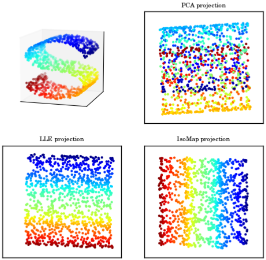
Autoencoders

Suchismit Mahapatra

Motivation

Massive amounts of data

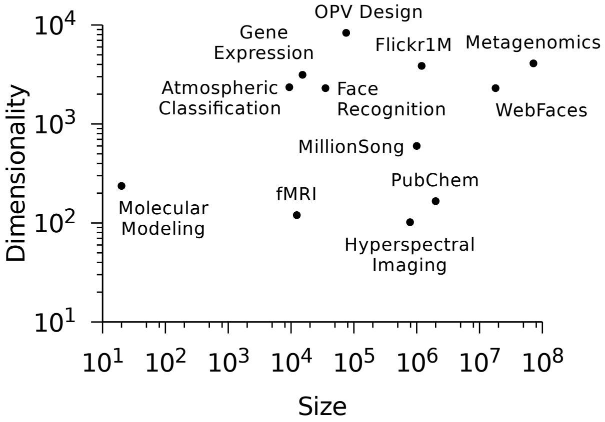
- Natural data tends to be generated by systems (physical or non-physical) that have very few degrees of underlying freedom.
- Real-world data is typically a result of complex non-linear processes, but can often be described by a low-dimensional manifold.



[Credit: Raymond Fu]

Motivation

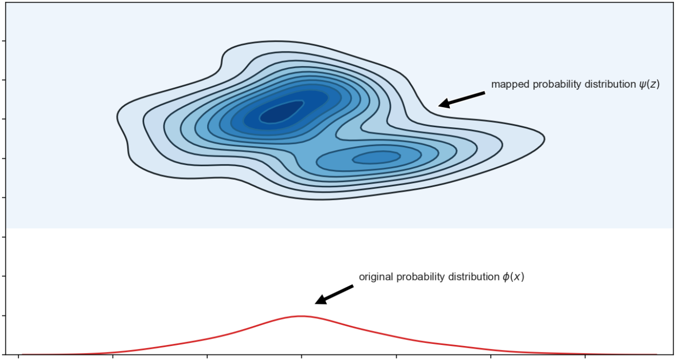
Massive amounts of data



Topology of high-dimensional, massive datasets

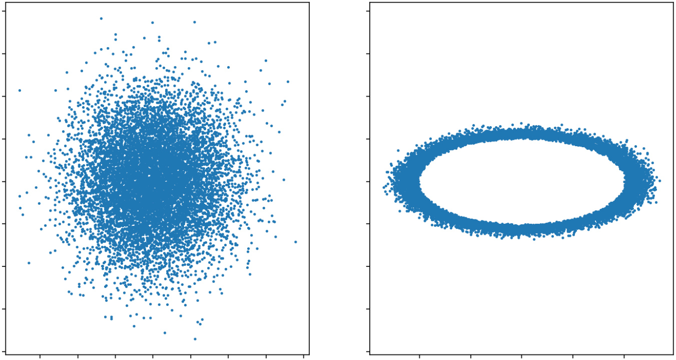
Motivation
Mapping between different probability distributions

Transforming between different probability distributions using non-linear mapping $z = f(x)$



Motivation

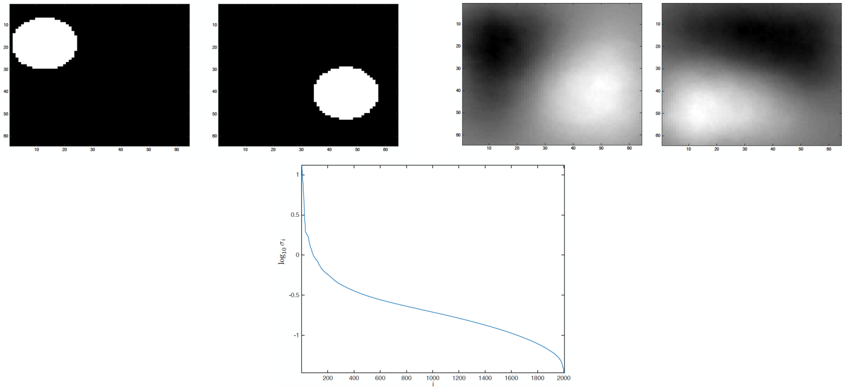
Creating order from noise



Using $\phi(x) = (\frac{x}{|x|} + \epsilon * x)$ to transform random Gaussian noise to a Torus

Motivation

Limitations of Linear Dimension Reduction methods

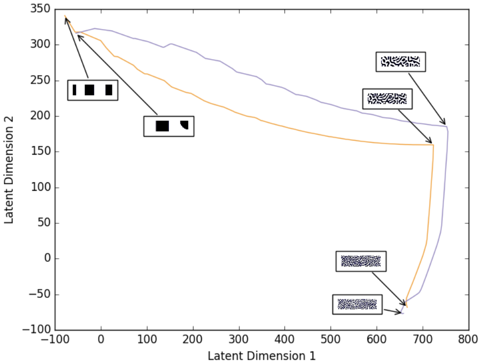


PCA on a simple data set and the intrinsic dimensionality it uncovers, even after using Fourier transformation.

[Credit: Mauro Maggioni]

Latent Spaces

Nonlinear Process Dynamics

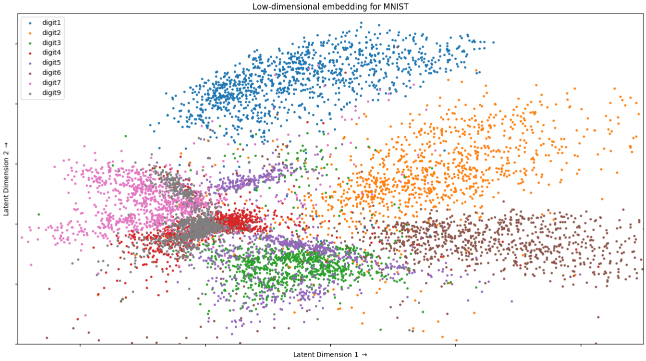


Morphological parametric trajectories for a non-linear process.

[Click here for simulation of all parametric trajectories][Click here for simulation of Manifold]

Manifolds in latent spaces

MNIST data mapped to 2-D



Notice the overlaps between the manifolds of the different digits

Manifolds in latent spaces

Caltech 101 Dataset

[Credit: Laurens van der Maaten]

Manifold Learning

Assumptions

- Distribution of data not uniform.
- Data lives on/near some low-dimensional manifold, typically embedded in high dimensions and separated by low-density regions.

Manifold Learning

Assumptions

- Distribution of data not uniform.
- Data lives on/near some low-dimensional manifold, typically embedded in high dimensions and separated by low-density regions.
- Typically used as a generic non-linear, non-parametric technique to approximate probability distributions in high-dimensional spaces.

Manifold

Properties

Definition

A manifold $\mathcal{M}$ is a metric space with the following property: if $x \in \mathcal{M}$, then there exists some neighborhood $\mathcal{U}$ of x and $\exists n$ such that $\mathcal{U}$ is homeomorphic to $\mathbb{R}^n$.

Manifold

Properties

Definition

A manifold $\mathcal{M}$ is a metric space with the following property: if $x \in \mathcal{M}$, then there exists some neighborhood $\mathcal{U}$ of x and $\exists n$ such that $\mathcal{U}$ is homeomorphic to $\mathbb{R}^n$.

- Global structure can be more complicated.
- Usually embedded in high dimensional spaces, but the intrinsic dimensionality is typically low due to fewer degrees of freedom.

Manifold

Properties

Definition

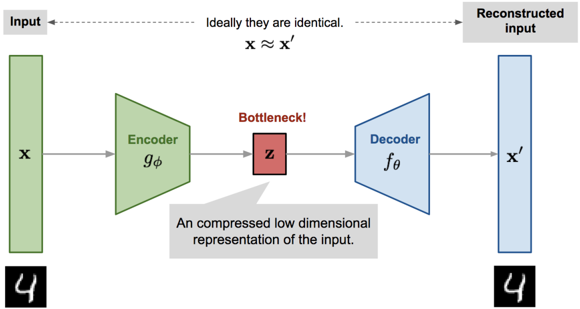
A manifold $\mathcal{M}$ is a metric space with the following property: if $x \in \mathcal{M}$, then there exists some neighborhood $\mathcal{U}$ of x and $\exists n$ such that $\mathcal{U}$ is homeomorphic to $\mathbb{R}^n$.

- Global structure can be more complicated.
- Usually embedded in high dimensional spaces, but the intrinsic dimensionality is typically low due to fewer degrees of freedom.
- Examples
  - Set of queries/product descriptions
  - Image data sets

Non-linear Dimension Reduction

Autoencoder

- Designed to learn the identity function so as to reconstruct the original input while compressing the data to discover a more efficient representation.



[Credit: Lilian Weng]

Autoencoder

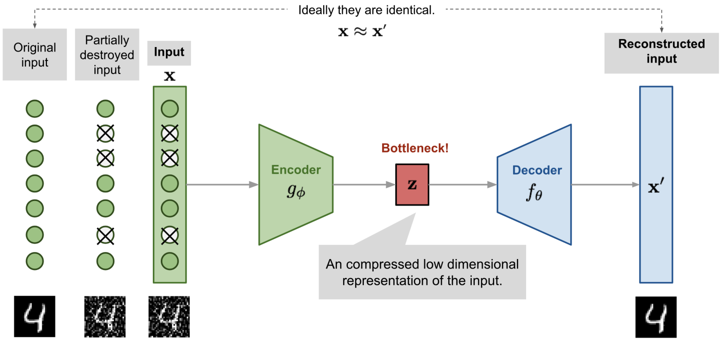
Terminology

- Encoder (g_ϕ) - Translates the original high-dimensional input into latent low-dimensional code.
- Decoder (f_θ) - Recovers data from the latent code.
- Objective function - (ϕ, θ) are learned together to output a reconstructed data sample same as the original input i.e.

$$(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*) = \arg\min_{\phi,\theta} \mathcal{L}_{AE}(\phi, \theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i - f_\theta(g_\phi(x_i)))^2$$

Autoencoder types

Denoising Autoencoder



[Credit: Lilian Weng]

Autoencoder types

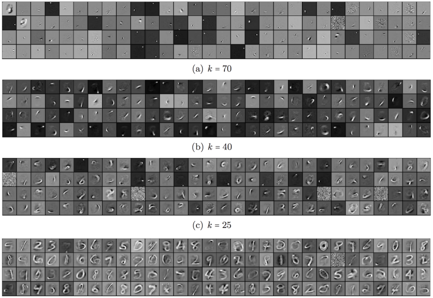
Sparse Autoencoder

- Applies a "sparse" constraint on the hidden unit activation to avoid over-fitting and improve robustness.
- Forces the model to only have a small number of hidden units being activated at the simultaneously.

Autoencoder types

k-Sparse Autoencoder

- Sparsity is enforced by only keeping the top-k highest activated units in the bottleneck layer using a linear activation function.



(a) $k = 70$

(b) $k = 40$

(c) $k = 25$

(d) $k = 10$

[Credit: Lilian Weng]

Autoencoder types

Contractive Autoencoder

- Encourages learnt representations to stay in a "contractive" space.
- Adds regularization term in the loss function to penalize the representation from being too sensitive to the input.

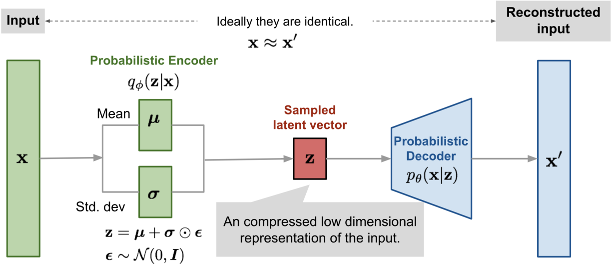
$$\left\| \mathcal{J}_f(x) \right\|_F^2 = \sum_{i,j} \left(\frac{\partial h_j(x)}{\partial x_i} \right)^2$$

- Thus improves the robustness to small perturbations.

Autoencoder types

Variational Autoencoder

- Instead of mapping the input to a fixed vector, the variational autoencoder maps it into a distribution.



[Credit: Lilian Weng]

Variational Autoencoder

Overview

The relationship between high-dimensional input x and the latent code z can be fully defined using :-

- Prior $p_\theta(z)$
- Likelihood $p_\theta(x|z)$
- Posterior $p_\theta(z|x)$

Variational Autoencoder

Overview

- The conditional probability $p_\theta(x|z)$ defines a generative model, similar to decoder $f_\theta(x|z)$. Also known as probabilistic decoder.
- The approximation function $q_\phi(z|x)$ is the probabilistic encoder, similar to $g_\phi(x|z)$.

Variational Autoencoder

Encoder

- Typically implemented using a neural network $q_\phi(z|x)$.
- Takes high-dimensional input x as input and outputs encoding z drawn from a Gaussian distribution parametrized by ϕ.
- Usually $\|z\| \ll \|x\|$.

Variational Autoencoder

Decoder

- Typically implemented using a neural network $p_\theta(x|z)$.
- Takes encoding z as input and outputs a reconstruction x' of the original high-dimensional input x.

Variational Autoencoder

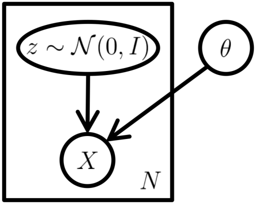
Loss function

- x' is reconstructed from latent code z.
- Ascertains how much information is lost during transition from x to z to x'.
- Measured using reconstruction log-likelihood $\log p_\phi(x|z)$.
- Tells us concretely how well the decoder has reconstructed the original input from the latent code z.

Variational Autoencoder

Probabilistic Model Perspective

- Data x and latent variables z.
- Joint PDF of model $p(x,z) = p(x|z)p(z)$.
- Decomposes it into likelihood $p(x|z)$ and prior $p(z)$ terms.
- Generative process :-
  - Draw $z \sim p(z)$
  - Draw $x \sim p(x|z)$



[Credit: Carl Doersch]