
  - Recovering/modelling latent explanatory factors via manifold learning
  - Learning data generation and recovering distribution via generative model
  - Disentangling underlying factors of variation via hybrid learning

- **Dimension Reduction**

  A powerful weapon to tackle high-dimensional data, including two methodologies
  - Feature selection: choosing a subset of M features from all N features ($M < N$)
  - Feature extraction: creating a low-dimensional representation of M new yet more informative features by a transformation from all N features ($M < N$)



$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature selection}} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_M} \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}\right)$$

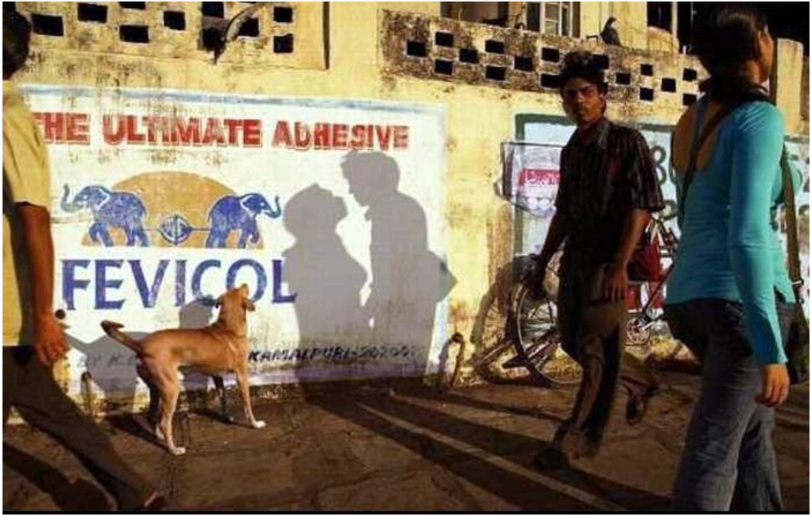
General Aspect

- **Dimension Reduction**

  Caveat: an accidental 2-D projection of the intrinsical 3-D physical world ···

- **Dimension Reduction**

  Caveat: avoid misleading! – an inspirational guideline to dimension reduction



Reference

If you want to deepen your understanding and learn something beyond this lecture, you can self-study the optional references below.

[Bengio et al., 2013] Bengio Y., Courville A., and Vincent P. (2013): Representation learning: A review and new perspective, *IEEE Trans. Pattern Anal. Mach. Intell.* 35(8): 1798–1827.

[Wikipedia] Curse of Dimensionality, *Wikipedia*.
https://en.wikipedia.org/wiki/Curse_of_dimensionality