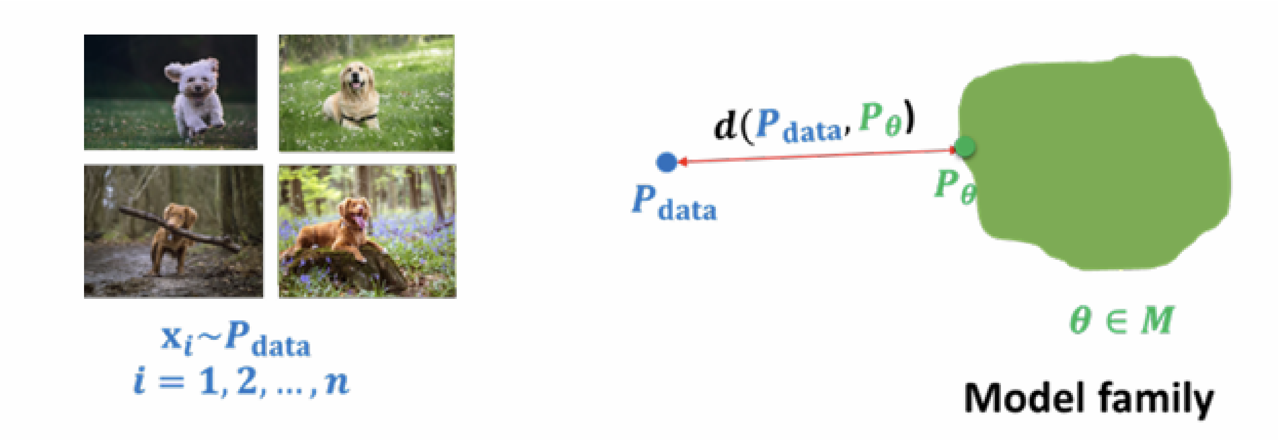
Notes1.Representation

1 1.What is a generative model?

We are given a training set of examples， the assumption is these data points are sampled form **unknown probability** distribution. We denote P_{data}

The problem is to basically come up with a good approximation of this data generating process.



Model family : All possible distribution(eg .Guassian with different μ and Σ)

The goal become to find a good approximation within the set.

define distance :loss function $\mathcal{L}$

optimization: We want P_θ relatively close to P_{data}

We want to learn a probability distribution p(x) over images x such that

Generation: If we sample $x_{new} \sim p(x)$, x_{new} should look like a dog (sampling)

Density estimation: $p(x)$ should be high if x looks like a dog, and low otherwise (anomaly detection)

Unsupervised representation learning: We should be able to learn what these images have in common, e.g., ears, tail, etc. (features)

1.1 How to present p(x)?

1.1.1 Example of joint distribution

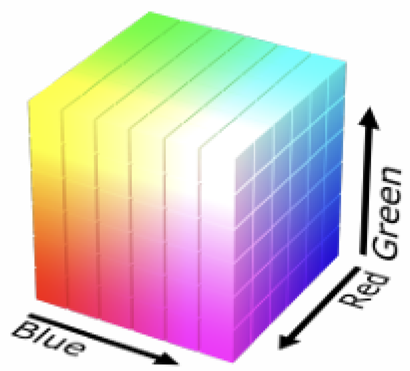
Modeling a single pixel’s color.

Three discrete random variables:

Red Channel R. Val(R) = $\{0,\cdots,255\}$

Green Channel G. Val(G) = $\{0,\cdots,255\}$

Blue Channel B. Val(B) = $\{0,\cdots,255\}$

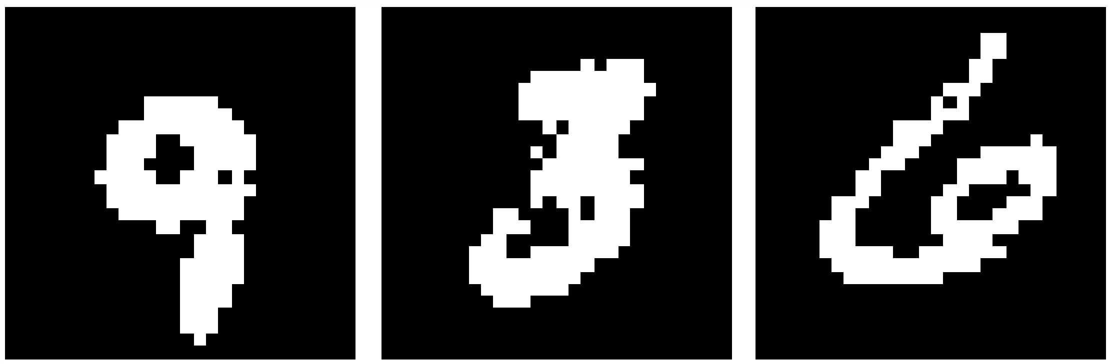


Sampling from the joint distribution (r,g,b) $\sim$ p(R,G,B)

randomly generates a color for the pixel.

How many parameters do we need to specify the joint distribution p(R = r,G = g,B = b)?

$256 * 256 * 256 - 1$



Suppose X1,...,Xn are binary (Bernoulli) random variables, i.e., Val(Xi) = {0,1} = {Black,White}.

How many possible images (states)?

$2 \times 2 \times \cdots \times 2$ (n times) $= 2^n$

Sampling from p(x1,...,xn) generates an image

How many parameters to specify the joint distribution p(x1,...,xn) over n binary pixels?

$2^n - 1$

A strong assumption is that all the r.v. are idpendent,then

We can use much less parameters!!!

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2) \cdots p(x_n)$$

$2^n - 1 \rightarrow n$

However the assumption is too strong that we might ignore the structure

So we need the rules: Chain rules and Bayes Rule

If we use the chain rules

$p(x_1, \cdots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \cdots, x_{n-1})$

We still need 2^{n-1} parameters(in fact we don't do any assumption,No free launch)

p(x_1) requires 1 param

$p(x_2|x_1)$ require 2 params x1 =0 1param x1=1 1param total 2params

...

Markov chain:

Now suppose $X_{i+1} \perp X_1, \ldots, X_{i-1} \mid X_i$, then

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid \cancel{x_1}, x_2) \cdots p(x_n \mid \cancel{x_1, \cdots}, x_{n-1}) \\ &= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2) \cdots p(x_n \mid x_{n-1}) \end{aligned}$$

Now we just need 2n-1 !!!

2 2. Representing probability distributions

2.1 Curse of dimensionality

2.2 Crash course on graphical models (Bayesian networks)

2.3 Generative vs discriminative models

2.4 Neural models