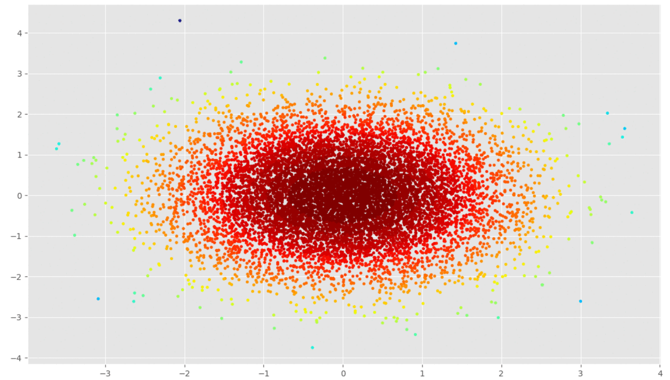
Normalizing Flows

November 4, 2020

Table of contents

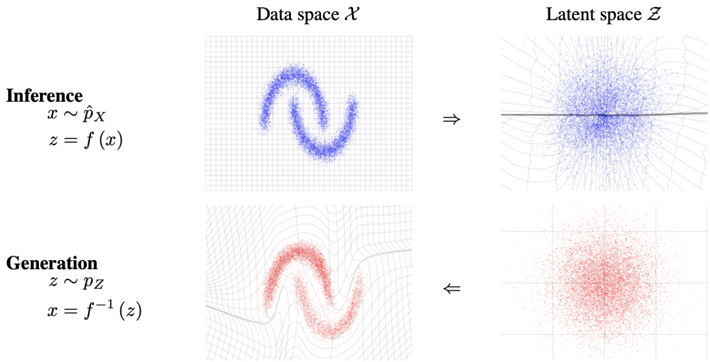
Overview

Main Idea



Normalizing flows provide a general mechanism for defining expressive probability distributions, only requiring the specification of a (usually simple) base distribution and a series of bijective transformations.

Animation Credit: Mike Slawinski, Ph.D., Cylance, Inc.



We learn an invertible mapping between a data distribution $\widehat{p}_X$ and a latent distribution p_Z (typically a Gaussian). The technique allows one to perform anomaly detection and sample generation even for complex and high-dimensional distributions.

Image Credit: Dinh et al. (2017), ICLR.

Density estimation and change of variables

Consider a parametric function mapping continuous random variable X to continuous random variable Z

$$\begin{aligned} f_\theta : X &\to Z \\ x &\mapsto z \end{aligned}$$

where x is an observed sample and $\boldsymbol{z}$ is a latent variable. Suppose p_Z is given. Then by the change of variables theorem, we have[1]

$$\underbrace{\boxed{p_X(x)}}_{\textit{data space}} = \underbrace{\boxed{p_Z(f_\theta(x)) \left| \det \frac{\partial f_\theta(x)}{\partial x} \right|}}_{\textit{transformed latent space}}$$

[1] assuming the conditions of the theorem are met

The goal of density estimation can be posed as follows: learn θ to model unknown data density p_X in terms of assumed latent variable density p_Z.

Normalizing Flow

Definition

(Normalizing Flow) A *(normalizing) flow*, $f = h_\theta^1 \circ \ldots \circ h_\theta^K$, is a sequence of invertible transformations which maps an observed data point, x, to a latent state representation, z.

If we allow ourselves this abuse of notation[2]

$$h_\theta^0 := x$$
$$h_\theta^K := z$$

Then, since $\det \prod_i A_i = \prod_i \det A_i$, the likelihood becomes

$$p_X(x) = p_Z(f_\theta(x)) \prod_{k=1}^{K} \left| \det \frac{\partial h_\theta^k}{\partial h_\theta^{k-1}} \right| \tag{3.1}$$

[2] More generally, we will use the same notation to refer to the function itself as well as its evaluation at a point.

Real NVP

Definition

(Real NVP) A *real NVP* is a normalizing flow (Def. 1) where $f = h_\theta^1 \circ ... \circ h_\theta^K$ is structured such that:

$$h^{i+1} = b^i \odot h^i + (1 - b^i) \odot \left(h^i \odot \exp\left(s_\theta^i(b^i \odot h^i)\right) + t_\theta^i(b^i \odot h^i) \right)$$

where $b^1, ..., b^K$ is a sequence of binary masks, $\odot$ is the Hadamard product or element-wise product, and s and t stand for scale and translation.

Definition

(Affine Coupling Layer) An affine coupling layer is one element of the sequence of invertible transformations in a real NVP; i.e. it is h_i for some $i \in \{1, ..., K\}$ in Def. 2.

Example

If random variables are D dimensional, and $b^i := [1, ...1, 0, ...0]$, where the 0 entries begin at the d_{i+1}st element, then the affine coupling layer is given by

$$\begin{aligned} h^{i+1}_{1:d_i} &= h^i_{1:d_i} \\ h^{i+1}_{d_i+1:D} &= h^i_{d_i+1:D} \odot \exp\left(s^i_\theta(h^i_{1:d_i})\right) + t^i_\theta(h^i_{1:d_i}) \end{aligned}$$

Note that the real NVP allows for efficient computation of the determinant of the Jacobians, since

$$\frac{\partial h_\theta^{i+1}}{\partial h_\theta^i} = \begin{pmatrix} \mathbb{I}_d & 0 \\ \frac{\partial h_{d_i+1:D}^{i+1}}{\partial h_{1:d_i}^i} & \operatorname{diag}\left(\exp\left(s_\theta(h_{1:d_i}^i) \right) \right) \end{pmatrix}$$

The bottom left term can be arbitrarily complex; we don't have to compute it, since the determinant of a triangular matrix is the product of the diagonals:

$$\det \frac{\partial h_\theta^{i+1}}{\partial h_\theta^i} = \exp\left(\sum_j s_\theta^i \left(h_{1:d_i}^i \right)_j \right)$$

So, by Equation 3.1, the log likelihood with real NVP normalizing flow applied to a single data sample, x, is

$$\log p_X(x) = \log p_Z(f_\theta(x)) + \sum_{\substack{i \\ \textit{(LAYERS)}}} \sum_{\substack{j \\ \textit{(FEATURES)}}} s_\theta^i \left(h_{1:d_i}^i\right)_j$$

And the log likelihood applied for a collection of samples, assumed *i.i.d.*, is the sum of individual log likelihoods.