NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

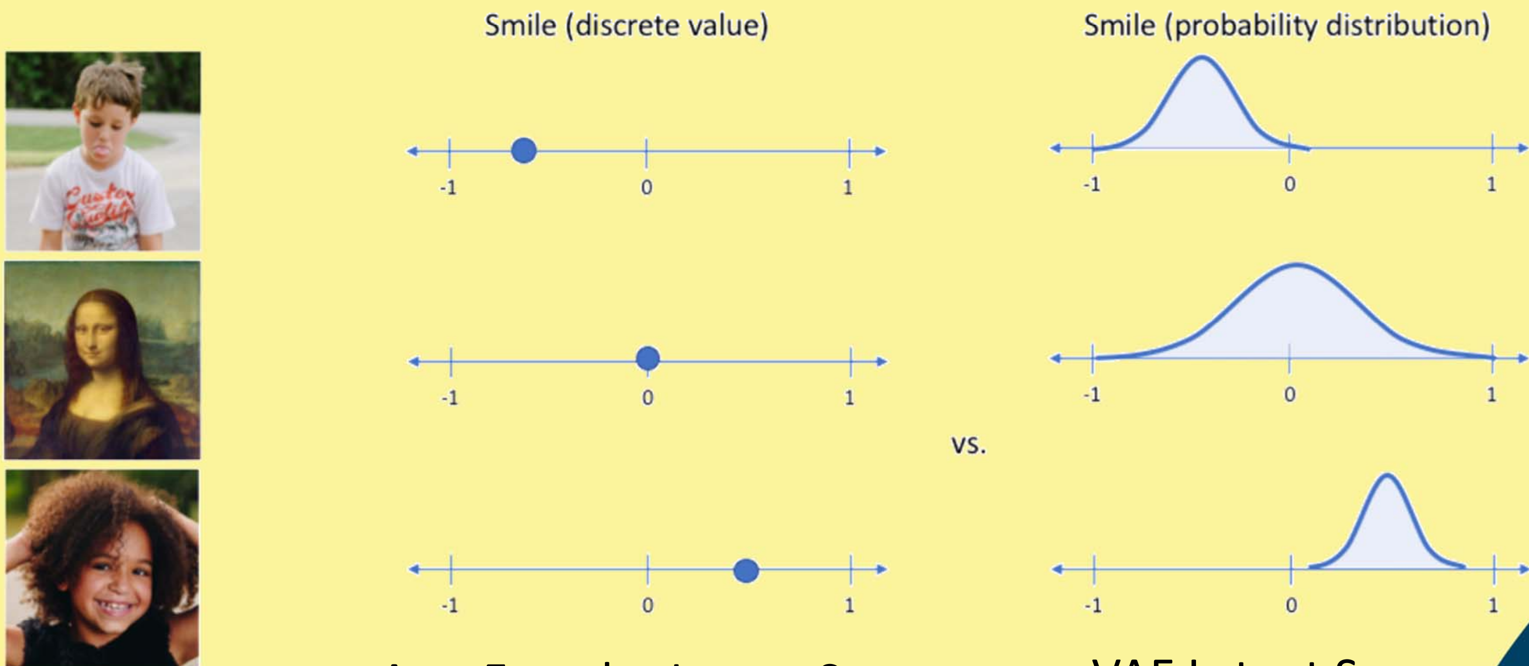
Lecture 58: Variational Autoencoder - II

CONCEPTS COVERED

Concepts Covered

- ❑ Generative Model
- ❑ Limitations of usual auto-encoder
- ❑ Intuitions behind VAE
- ❑ Variational Inference
- ❑ Practical Realization of VAE

Autoencoder Intuition vs. VAE Latent Space



AutoEncoder Latent Space

VAE Latent Space

https://www.jeremyjordan.me/variational-autoencoders/

Variational Autoencoder Intuition

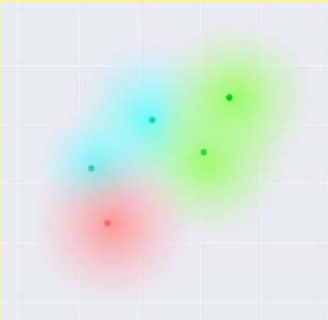
- Instead of deterministic latent code we might be interested to learn a distribution over the latent code
- For example, it is more intuitive to determine a range of “smile” value for a face instead of an absolute “smile” value
- Instead of deterministic code, we will now output the mean and standard deviation of each component of the vector (assuming each component is independent of each other)

Variational Autoencoder Intuition

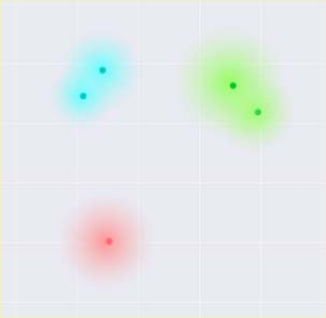
- With this setup we can represent each latent factor as a probability distribution
- We can sample from such distribution
- Then the sampled vector can be passed through Decoder (Generator) to generate an image

Variational Autoencoder Intuition

- For smooth interpolations, ideally, we want overlap between samples that are not very similar too, in order to interpolate between classes.
- However μ and σ can take any value and learn to cluster the mean vectors of different classes far apart (and minimize σ) to reduce uncertainty for the Decoder



Our goal



Network might converge to

Image Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

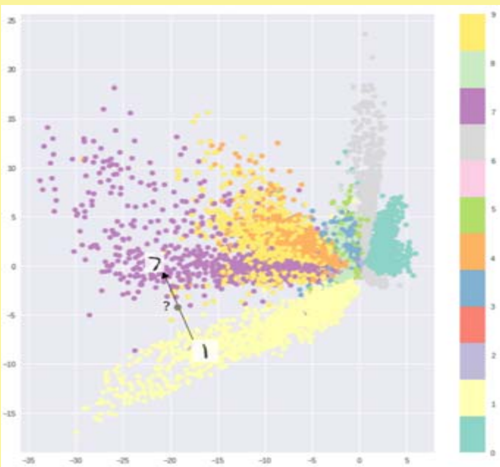
Variational Autoencoder Intuition

- In order to enforce smooth transition we will apply Kullback–Leibler divergence (KL divergence) between the distribution of encoded vectors and a prior distribution asserted on latent distribution space
- KL divergence between two probability distributions simply measures how much they diverge from each other.
- Minimizing the KL divergence here means optimizing the probability distribution parameters (μ and σ) to closely resemble that of the target distribution.

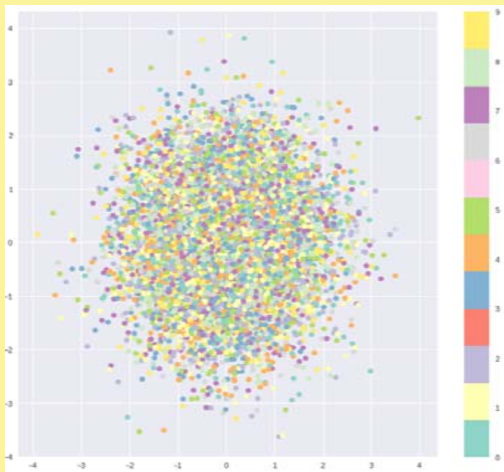
Variational Autoencoder Intuition

- In VAE, it is usually assumed that the distribution of the latent space follows a zero mean Normal distribution with diagonal covariance matrix (each component is independent of the other)
- KL divergence loss will encourage encodings from different inputs to be clustered about the center of the latent space
- If network creates clusters in specific regions then KL divergence loss will penalize such clusters formation

Variational Autoencoder Intuition



Reconstruction Loss



KL Divergence Loss

KL Divergence + Reconstruction Loss

Image Source: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Variational Autoencoder Intuition

- This equilibrium is attributed to cluster-forming nature of the reconstruction loss, and the dense packing nature of the KL loss
- It means when randomly generating, if you sample a vector from the prior distribution, $P(z)$ of latent space, the Decoder will successfully decode it.
- For interpolation, since there is no sudden gap between clusters, but a smooth mix of features, a Decoder can understand.

Variational Autoencoder : Variational Inference

- In VAE, we assume that there is a latent (unobserved) variable, z, generating our observed random variable, x.

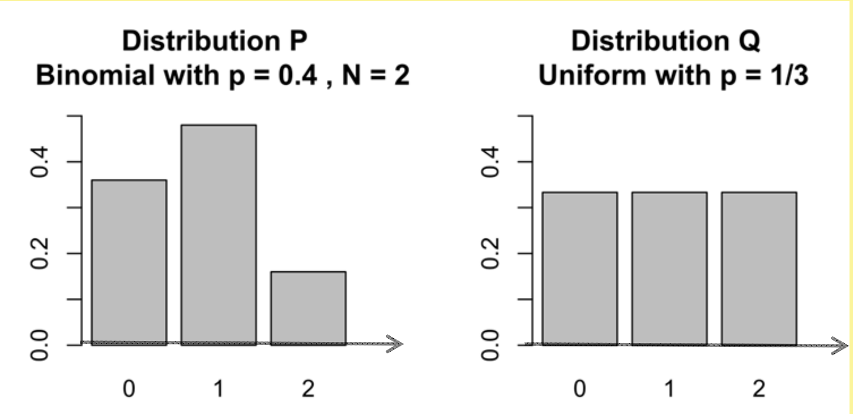
- Our aim: To compute the posterior $P(z|x) = \frac{P(X|Z)P(z)}{P(x)}$
- $P(x) = \int P(x|z)P(z)dz$ $\Longrightarrow$ Intractable

Variational Autoencoder : Variational Inference

- Let's assume there is a tractable distribution Q, such that $P(z|x) \approx Q(z|x)$
- We want $Q(\cdot)$ to be in the family of tractable distributions (Gaussian for example) such that we can play around with its parameters to match P$(z|x)$
- So, we will aim towards minimizing KL divergence of P$(z|x)$ with respect to $Q(z|x)$
- Our objective: minimize KL($Q(z|x)$ || P$(z|x)$)

KL Divergence

KL$(Q(z|x) \; || \; \mathrm{P}(z|x)) = \sum_x Q(x) log \frac{Q(x)}{P(x)}$



x	0	1	2
P(x)	0.36	0.48	0.16
Q(x)	0.33	0.33	0.33

KL Divergence

$$KL(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

$$= 0.36 \log\left(\frac{0.36}{0.33}\right) + 0.48\ log\left(\frac{0.48}{0.33}\right) + 0.16\ log\left(\frac{0.16}{0.33}\right) = 0.0414$$

x	0	1	2
P(x)	0.36	0.48	0.16
Q(x)	0.33	0.33	0.33

$$KL(Q||P) = \sum_x Q(x) log \frac{Q(x)}{P(x)}$$

$$= 0.33 \log\left(\frac{0.33}{0.36}\right) + 0.33\ log\left(\frac{0.33}{0.48}\right) + 0.33\ log\left(\frac{0.33}{0.16}\right) = 0.0375$$

KL Divergence

Minimize

$$KL(Q(z|x)||P(z|x))$$

KL Divergence

$$KL(Q(z|x)||P(z|x))$$

$$= -\sum_{z} Q(z|x) \log \frac{P(z|x)}{Q(z|x)}$$

$$= -\sum_{z} Q(z|x) \log \frac{P(x,z)}{P(x) * \ Q(z|x)}$$

KL Divergence

$$= -\sum_z Q(z|x)\left\{log\frac{P(x,z)}{Q(z|x)} - logP(x)\right\}$$

$$= -\sum_z Q(z|x)\log\frac{P(x,z)}{Q(z|x)} + \sum_z Q(z|x)\log P(x)$$

$$= -\sum_z Q(z|x)\log\frac{P(x,z)}{Q(z|x)} + logP(x)$$

KL Divergence

$$KL(Q(z|x)||P(z|x)) = -\sum_{z} Q(z|x)\log\frac{P(x,z)}{Q(z|x)} + \ logP(x)$$

$$\downarrow$$

$$\log P(x) = KL(Q(z|x)\|P(z|x)) + \sum_{z} Q(z|x)\log\frac{P(x,z)}{Q(z|x)}$$

KL Divergence

$$\log P(x) = KL(Q(z|x)||P(z|x)) + \sum_z Q(z|x) \log \frac{P(x,z)}{Q(z|x)}$$

- Since, x is given, LHS is constant.
- Aim is to minimize $KL(Q(z|x)||P(z|x))$
- This is same as maximizing $\sum_z Q(z|x) \log \frac{P(x,z)}{Q(z|x)}$

NPTEL ONLINE CERTIFICATION COURSES

Thank you