Resources:

Demo: http://dpkingma.com/sgvb_mnist_demo/demo.html

https://www.youtube.com/watch?v=9zKuYvjFFS8&t=26s

https://www.youtube.com/watch?v=vy8q-WnHa9A

https://www.youtube.com/watch?v=nKM9875PVtU&t=8s

https://www.youtube.com/watch?v=YgSWrafXI8U&list=PLTKMiZHVd_2KJtIXOW0zFhFfBaJJilH51&index=142

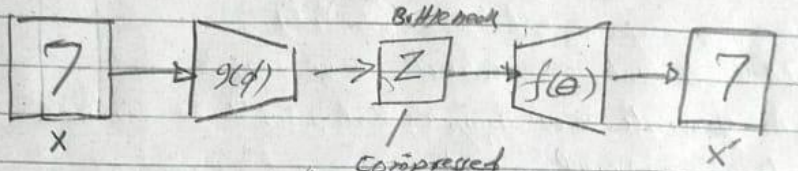
https://www.youtube.com/watch?v=w8F7_rQZxXk&list=PLdxQ7SoCLQANQ9fQcJ0wnnTzkFsJHlWEj&index=19

Variational Autoencoder

- Review of stacked Autoencoder
- Basic of probability
- KL Divergence & its significance
- Derivation of loss function for variational autoencoder

Basic Autoencoder

- Architechre: Same image reconstruction



- Applications
  1. Generate compressed image (Dimensional Reduction like PCA)
  2. Denoising (noise removal)
  3. Image inpainting
  4. Image segmentation

$$L(\theta, \phi) = \frac{1}{m} \sum_{i=1}^{m} \left[x^{(i)} - f_\theta(g_\phi(x^{(i)})) \right]^2$$

- If no Activation Function is used in autoencoder with 1 hidden layer, this model is sort of identical to PCA

→ In PCA $x \in R^n$

$$Z_{P\times 1} = U^T_{P\times n}\, x_{n\times 1}$$

$Z \in R^P$

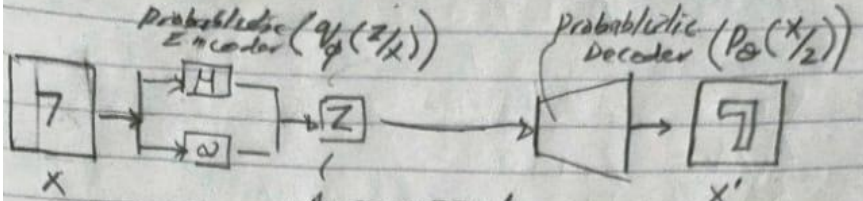
$$\hat{x} = U_{n\times P}\, Z_{P\times 1}$$

$\hat{x} \in R^n$

$\min \| x - UU^T \hat{x} \|$

[Due to activation function]

Variational Auto encoder.



- An compressed low dimensional representation of the input
- Sampled latent vector

$$z = \mu + \omega * \epsilon \quad \text{(Sample Z from Gaussian distribution)}$$
$$\epsilon \sim N(0, I)$$

- VAE is not reconstruct same sample, but it generate new sample that should be similar to the original sample.

$$L(\theta, \phi) = \underset{z \sim q_\phi(z/x)}{E} [P_\theta(x/z)] + D_{KL}(q_\phi(z/x) \| P_\theta(z))$$

Encoder part — Encoder part

learned distribution ≈ prior distribution

- See difference b/w autoencoder (AE) and variational auto encoder (VAE)

- $P(x)$: Define the probability of random variable x.

Example: When a die is tossed once, what is the probability of getting 3.

$X = \{1, 2, 3, 4, 5, 6\}$

X : random variable. Sample space of experiment

R.V is a function that assign real no to the sample space of Experiments

$P(3) = 1/6$

- $P(X/Y)$: Define as the prob. of random variable X provided Y has happened. Also called an conditional probability

Example: In tossing a fair die, what is the prob. the 3 has occured conditioned on the face being odd.

$X = \{1, 3, 5\}$

sample space of experiment (only odd value)

$P(3/X) = P(3/(1,3,5)) = 1/3$

* $P(3)$ is increased in case of $P(3/x \text{ being odd})$

Bayes Theorm:

$$P(y/x) = \frac{P(x/y)\,P(y)}{P(x)}$$

Posterior prob — likelihood ratio — prior prob (prior info. readily available in the system)

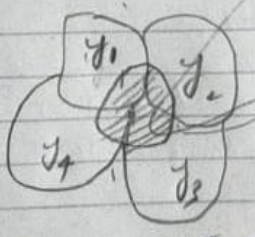
$$P(y/x) = \frac{P(x,y)}{P(x)} \quad \text{(joint prob.)} \quad \text{—— (1)}$$

Theorm of Total Probability

Let $y_1, y_2 - - y_n$ be a set of mutual exclusive events (i.e $y_i \cap y_j = \phi$) and event (x) is the union of n mutually exclusive events, then

$$P(x) = \sum_{i=1}^{n} P(x/y_i)\,P(y_i) \quad \text{—— (2)}$$

- Let y_1, y_2, y_3, y_4 are mutual Exclusive events



Union of 4 ~~events~~ ME

$$P(x) = \sum_{i=1}^{4} P(x, y_i)$$

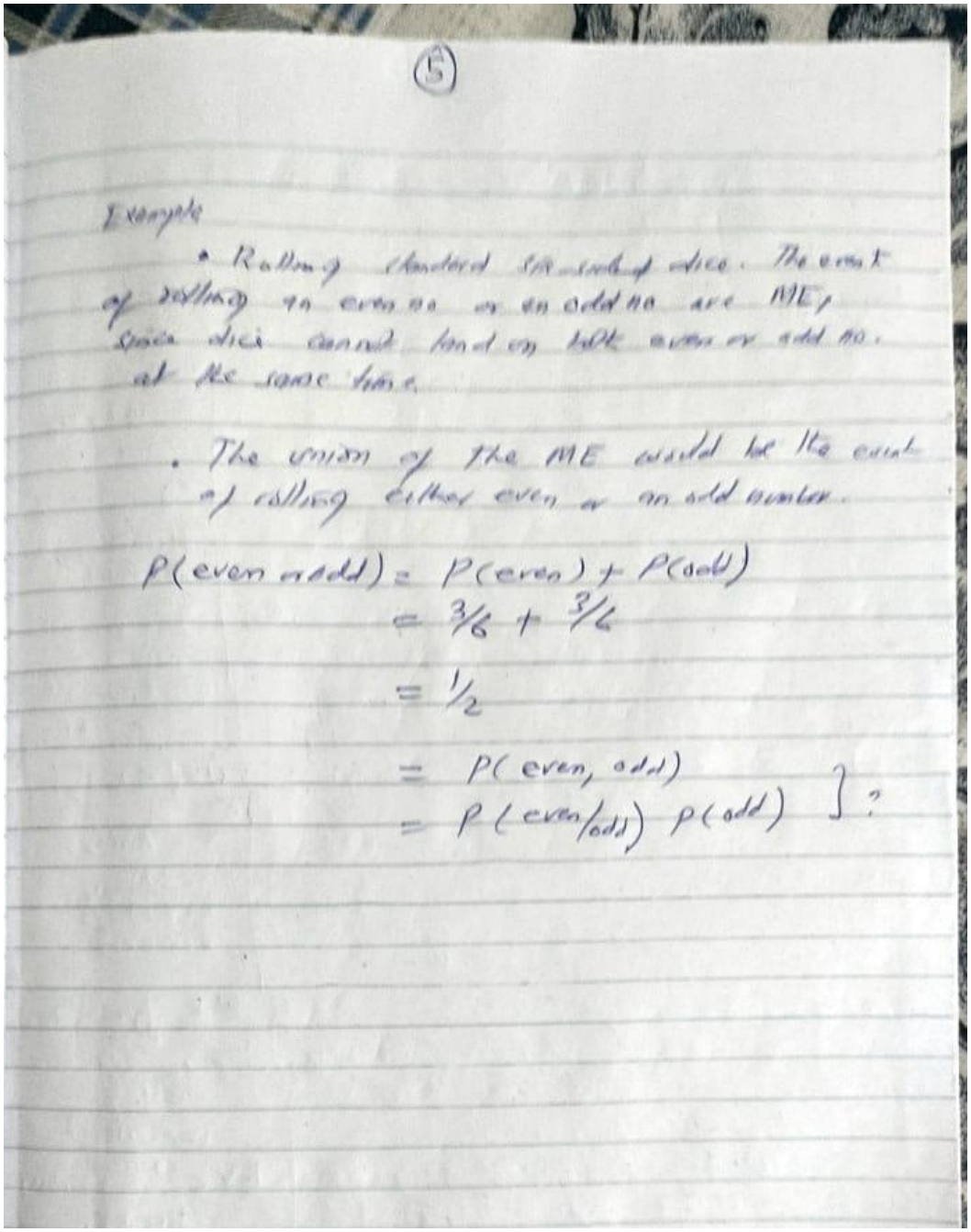
$$P(x) = P(x,y_1) + P(x,y_2) + P(x,y_3) + P(x,y_4)$$

or $P(x) = P(x/y_1)P(y_1) + P(x/y_2)P(y_2) + P(x/y_3)P(y_3) + P(x/y_4)P(y_4)$

By substituting Eq (2) in Eq (1)

$$P(y/x) = \frac{P(x,y)}{\sum_{i=1}^{n} P(x/y_i)P(y_i)} = \frac{P(x/y)\,P(y)}{\sum_{i=1}^{n} P(x/y_i)P(y_i)}$$

- **Example**:
    - Rolling a standard six sided die. The events of rolling an **even** number ({2, 4, 6}) and an **odd** number ({1, 3, 5}) are mutually exclusive. Since die can not land on both even and odd number at the same time.



weighted average

Expectation of random variable X. $E(X)$

Expectation value of random variable is a weighted average of the possible values of X that it can take, each value being weighted according to the probability of that event defined as

$$\underset{p}{E(x)} = E(x) = \sum_{i=1}^{k} x_i \underbrace{P(X = x_i)}_{\text{weight}}$$

(x_i: state) ← // for discrete RV

x has the Prob. density function p

continuous RV

$$E(X) = \int_{\ell} x_i P(X = x_i)$$

Example. let x represent the outcome of a fair die. What is the $E(x)$?

$$x = \{1, 2, 3, 4, 5, 6\}$$

$$\underset{p}{E(x)} = \sum x_i P(x_i) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6}$$

$$= 3.5$$

↔ also

we can compute average of Function of random variable

$$\underset{p}{E}[h(x)] = \sum_{i=1}^{k} h(x_i) P(X = x_i)$$ // For Function discrete RV

$$\underset{p}{E}[h(x)] = \int_{R} h(x) \cdot P(x)\, dx$$ // for Function continuous RV

eg $\underset{p}{E}\left[\log\left(\frac{P(x)}{q(x)}\right)\right] = \sum_{i=1}^{\infty} \log\left(\frac{P(x)}{q(x)}\right) \times P(x = x_i)$

$= P(x) \times \sum \log \frac{P(x)}{q(x)} = \frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{P(x)}{q(x)}\right) \times (1 \times 1)$ $= P(x = x_i) \times \sum_{i=1}^{\infty} \log\left(\frac{P(x)}{q(x)}\right)$

Information

- We can quantify the information that a statement can carry by the following expression

$$I = -\log P(x) \qquad \text{// log likelihood}$$

x = is a certain event
$P(x)$ = is the prob of x; it has value $0 \to 1$
If $P(x)$ is small, $-\log(P(x))$ is very large
If $P(x)$ is large, $-\log(P(x))$ is very small
- unusual events I has high information

eg
(a) → If we have a statement
"In July the weather will be very hot"

its $P(x) = 1$
$\therefore I = -\log P(x) = -\log 1 = 0$

- It is very much likely in July weather will be very hot so $P(x) \to 1$ & $I = 0$, there will be no information

(b) But if we say in July there is a snow, It is very unlikely, but it contains lot of information.

$P(x) \simeq 0$ (as it is very unlikely)
$I = -\log p(x) = -\log(\text{small no})$
$I =$ large value

In a fair coin, occurrence of head or tail is an event

∵ unpredictable / unusual event contains lot of information.

ENTROPY

Average of Information / Expectation of information is called ENTROPY.

$$H = -\sum P(x) \log P(x) \rightarrow \text{Information}(I=x)$$

weighted sum

$E[X] = x * P(x)$
$= Px * x$

KL DIVERGENCE (Kullback-Leibler)

- It has close connection with Information or Entropy
- It is also called relative entropy.
- It show dissimilarity b/w 2 different distribution
- It is a measure of how one prob. distribution is different from the 2nd, reference distribution
- For the discrete prob. distributions P & Q, the KL divergence b/w P & Q is defined as

$$D_{KL}(P||Q) = +\sum_x P(x=x) \log P(x) - \sum_x Q(x=x) \log Q(x=x)$$

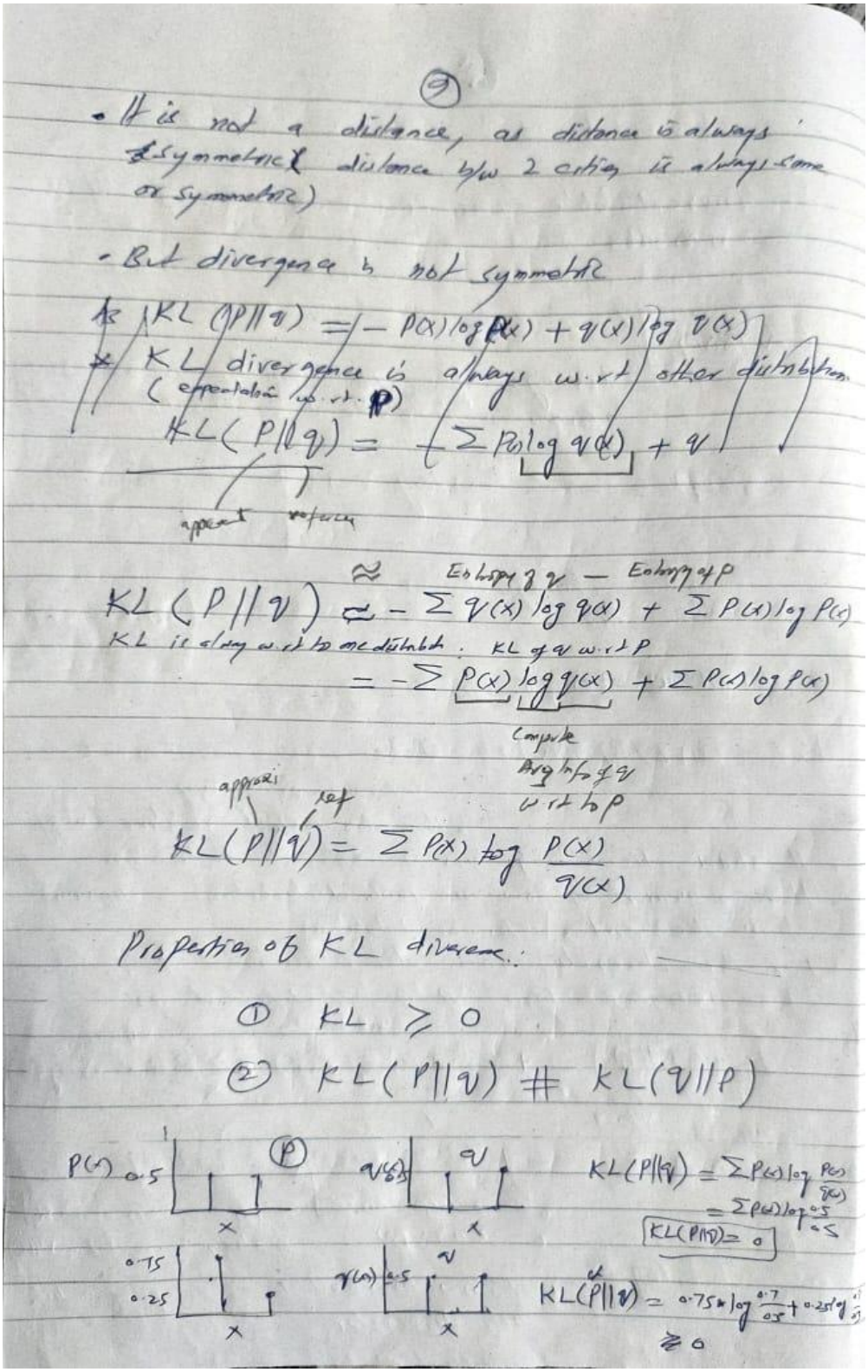
(P: approximate, Q: reference)

$$= +\sum_x P(x) \log P(x) - \sum_x Q(x) \log(Q(x))$$

→ [average info about P] − [avg info about Q]

→ Entropy P − Entropy Q ⇒ KL(P||Q)

→ it show how much the average of 2 dist. are dissimilar



1) KL(p||q) = 0.5 log (0.5/0.5)+ 0.5 log (0.5/0.5) =0

KL(p||q) = 0.75 log (0.75/0.5)+ 0.25 log (0.25/0.5) = 0.0568

2) **KL(p||q) not equal to KL(q||p), may find by solving second expression for**
KL(q||p) = 0.5 log (0.5/0.75)+ 0.5 log (0.5/0.25) = 0.0624

Suppose we have 2 multivariate normal distributions denoted as

$$p(x) = N(x; \mu_1, \Sigma_1)$$

$$q(x) = N(x; \mu_2, \Sigma_2)$$

where μ_1 & μ_2 are means and $\Sigma_1 \neq \Sigma_2$ are the covariance matrix.

- The multivariate normal density function is denoted as

$$p(x) = N(x; \mu_1, \Sigma_1) = \frac{1}{\sqrt{(2\pi)^k |\Sigma_1|}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)\right)$$

$$q(x) = N(x; \mu_2, \Sigma_2) = \frac{1}{\sqrt{(2\pi)^k |\Sigma_2|}} \exp\left(-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2)\right)$$

- If the bolt have some dimension k.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$ ← each variable follows $N(0,1)$

Recap:

- Multivariate PD is used to describe the behaviour of multiple R.V that are correlated. It is extension of univariate normal distribution to 2 or more variables where each variable follows a normal distribution & joint distribution can be represented by mean vector & covariance matrix.

- PDF for single variable $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Now KL divergence b/w P & Q

$$D_{KL}(P(x) \| q(x)) = \frac{1}{2}\left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - K + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2)\right]$$

Prove ?

Proof

$$KL(P(x) \| Q(x)) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad -(1)$$

we know multivariate normal distribution can be defined as

$$P(x) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_1|}} \exp\left(-\frac{(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)}{2}\right)$$

↑ determinant of Σ_1

$$P(x) = (2\pi)^K |\Sigma_1|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right)$$

$$P(x) = (2\pi)^{-K/2} \cdot |\Sigma_1|^{-1/2} \cdot \exp\left[\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right]\right.$$

Taking log on both sides

$$\log P(x) = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \left[\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right] \quad -(2)$$

Similarly for $Q(x)$ prob. density function

$$\log Q(x) = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_2| - \left[\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right] \quad -(3)$$

Equation ① can be written as

$$KL(P(x) \| Q(x)) = \sum_x P(x)[\log P(x) - \log Q(x)]$$

Now Put equation ② & ③ in above equation

$$KL(P(x) \| Q(x)) = \sum_x P(x)\left[-\tfrac{K}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma_1| - \tfrac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \tfrac{K}{2}\log(2\pi) + \tfrac{1}{2}\log|\Sigma_2| + \tfrac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$$

$$KL(P(x) \| Q(x)) = \sum_x P(x)\left[\tfrac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} + \tfrac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) - \tfrac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right]$$

(Y)

How To solve these ② expression
let consider part by part

$$\sum_x P(x)\cdot\tfrac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) = E_p\left[\tfrac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]$$

As we know

- Expectation of RV

$$E_p[x] = \sum_{i=1}^{\infty} x_i \cdot \underbrace{P(x)}_{\text{weight}}$$

(i → state)

- Expectation of Function of RV

$$E_p[h(x)] = \sum_{i=1}^{\infty} h(x)\cdot P(x)$$

Note

$$\sum_x P(x)\tfrac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} = E_p\left[\tfrac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}\right] = \tfrac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|}$$

as $E_p[\text{const}] = \text{const}$

Recap linear Algebra.

- If X is square matrix

① $\begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}$ $\sum_{i=1}^{n} d_i$

Trace & Expectation trick

Now

$E(X^T A X) = E(tr(X^T A X))$ — ①

$= E(tr(A X X^T))$ — ②

$= tr(E(A X X^T))$ — ③

Let's rewrite the expression of previous page

$\frac{1}{2} E_p[(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)]$

$= E_p[tr(\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)]$

$= E_p[tr(\frac{1}{2}(x-\mu_1)(x-\mu_1)^T \Sigma_1^{-1})$ — ④

$= tr(E_p[\frac{1}{2}(x-\mu_1)(x-\mu_1)^T]\Sigma_1^{-1})$ — ⑤

(Covariance matrix ⑦)

$= tr[\Sigma_1 \frac{1}{2} \Sigma_1^{-1}] = tr[I_k]$

$= k$

$k =$ $\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{bmatrix}$

Trace & Expectation trick

→ If x is scalar, then

$E(x) = E(tr(x))$ — ⑧

• since trace of x is scalar

→ $tr(AB) = tr(BA)$ — ⑥

→ $tr(ABC) = tr(BCA) = tr(CAB)$ — ⑨

→ $tr(ABC) \neq tr(ACB)$

→ $E(tr(x)) = tr(E(x))$ — ⑩

Now solve the other expression from (1)

$$\sum_x P(x)\left[\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]$$

→ Note add & subtract μ_1

$$= \sum_x P(x)\left[\frac{1}{2}\underbrace{[(x-\mu_1)}_{A} + \underbrace{(\mu_1-\mu_2)}_{B}]^T\Sigma_2^{-1}[\underbrace{(x-\mu_1)}_{A}+\underbrace{(\mu_1-\mu_2)}_{B}]\right]$$

$$= \sum_x P(x)\left[\frac{1}{2}(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1) + \frac{2}{2}(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2) + \frac{1}{2}(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

$$(A+B)^T\Sigma^{-1}(A+B)$$
$$= (A^T+B^T)\Sigma_2^{-1}(A+B)$$
$$= A^T\Sigma_2^{-1}A + B^T\Sigma_2^{-1}B + B^T\Sigma_2^{-1}A + A^T\Sigma_2^{-1}B$$
$$= A^T\Sigma_2^{-1}A + B^T\Sigma_2^{-1}B + 2A^T\Sigma_2^{-1}B^T$$

$$= E_p\left[\frac{1}{2}(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1) + (x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2) + \frac{1}{2}(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

$$= E_p\left[\frac{1}{2}(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)\right] + E_p\left[(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right] + E_p\left[(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

(constant)

$$= E_p\left(\mathrm{tr}\left\{\frac{1}{2}(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)\right\}\right) + \quad + \quad E_p[\text{const}] = \text{const}$$

$$= \mathrm{tr}\left\{\frac{\Sigma_2^{-1}\Sigma_1}{2}\right\} + 0 + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) \quad \text{— (B)}$$

(as did previously)

$$E_p[(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)] = [E_p(x)-E_p(\mu_1)]^T\Sigma_2^{-1}(E_p(\mu_1)-E_p(\mu_2))$$

$$= (\mu_1-\mu_1)\Sigma_2^{-1}(\mu_2-\mu_1) = \boxed{0} \text{ proved}$$

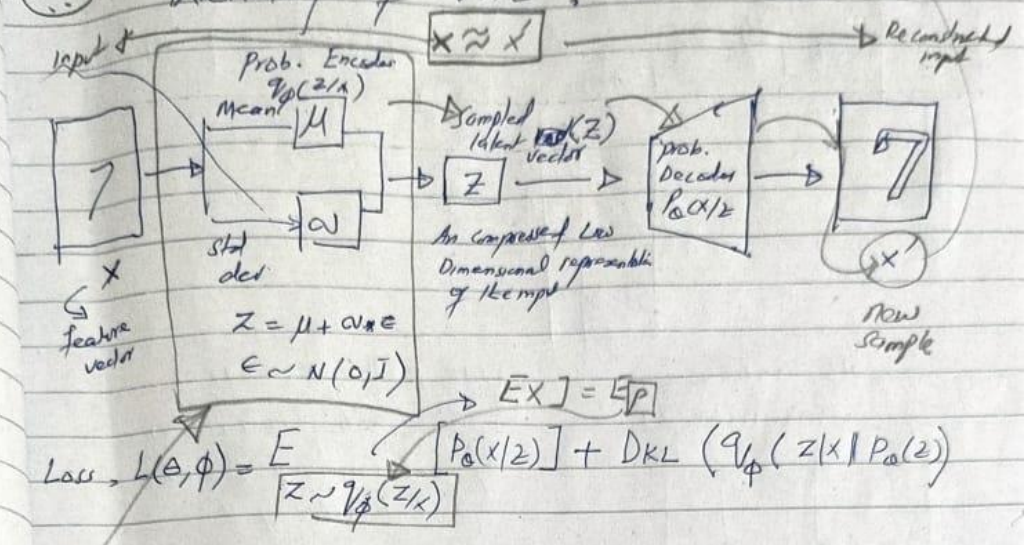
Substitute (A), (B) in 4 we get

$$\boxed{KL(P(x)\|Q(x)) = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - K + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]}$$

- We proved that the two Gaussian having $N(\mu_1, \Sigma_1)$ & $N(\mu_2, \Sigma_2)$ have the KL divergence as shown in previous page.
- It is important expression & we will see it

Dataset → Generative sample

Working of VAE : Top Level view



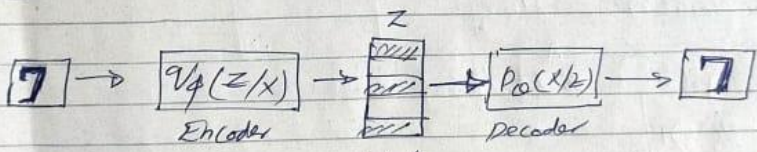
$$E[X] = E_{p}$$

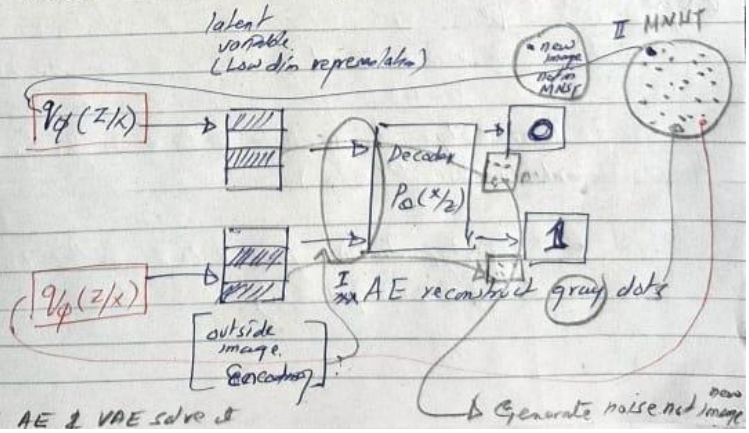
$$\text{Loss}, L(\theta, \phi) = E_{z \sim q_\phi(z/x)}\left[P_\theta(x/z)\right] + D_{KL}\left(q_\phi(z|x) \,\|\, P_\theta(z)\right)$$

why we need VAE

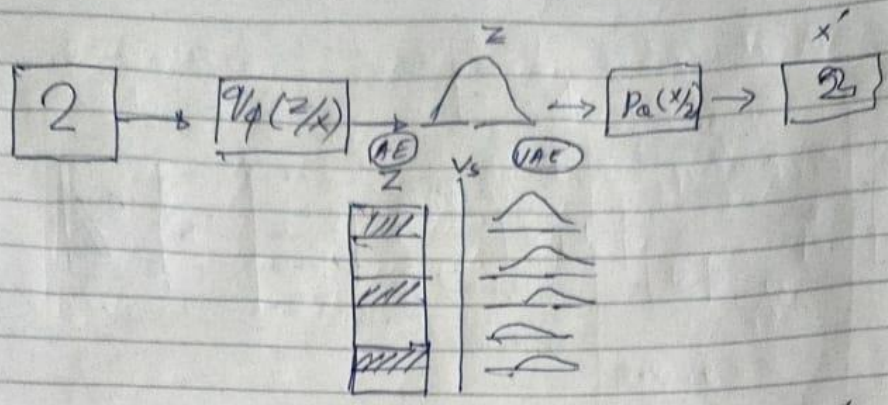
The Goal of VAE: The Goal of VAE is to find a distribution $q_\phi(z/x)$ of some latent variable (z), which can sample from $z \sim q_\phi(z/x)$ to generate new sample $x' \sim P_\theta(x/z)$

Typical AE





It is Problem of AE & VAE solve it

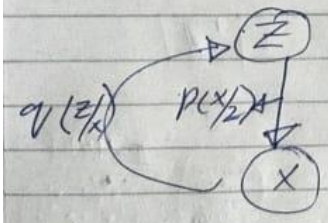


- AE produce prob. vector (z)
- VAE produce the range of value (prob density function) of each encoding of each encoded value of z.
- Now if input image to VAE is different from MNIST it still generate reasonable new image

↳ we can't classify them

LATENT VARIABLE | z ⇒ hidden variable

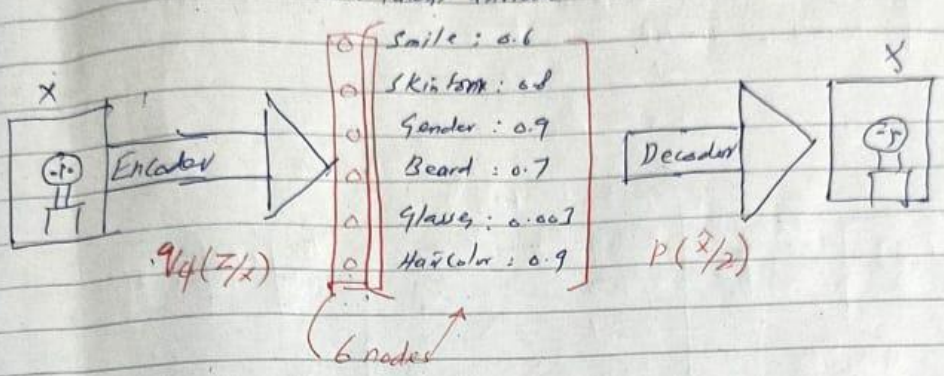
We model system as a collection of random variables. Here



- The edge drawn from z to x is the conditional distribution $P(x/z)$
- The edge drawn from x to z is the variational distribution $q(z/x)$

(8)

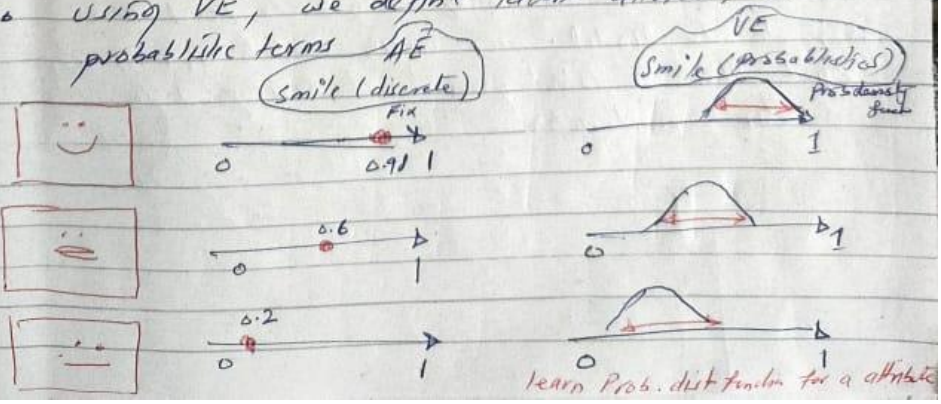
① Latent variables corresponding to a real feature of the object that have not been measured (may be technology is not available to do that)



- In above examples, we trained AE on larger dataset of faces with encoding dimension of 6. An ideal AE will learn the descriptive attributes of faces such as ~~face~~ Smiles, skin tone etc. in order to describe observation in some compressed form.

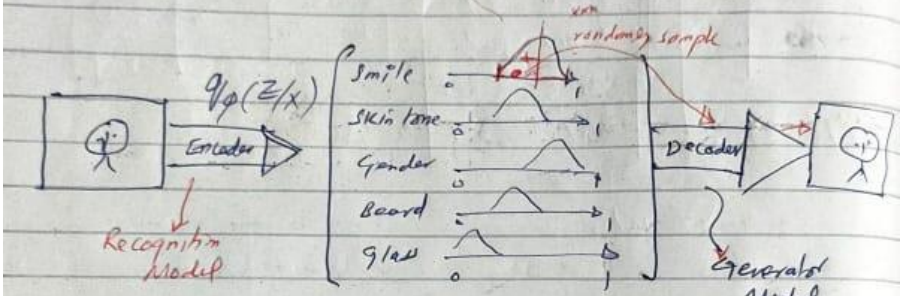
- In above example, we have described the input image in term of latent variables using single value to describe each attributes. For instance, what single value you will assign for photo of smiling?

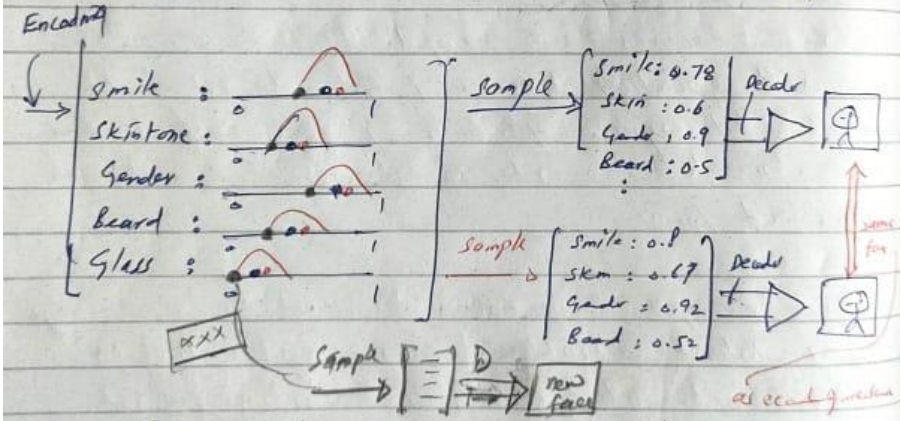
- Using VE, we define latent attributes in probabilistic terms



during Training we use that range of values to generate images so it works

With this approach we now represent each latent attribute for a given input as a probability distribution. When decoding, we will randomly sample from each latent state distribution to generate a vector as input for decoder model.





- By constructing our encoder model to output range of possible values (as statistically distributed) from which we will randomly sample to feed into our decoder model. The values which are near each other in latent space must correspond to similar reconstruction.

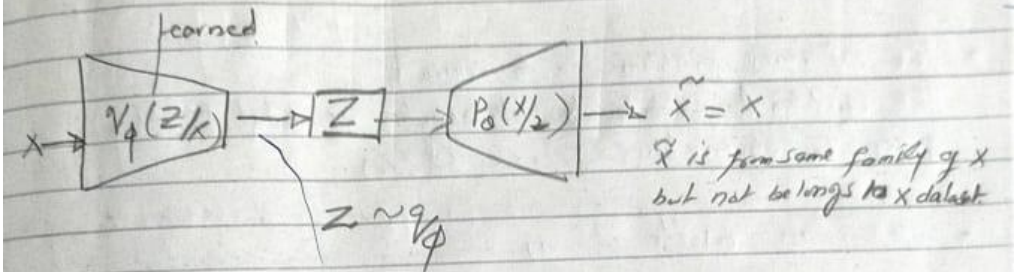
if

Derivation of the Loss Function

Goal

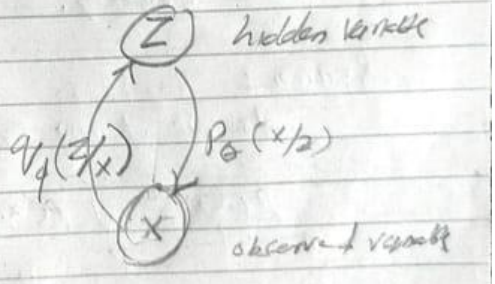
Let's recall the Goal of VAE:

The Goal of VAE is to find a distribution $q_\phi(z/x)$ of some latent variables (z), which we can sample from $z \sim q_\phi(z/x)$ to generate new samples x' from $p_\theta(x/z)$



ϕ, θ are learned through B.P

Graphical Model



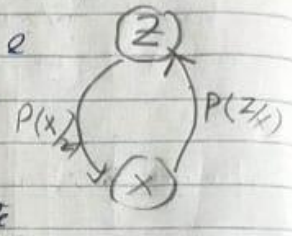
* x is feed to encoder, ,
* we have to learn $q_\phi(z/x)$ & then samp

$z \sim q_\phi(z/x)$. Feed z to $p_\theta(x/z)$ to generate the new sample $(\tilde{x})$

- The Derivation of loss function of VAE is solved through the problem approximate inference.

The Problem Of Approximate Inference :

- let x be a set of observed variables & let z be the set of latent variables with joint distribution $p(z, x)$.



- The inference problem is to compute the conditional ~~probability~~ distribution of latent variables given the observations (x) . ie $p(z/x)$

- we can write it as

$$p(z/x) = \frac{p(x/z)\,p(z)}{p(x)} \quad \text{---(A)}$$

Evaluating (A) is difficult because $p(x)$ cannot be solved properly.

Reason :

As we know by the law of total probability

$$p(x) = \int_z p(x/z)\,p(z)\,dz$$

$$= \int_z p(x, z)\,dz$$

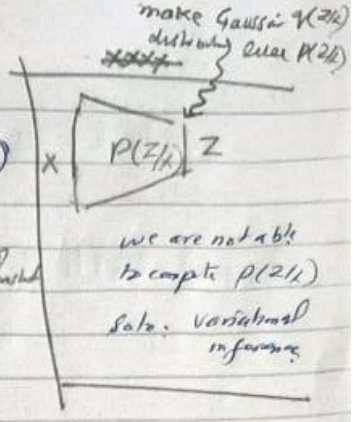
This integral is not available in close form or in intractable (ie requires exponential time to compute) due to multiple integrals involved for latent variable vector z.

$z = \begin{bmatrix} 0 & \text{smile} \\ 0 & \text{skinton} \\ 0 & \text{glasses} \\ 0 & \vdots \\ 0 & \\ 0 & \end{bmatrix}$

Variational Inference (VI)

$$P(z|x) = \frac{P(x|z)P(z)}{P(x)}$$ (posterior, likelihood, prior, normalizing constant)

- (VI) is used to solve the problem $P(x)$ is not tractable
- Approximating $P(z|x)$ with another distribution $Q_\phi(z|x)$ (i.e. $Q_\phi(z|x) \approx P(z|x)$)
  - Prediction



- Goal is to approximate intractable posterior distribution ($P(z|x)$) with simpler distribution ($Q_\phi(z|x)$), such as Gaussian distribution

By choosing tractable distribution (like gaussian distribution), let say $Q(z|x)$ and play with the parameters (mean, variance) of $Q(z|x)$ distribution it become closer to $P(z|x)$

IDEA: we make $Q_\phi(z|x) \approx P(z|x)$, by minimizing $KL(Q\|P)$ [as close as possible]

as $KL(Q\|P)$ minimization make $Q \approx P$.

Q_ϕ: predict, P_θ: g.t

$$D_{KL}(Q_\phi(z|x) \| P_\theta(z|x)) = \sum_z Q_\phi(z|x) \log \frac{Q_\phi(z|x)}{P_\theta(z|x)}$$

$$= E_{z \sim Q_\phi(z|x)}\left[\log \frac{Q_\phi(z|x)}{P_\theta(z|x)}\right]$$

$$= E_{z \sim Q_\phi(z|x)}\left[\log(Q_\phi(z|x)) - \log(P_\theta(z|x))\right]$$

$$= E_{z \sim Q_\phi(z|x)}\left[\log Q_\phi(z|x)\right] - E_{z \sim Q_\phi(z|x)}\left[\log(P_\theta(z|x))\right.$$

(27) $z \sim z \sim Q_\phi(z/x)$

Substituting (A) in (B)

$$D_{KL}(Q_\phi(z/x) \| P_\theta(z/x)) = E_z[\log Q_\phi(z/x)] - E_z\left[\log \frac{P_\theta(x/z)\,P(z)}{P(x)}\right]$$

(Bayes' Rule)

$$= E_z[\log Q_\phi(z/x)] - E_z\left[\log \frac{P(z,x)}{P(x)}\right]$$

$$= E_z[\log Q_\phi(z/x)] - E_z[\log P(z,x)] + E_z[\log P(x)]$$

$$= E_z[\log Q_\phi(z/x)] - E_z[\log P(z,x)] + \int_z Q_\phi(z/x) \log P(x)\,dz$$

$$= E_z[\log Q_\phi(z/x)] - E_z[\log P(z,x)] + \log P_\theta(x) \int_z Q_\phi(z/x)\,dz$$

$$= E_z[\log Q_\phi(z/x)] - E_z[\log P_\theta(z,x)] + \log P_\theta(x)$$

$$\log P_\theta(x) = \underbrace{E_z[\log P(z,x)] - E_z[\log Q_\phi(z/x)]}_{(2)} + D_{KL}(Q_\phi(z/x) \| P_\theta(z/x))$$

Fixed

θ

Component 1

max

Component 2

min

we want to min. it

- In order to make the above equation equally true if we minimized component (2) we have to maximized the component 1 (2)
- Component 1 (2) is called Evidence lower bound (ELBO)
- Now if we maximized ELBO, it indirectly minimized the KL divergence.

- Why $\mathcal{L}$ is called lower bound

As $KL \geqslant 0$ (it is property of KL)

$$\therefore \log P_\theta(x) = \mathcal{L} + \underbrace{(+ve\ no \geqslant 0)}_{KL}$$

$$\mathcal{L} = \log P_\theta(x) - KL$$

- If $KL = 0$

$$\mathcal{L} = \log P_\theta(x)$$

- If $KL > 0$

$$\mathcal{L} < \log P_\theta(x)$$

$$\therefore \mathcal{L} \leq \log P_\theta(x)$$

so it is lower bound of $\log P_\theta(x)$

NOW As

$$\mathcal{L} = ELBO = E_z[\log P(Z, X)] - E_z[\log(Q_\phi(z/x))]$$

$$= E_z[\log P(x/z) P_\theta(z)] - E_z[\log Q_\phi(z/x)]$$

$$= E_z[\log P(x/z)] + E_z[\log(P_\theta(z))] - E_z[\log Q_\phi(z/x)]$$

$$= E_z[\log P(x/z)] - E_z\left[\log \frac{Q_\phi(z/x)}{P_\theta(z)}\right]$$

$$\mathcal{L} = E_z[\log P(x/z)] - D_{KL}[Q_\phi(z/x) \,\|\, P_\theta(z)]$$

← expected reconstruction error (first term); learn distribution ≈ prior distribution (second term)

[if z is gaussian, then reconstruction is square error]

$$P(x/z) = e^{-|x-x|^2}$$

$$\log P(x/z) = |x - x|^2$$

if z is bernoulli, then reconstruction error is cross entropy

Expectation of random variable X (E(X))

Expectation

Random variable is a function that assign real no. to the sample space of experiment

eg

Experiment

- Pick Red Ball score is 10
- pick Blue ball score is 20
- pick Green ball score is 40

$U = \{Red, Blue, Green\}$

X = Score of pick

X(Red) = 10

X(Blue) = 20

X(Green) = 40

$X = \{10, 20, 40\}$

Random variable

RV
- ① Discrete (Countable) — use (PMF)
- ② Continuous — use (PDF)

prob distribution is a table that show possible values of a variable & how often it occurs