

Normalizing Flows

November 4, 2020

Table of contents

1. Overview
2. Density estimation and change of variables
3. Normalizing Flow
4. Real NVP

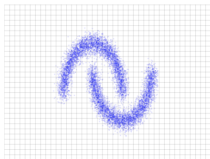
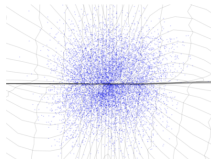
Overview

Main Idea

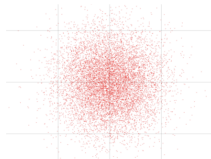
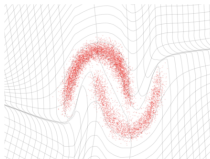
Normalizing flows provide a general mechanism for defining expressive probability distributions, only requiring the specification of a (usually simple) base distribution and a series of bijective transformations.

Inference

$$x \sim \hat{p}_X$$
$$z = f(x)$$

Data space \mathcal{X} Latent space \mathcal{Z} **Generation**

$$z \sim p_Z$$
$$x = f^{-1}(z)$$



We learn an invertible mapping between a data distribution \hat{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

$$f_{\theta} : X \rightarrow Z$$

$$x \mapsto z$$

where x is an observed sample and z is a latent variable. Suppose p_Z is given. Then by the change of variables theorem, we have¹

$$\underset{\text{data space}}{p_X(x)} = \underset{\text{transformed latent space}}{p_Z(f_{\theta}(x)) \left| \det \frac{\partial f_{\theta}(x)}{\partial x} \right|}$$

¹assuming 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, $f = h_{\theta}^1 \circ \dots \circ h_{\theta}^K$, 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 notation²

$$h_{\theta}^0 := x$$

$$h_{\theta}^K := z$$

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

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

²More 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 $f = h_{\theta}^1 \circ \dots \circ h_{\theta}^K$ is structured such that:

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

where b^1, \dots, 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_k for some $k \in \{1, \dots, K\}$ in Def. 2.

Example

If random variables are D dimensional, and $b^k := [1, \dots, 1, 0, \dots, 0]$, where the 0 entries begin at the $(d_k + 1)$ st element, then the affine coupling layer is given by

$$\begin{aligned}h_{1:d_k}^{k+1} &= h_{1:d_k}^k \\h_{d_k+1:D}^{k+1} &= h_{d_k+1:D}^k \odot \exp(s_{\theta}^k(h_{1:d_k}^k)) + t_{\theta}^k(h_{1:d_k}^k)\end{aligned}$$

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

$$\frac{\partial h_{\theta}^{k+1}}{\partial h_{\theta}^k} = \begin{pmatrix} \mathbb{I}_{d_k} & 0 \\ \frac{\partial h_{d_k+1:D}^{k+1}}{\partial h_{1:d_k}^k} & \text{diag}\left(\exp(s_{\theta}(h_{1:d_k}^k))\right) \end{pmatrix}$$

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:

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

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

$$\log p_X(x) = \log p_Z(f_\theta(x)) + \sum_k \sum_j s_\theta^k (h_{1:d_k}^k)_j$$

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