Introduction to Variational AutoEncoder (VAE)

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

SEOULTECH 서울과학기술대학교 데이터사이언스학과

1. I. Information Theory

2. II. AutoEncoder

3. III. VAE

4. IV. Practice

I. Information Theory

Entropy

Definition

- [Extensional definition] A state function used to describe the flow of unuseful energy
  - “unuseful energy” means it cannot be converted to work
  - According to the 2nd law of thermal dynamics, entropy
- [Statistical definition] Obtaining the expected value of a log index that reflects how important an event is in terms of information.
  - (Info theory) An informational importance of an event occurred increase over time
- if a coin flip is head, is this important info?
- if an earthquake occurs, is this important info?

$$H(X) = \mathbb{E}[log\frac{1}{P(x)}]$$

For a Discrete (r.v X) : w$/pmf \quad p(x) = \mathbb{P}(X = x)$

$$\begin{aligned} H(X) &= \mathbb{E}[log\frac{1}{P(x)}] \\ &= \sum p(x) log\frac{1}{P(x)} \end{aligned}$$

For a Continuous (r.v X): w$/pdf \quad f(x)$

$$\begin{aligned} H(X) &= \mathbb{E}[log\frac{1}{P(x)}] \\ &= \int f(x) log\frac{1}{f(x)} \; dx \end{aligned}$$

- Example
  - A trial of randomly picking a fruit from a pocket
  1. pocket: Banana (0.5), Strawberry (0.3), Grape (0.2)
  - $H(x) = 0.5 \times log(0.5)^{-1} + 0.3 \times log(0.3)^{-1} + 0.2 \times log(0.2)^{-1}$
  - $\quad = 0.5 \times 0.3010 + 0.3 \times 0.5229 + 0.2 \times 0.6990$
  - $\quad = 0.44717$

  2. pocket: Banana($\frac{1}{3}$), strawberry($\frac{1}{3}$), Grape($\frac{1}{3}$)
  - $H(x) = \frac{1}{3} \times log(\frac{1}{3})^{-1} + \frac{1}{3} \times log(\frac{1}{3})^{-1} + \frac{1}{3} \times log(\frac{1}{3})^{-1}$
  - $\quad = \frac{1}{3} \times 0.4771 + \frac{1}{3} \times 0.4771 + \frac{1}{3} \times 0.4771$
  - $\quad = 0.4771$

- Example
  - A trial of randomly picking a ball from a pocket (100 balls in a pocket)
  1. pocket: red(1), black(99)
  2. pocket: red(50), black(50)

  - Intuitively, you might think that the probability of getting black in the first case is high, so there is less uncertainty. Actually calculating entropy gives the latter much larger.

  1. $H(x) = 0.01 \times log(0.01)^{-1} + 0.99 \times log(0.99)^{-1}$
  - $= 0.01 \times 2 + 0.99 \times 0.004$
  - $= 0.02396$
  2. $H(x) = 0.5 \times log(0.5)^{-1} + 0.5 \times log(0.5)^{-1}$
  - $= 0.5 \times 0.3010 + 0.5 \times 0.3010$
  - $= 0.3010$
- As a result, case 2 have bigger entropy than case 1.

Cross Entropy

Definition 1

Definition : It shows the usefulness of obtaining information according to the predicted distribution by using the feature that the actual distribution and the predicted distribution($Q(x)$) are different after assuming the actual distribution($P(x)$).

$$\begin{aligned} H(P,Q) &= \sum_{x \in \mathbf{x}} P(x) log(\frac{1}{Q(x)}) \\ &= -\sum_{x \in \mathbf{x}} P(x) log Q(x) \\ &= -\frac{1}{n} \sum_{x} P(x) log Q(x) + (1 - P(x)) log(1 - Q(x)) \end{aligned}$$

Cross Entropy

$$H(P,Q) = -\sum_{i=1}^{n} P(x_i) log\ Q(x_i)$$

- Cross entropy is to predict q through p, which is a distribution obtained through modeling, without knowing about the actual distribution q. **Since both q and p are used, it is called cross entropy**.
- In the case of machine learning, there are cases where both the value and q of the real environment are known and the predicted value (observed value) p. **It is used when the machine learning model predicts with a certain percentage of probability, but knows that the actual probability is what percent!**
- In cross entropy, when the actual value and the predicted value match, the value converges to 0, and when the value is different, the value increases. Therefore, it can be called **entropy to reduce the difference between the actual value and the predicted value**.

Cross Entropy proof

$$-\sum_{x}^{N} p(x) log\ q(x) = -(p(x_1) log\ q(x_1) + p(x_2) log\ q(x_2) + ... + p(x_n) log\ q(x_n))$$

$$= -\frac{1}{N}[(p(x_1) log\ q(x_1) + \underbrace{(1 - p(x_1)) log(1 - q(x_1)))}_{= p(x_2) log\ q(x_2) + ... + p(x_n) log\ q(x_n)}$$
$$+ (p(x_2) log\ q(x_2) + (1 - p(x_2)) log(1 - q(x_2))) + ...]$$

$$= -\frac{1}{N} \sum_{x}^{N} p(x) log\ q(x) + (1 - p(x)) log(1 - q(x))$$

For example sample number is 3.

$$\begin{aligned}
-\sum_{x}^{3} p(x) log\ q(x) &= -(p(x_1) log\ q(x_1) + p(x_2) log\ q(x_2) + p(x_3) log\ q(x_3)) \\
&= -\frac{1}{3}[(p(x_1) log\ q(x_1) + (1-p(x_1)) log(1-q(x_1))) \\
&\quad + (p(x_2) log\ q(x_2) + (1-p(x_2)) log(1-q(x_2))) \\
&\quad + (p(x_3) log\ q(x_3) + (1-p(x_3)) log(1-q(x_3)))] \\
&= -\frac{1}{3}\sum_{x}^{3} p(x) log\ q(x) + (1-p(x)) log(1-q(x))
\end{aligned}$$

KL Divergence

1. Definition of Kullback-Leibler(KL) Divergence

Definition 2

A function used to calculate the difference between two probability distributions. For an ideal distribution, another distribution that approximates the distribution is used. Calculate the information entropy difference that can occur if you sample.

- Discrete : $D_{KL}(P||Q) = \mathbb{E}[log(\frac{P(X)}{Q(X)})] = \sum_{x \in \mathbf{x}} P(x)log(\frac{P(x)}{Q(x)})$
- Continuous: $D_{KL}(P||Q) = \int_{-\infty}^{\infty} P(x)log(\frac{P(x)}{Q(x)})$

KL Divergence

2. Relationship between KL Deivergence nad Entropy

For discrete variables, KL-Divergence can be expressed as the difference between the cross entropy of another probability distribution (Q) for a specific probability distribution (P) and the entropy of a specific probability distribution (P).

$$D_{KL}(P||Q) = -\sum_{x \in \mathbf{x}} p(x)log(q(x)) - [-\sum_{x \in \mathbf{x}} p(x)log(p(x))]$$

$$= H(P,Q) - H(P)$$

$$= -\sum P(x)log(\frac{Q(x)}{P(x)})$$

Distance measure

1. $d(X, X) = 0$
2. $dist(X, Y) = d(Y, X)$
3. $dist(X, Y) \leq d(X, Z) + d(Z, Y)$

Characteristic of KL Divergence

1. Does not have negative values

$$D_{KL}(P||Q) \geq 0$$

2. 'The value of KL Divergence is 0' and 'probability distribution P and Q are the same' are equivalent.

$$D_{KL}(P||Q) = 0 \Leftrightarrow P(x) = Q(x), \forall_x \in \mathbf{x}$$

3. In general, the value of KL Divergence of two distributions varies depending on which distribution is used (asymmetry).

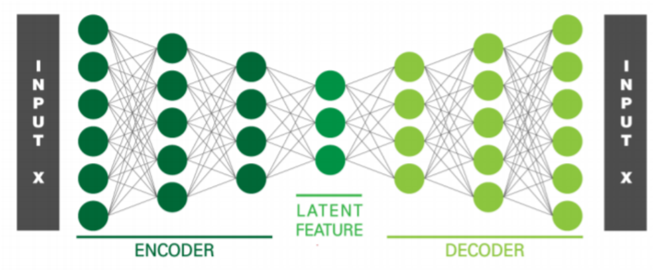
$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

We can know that follow things from previous proof and theorem:

- $D_{KL}(P||Q) = H(P,Q) - H(P)$
- $D_{KL}(P||P) = 0$
- $\sum P(x)log(\frac{P(x)}{Q(x)}) \neq \sum Q(x)log(\frac{Q(x)}{P(x)})$

II. AutoEncoder

Definition and Structure of AutoEncoder



- **ENCODER**
  - input $\xrightarrow{\text{encoder}}$ feature value
- **DECODER**
  - input $\xleftarrow{\text{decoder}}$ feature value
- **LATENT FEATURE**
  - feature value

Representation Learning / Feature Learning

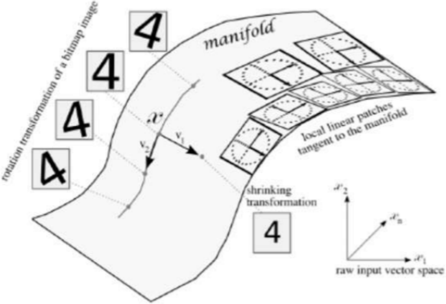
- It refers to the totality of systematic techniques for automatically discovering feature points necessary for feature detection or classification.

Taxonomy of Representation Learning

- Unsupervised Learning
    - K-means Clustering
    - Principal Component Analysis(PCA)
    - Local Linear Enbedding
    - Independent Component Analysis
    - Unsupervised Dictionary Learning
- Supervised Learning
    - Supervised Dictionary Learning
    - Neural Networks
- multi-layered learning
    - Restricted Boltzman Machine
    - **Autoencoder**

Manifold Hypothesis

- It is assumed that high-dimensional data can be represented in low-dimensional manifolds.
- If you project to a manifold how much it has rotated or how much it has been reduced or enlarged for a single shape, you can represent the values of the discussed elements in a **two-dimensional** form.



III. VAE

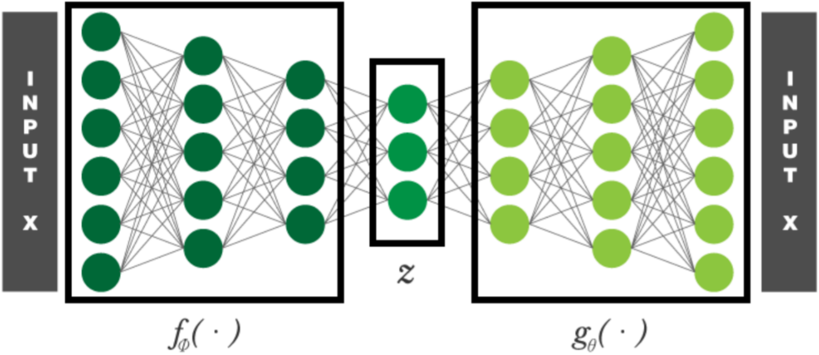
Definition of AutoEncoder

- It refers to a generative neural network that implements variational inference through the structure of an auto-encoder.

- **what is variational inference?**

- It refers to approximating the posterior probability distribution to a simpler probability distribution.

Stucture of VAE



$$x = g_\theta(z)$$

$$p(x) = p(x|z)$$

$$p(x) = E_{z-p_\theta(z)}[p(x|z)$$

$$p(x) = E_{z-p_\theta(z|x)}[p(x|z)] \approx E_{z-q_\Phi(z|x)}[p(x|z)]$$

$$
\begin{aligned}
log(p(x)) = log(p(x)) * 1 = log(p(x)) * \int_z q_\Phi(z|x)\ dx \\
&= \int_z log(p(x)) q_\phi(z|x)\ dz \qquad [p(x) = \frac{p(x,z)}{p(z|x)}] \\
&= \int_z log(\frac{p(z,x)}{p(z|x)}) q_\phi(z|x)\ dz = \int_z log(\frac{p(z,x)}{q_\phi(z|x)} \frac{q_\phi(z|x)}{p(z|x)}) q_\phi(z|x)\ dz \\
&= \int_z log(\frac{p(z,x)}{p(z|x)}) q_\phi(z|x)\ dz + \int_z log(\frac{q_\phi(x|z)}{p(z|x)}) q_\phi(z|x)\ dz \\
&(Let, q_\phi(x|z) = a, p(z|x) = b) \\
&\int_z log(\frac{q_\phi(x|z)}{p(z|x)}) q_\phi(z|x)\ dz = \int a * log(\frac{a}{b}) \\
&\int_z log(\frac{q_\phi(x|z)}{p(z|x)}) q_\phi(z|x)\ dz = D_{KL}(q_\phi(z|x)||p(z|x)) \\
&= \underbrace{\int_z log(\frac{p(z,x)}{p(z|x)}) q_\phi(z|x)\ dz}_{ELBO; Evidence Lower Bound} + D_{KL}(q_\phi(z|x)||p(z|x))
\end{aligned}
$$

$$\underbrace{\int_z log(\frac{p(z,x)}{p(z|x)})q_\phi(z|x)\ dz}_{ELBO;Evidence Lower Bound} + D_{KL}(q_\phi(z|x)||p(z|x))$$

ELBO

$$\begin{aligned}ELBO &= \int_z log(\frac{p(z,x)}{p(z|x)})q_\phi(z|x)\ dz = \int_z log(\frac{p(z)p(x|z)}{q_\phi(z|x)})q_\phi(z|x)\ dz \\ &= \int_z log(p(x|z))q_\phi(z|x)\ dz - \int_z log(\frac{q_{phi}(z|x)}{p(z)})q_\phi(z|x)\ dz \\ &= E_{q_\phi(z|x)}[log(p(x|z))] - D_{KL}(q_\phi(z|x)||p(z))\end{aligned}$$

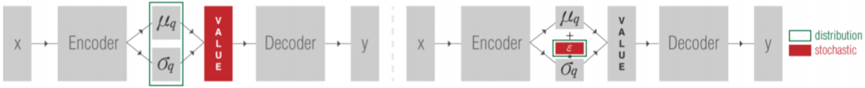
ELBO : $-E_{q_\phi(z|x)}[log(p(x|z))]$

$-E_{q_\phi(z|x)}[log(p(x|z))] = -\int q_\phi(z|x)log(p(x|z)) \approx \underbrace{-\frac{1}{L}\sum_1^L log(p(x|z_1))}_{\text{Monte Carlo Estimation}} \, where \; z_1 \sim q_\phi(z|x)$

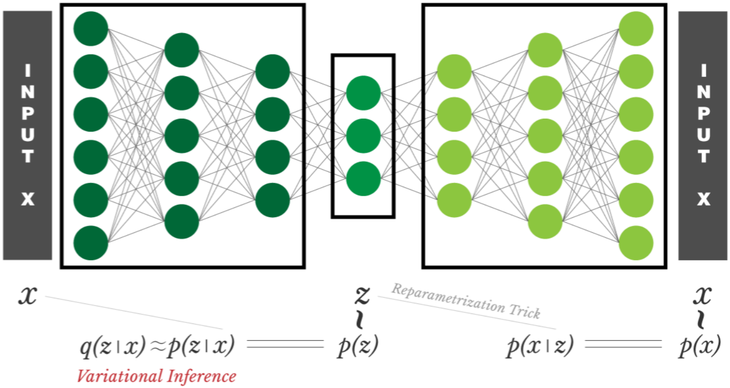
But when using the MonteCarlo technique, we can't differentiate so that it is stochastic .

For that reason, we use Reparametrization Trick; 'z' is replaced with '$\mu_q + \sigma_q\epsilon$'

$$-E_{q_\phi(z|x)}[log(p(x|z))] = -E_{\epsilon\sim N(0,1)}[log(p(x|\mu_q + \sigma_q\epsilon))]$$



Summary


INPUT X
INPUT X
x
z
$\wr$
Reparametrization Trick
x
$\wr$
$q(z|x) \approx p(z|x)$
$p(z)$
$p(x|z)$
$p(x)$
Variational Inference

Summary

1. What creates image X well
2. X is more likely to appear
3. Neural networks produce results based on Z-space
4. Assume that the Z-space is uniformly distributed.
5. **Inferring that constant distribution as a simpler distribution** (Variational Inference)

IV. Practice

IV. Practice

kolab

$\cdots\cdots.$

```
"Success isn't permarnent, and failure isn't fatal. - Mike Ditka"
```