Deep Learning

Information Theory

Kyungwoo Song

Information Theory

- Discrete random variable x에 대해서 생각해보겠습니다.
  - x의 정보량, (amount of information, degree of surprise)는 어떻게 측정할 수 있을까요?
  - 이러한 정보량을 $h(x)$라고 하겠습니다.
- $p(x = a)$가 낮은데, 우리가 $x = a$를 관측했다고 하면, 정보량이 크다고 할 수 있습니다.
  - 예를 들어, 항상 100점 맞는 친구가, 또 100점을 맞는것을 보는것과, 99점을 맞는 것을 보는 것은 다른 정보량이겠죠?
  - **1) 그 까닭에, $h(x)$는 $p(x)$에 영향을 받을 수 밖에 없습니다.**
- Independent random variable x와 y에 대해서 생각해보겠습니다.
  - $p(x, y) = p(x)p(y)$
  - **2) 정보량의 경우, $h(x, y) = h(x) + h(y)$**
    - 각각이 독립이니, 정보량 또한 각각 더해주면 됩니다.
- 이러한 1번과 2번을 모두 만족하는 식은 무엇이 있을까요?

Source: Pattern Recognition and Machine Learning

- **1) 그 까닭에, $h(x)$는 $p(x)$에 영향을 받을 수 밖에 없습니다.**
- **2) 정보량의 경우, $h(x, y) = h(x) + h(y)$**
- $\Rightarrow h(x) = -\log_2 p(x)$
  - Information은 0또는 양수가 됩니다.
  - $p(x)$가 낮을 수록, 높은 정보량에 해당됩니다.
  - 밑이 반드시 2일 필요는 없습니다. (밑 변환을 할 경우, rescale 여부)
  - Self-information 이라고도 부릅니다.

밑이 2일 때,
bit 라고도 부릅니다.

- 이러한 정보량의 기댓값을 계산해볼까요?
  - $H(x) = -\sum_x p(x) \log_2 p(x)$
  - **(Shannon) Entropy** of the random variable x

NOTE)
$\lim_{p \to 0^+} p \ln p = 0$
We shall take $p(x) \ln p(x) = 0$ whenever we encounter a value for x such that $p(x) = 0$

- Example
  - 만약 x가 가질 수 있는 값이 $\{a, b, c, d, e, f, g, h\}$이고, 각각의 확률값이 $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$라고 한다면, Entropy 는?
  - 2

Source: Pattern Recognition and Machine Learning

- Consider a r.v. x having 8 possible states, each of which is equally likely
  - We would need to transmit a message of length 3 bits
  - Entropy: $H(x) = -8 \times \frac{1}{8}\log_2 \frac{1}{8} = 3$
- Consider a r.v. x having 8 possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective probabilities are given by $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$
  - Entropy: $H(x) = 2$
  - Non-uniform distribution has a smaller entropy than the uniform one
  - Example) 0, 10, 110, 1110, 111100, 111101, 111110, 111111
    - ❖The average length of the code: $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2$ bits
    - ❖0, 10, 01, 11, … 식으로 하면 안될까요? ⇒ 0110 이 ada 인지, cb 인지 알 수 없음
- Noiseless coding theorem
  - The entropy is a lower bound on the number of bits needed to transmit the state of a r.v.

Source: https://www.princeton.edu/~cuff/ele201/kulkarni_text/information.pdf

Entropy

Information Theory

- $H(x) = -\sum_x p(x) \log p(x)$
  - $0 \le p(x) \le 1 \Rightarrow H(x) \ge 0$
  - It will equal its minimum value of 0 when one of the $p_i = 1$ and all other $p_{j \ne i} = 0$
- The maximum entropy can be found by maximizing $H(x)$ using a Lagrange multiplier
  - $\tilde{H} = -\sum_i p(x_i) \log p(x_i) + \lambda(\sum_i p(x_i) - 1)$
  - $\frac{\partial \tilde{H}}{\partial p(x_i)} = -\log p(x_i) - 1 + \lambda = 0$
  - $\frac{\partial \tilde{H}}{\partial \lambda} = \sum_i p(x_i) - 1 = 0$
  - $\Rightarrow p(x_i) = \exp(-1 + \lambda)$ where $\sum_i p(x_i) = 1$
  - $\Rightarrow p(x_i) = \frac{1}{M}$ where M is the total number of states x_i
  - Maximum value: $\log \mathrm{M}$
    - We need to check the second derivative of the entropy
    - $\frac{\partial \tilde{H}}{\partial p(x_i) p(x_j)} = -I_{ij} \frac{1}{p_i}$ where I_{ij} are the elements of the identity matrix

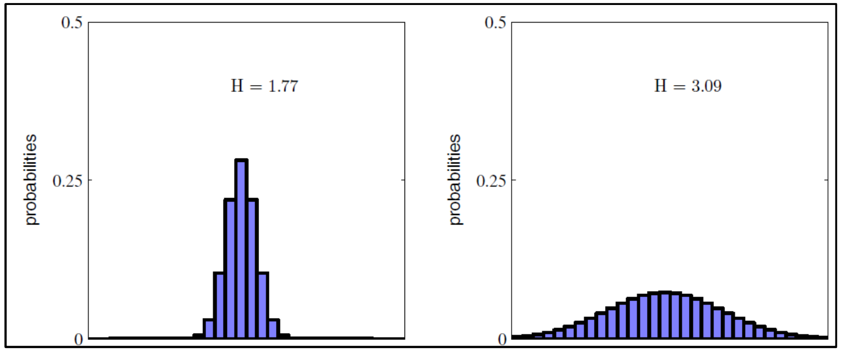
Lagrange Multiplier
Maximize $f(x,y)$ s.t. $g(x,y) = 0$
Lagrange function: $L(x,y,\lambda) = f(x,y) - \lambda g(x,y)$
Solve $\nabla_{x,y,\lambda} L(x,y,\lambda) = 0$

Source:

Amount of Information

Information Theory

- Entropy를 바라보는 다른 view도 존재합니다.
  - N개의 동일한 사물이 있다고 가정하겠습니다.
  - 그리고, i번째 bin에 n_i개를 넣는다고 가정하겠습니다.
  - 그럼 총 가능한 가지수는 $W = \frac{N!}{\prod_i n_i!}$
  - $H = \frac{1}{N}\ln W = \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i!$
  - 만약 $N \to \infty$라고 한다면, $\ln N! \approx N \ln N - N$
    - ❖Stirling's approximation
  - $H = -\lim_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln\left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$



- 이러한 Entropy가 언제 최솟값을 가질까요?
  - $p_i = 1$이며, $p_{j\neq i} = 0$
- 그렇다면, 이러한 Entropy는 언제 최댓값을 가질까요?
  - $p_i = \frac{1}{M}$ (M: bin 개수)

Source: Pattern Recognition and Machine Learning, https://math.stackexchange.com/questions/3448564/derivation-of-information-entropy-using-stirlings-approximation

Cross Entropy

Information Theory

- $h(x) = -\log_2 p(x)$
  - x의 정보량, (amount of information, degree of surprise)
- 이러한 정보량의 기댓값을 계산해볼까요?
  - $H(x) = -\sum_x p(x) \log_2 p(x)$
  - $H = -\lim\limits_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$
  - **Entropy** of the random variable x
- Cross-Entropy
  - $H(p,q) = -E_p[\log q]$
  - $H(p,q) = -\sum_x p(x) \log q(x)$
- Supervised Learning 기억나시나요? Classification도 기억나시나요?
  - y label이 주어지고, 우리 모델의 output 이 label 과 같아지도록 학습
  - $p(x)$: y label (one-hot encoding)
  - $q(x)$: 우리 모델의 output 값

Source:

Continuous Variable Entropy

- 이러한 Entropy는 continuous random variable 에 대해서도 정의할 수 있습니다.
  - $H(x) = -\int p(x) \ln p(x)\, dx$ (differential entropy)
- Example) Gaussian distribution: $p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
  - $\ln p(x) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$
  - $p(x) \ln p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \left(-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}\right)$
  - $-\int p(x) \ln p(x)\, dx = -\left\{\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \left(-\frac{1}{2} \ln 2\pi\sigma^2\right) dx - \int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx\right\}$
  - $\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \left(-\frac{1}{2} \ln 2\pi\sigma^2\right) dx = -\frac{1}{2} \ln 2\pi\sigma^2$
  - $\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = -\frac{1}{2\sigma^2} \int (x-\mu)^2 \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = -\frac{1}{2}$

Source:

Continuous Variable Entropy

Information Theory

- Example) Gaussian distribution: $p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
  - $\ln p(x) = -\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$
  - $p(x)\ln p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\left(-\frac{1}{2}\ln 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}\right)$
  - $-\int p(x)\ln p(x)\,dx = -\left\{\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\left(-\frac{1}{2}\ln 2\pi\sigma^2\right)dx - \int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx\right\}$
  - $\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\left(-\frac{1}{2}\ln 2\pi\sigma^2\right)dx = -\frac{1}{2}\ln 2\pi\sigma^2$
  - $\int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx = -\frac{1}{2\sigma^2}\int (x-\mu)^2 \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx = -\frac{1}{2}$
  - $\Rightarrow -\int p(x)\ln p(x)\,dx = \frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2}$
- 해석을 해보자면…
  - σ^2이 커질수록 (broader), entropy는 더욱 커지는 구조입니다.
  - + entropy가 음수도 될 수 있습니다. ($\sigma^2 < \frac{1}{2\pi e}$)

익숙한 형태죠?
Variance!
$E[(x-\mu)^2] = \sigma^2$

Source:

Continuous Variable Entropy

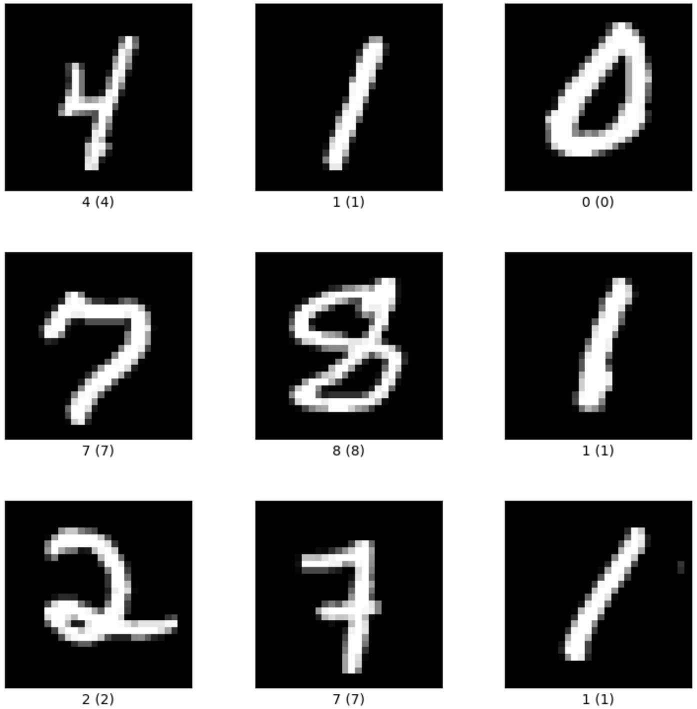
Information Theory

- Maximize the differential entropy with the three constraints
  - $H(x) = -\int p(x) \ln p(x)\, dx$
    - $\int_{\infty}^{-\infty} p(x) dx = 1$
    - $\int_{\infty}^{-\infty} x p(x) dx = \mu$
    - $\int_{\infty}^{-\infty} (x-\mu)^2 p(x) dx = \sigma^2$
  - Lagrange multipliers + calculus of variations,
    - $p(x) = \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\}$
    - $p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
  - Entropy: $\frac{1}{2} \ln 2\pi\sigma^2 + \frac{1}{2}$
    - It can be negative

Source:

이러한 Entropy가 어디에 쓰일까요?

- MNIST Dataset 기억나시나요?
- MNIST: handwritten digits
  - Training set: 60,000 examples
  - Test set: 10,000 examples
  - # class: 10
    - ❖0,1,2,3,4,5,6,7,8,9



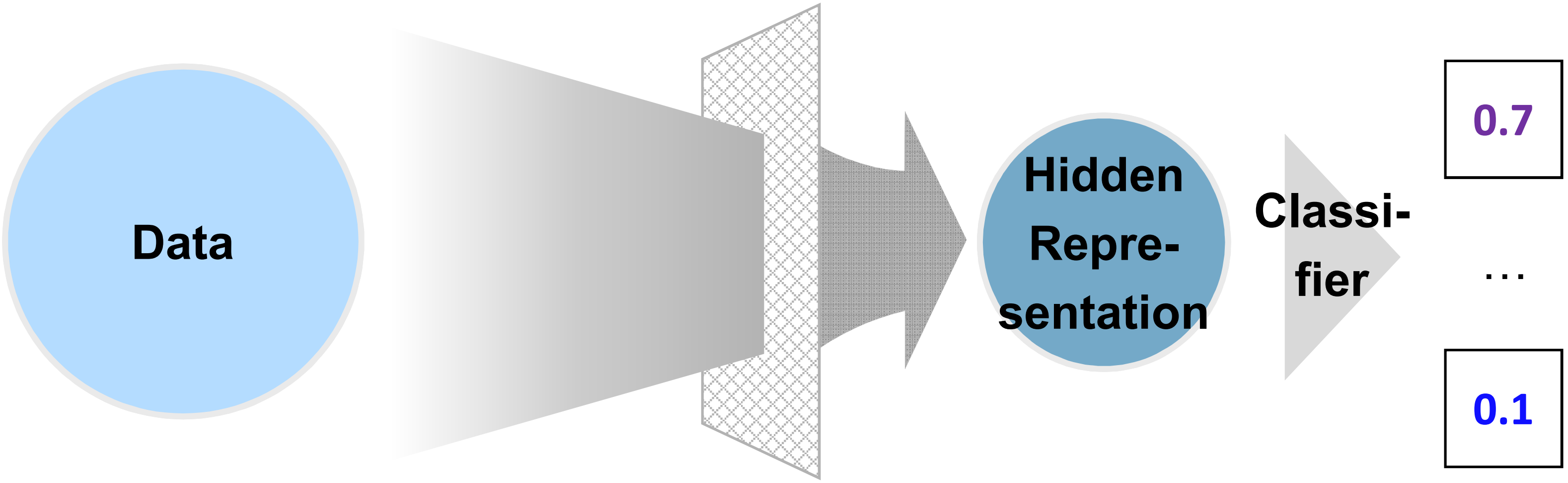
- Very easy dataset
  - Training Accuracy: 100%
  - Test Accuracy: 99% ↑

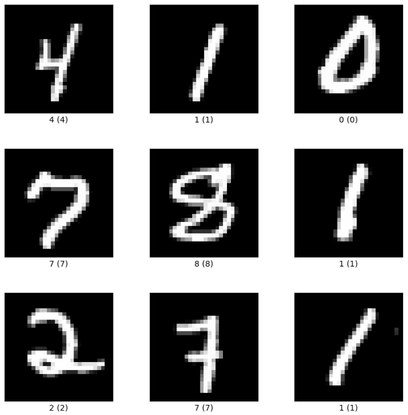
- When we develop a new model,
  - Fail on MNIST → Fail on other datasets (T.T)
  - Success on MNIST → ??? on other datasets (T.T)

Source:

이러한 Entropy가 어디에 쓰일까요?

Information Theory





- Neural Network captures meaningful representation
- Classifier outputs the probability (?) of each class
  - The probability of class 0: **0.7**
  - …
  - The probability of class 9: **0.1**
- Model **confidence** can be measured by
  - $max_i p_i$ for $i = 0, \ldots, 9$
  - **Negative Entropy**

Source:

Conditional Entropy

Information Theory

- 다시 돌아와서…
- Conditional Entropy 도 정의할 수 있습니다.
  - 즉, x 에 대한 것은 이미 알고 있을 때, y에 대한 entropy 입니다.
    - ❖정보량은 $-\ln p(y|x)$ 입니다.
    - ❖그렇다면, 평균적인 정보량을 나타내는 entropy $H(y|x)$의 경우에는 어떻게 될까요?
  - $H(y|x) = -\int\int p(y,x)\ln p(y|x)\ dydx$
    - ❖NOTE) $H(x) = -\int p(x)\ln p(x)\,dx$
    - ❖Conditional entropy of y given x
  - $H(x,y) = H(y|x) + H(x)$

$H(x,y)$: entropy of $p(x,y)$
$H(x)$: entropy of $p(x)$
즉, x 와 y를 표현하기 위해 필요한 정보량은, x를 표현하고, y given x를 표현 하는 정보량

❖$H(y|x) = -\int\int p(y,x)\ln p(y|x)\ dydx = -\int\int p(y,x)\ln\frac{p(x,y)}{p(x)}\ dydx$

$$= -\int\int p(y,x)\ln p(x,y)\ dydx + \int\int p(y,x)\ln p(x)\ dydx$$

$$= -\int\int p(y,x)\ln p(x,y)\ dydx + \int\int p(x)\ln p(x)\ dydx$$

$$= H(x,y) - H(x)$$

Source:

KL Divergence

Information Theory

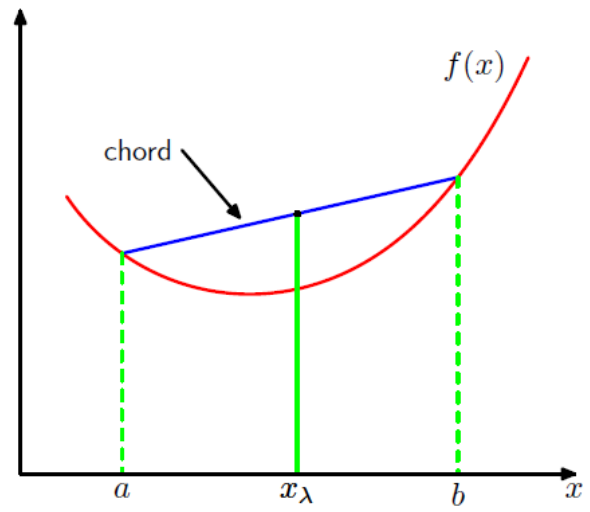
- 지금까지 우리는 정보량과 엔트로피에 대해서 살펴보았습니다.
- 이번에는 relative entropy, 또는 Kullback-Leibler (KL) divergence 에 대해서 살펴보겠습니다.
  - 우리는 unknown distribution $p(x)$에 관심이 많습니다.
  - 하지만, 모르기에, $q(x)$ 로 대신 모델링하고자 합니다.
  - 그때, 평균적으로 더 필요한 정보량은 어떻게 될까요?
  - $KL(p||q) = -\int p(x) \ln q(x)\, dx - \left(-\int p(x) \ln p(x)\, dx\right)$

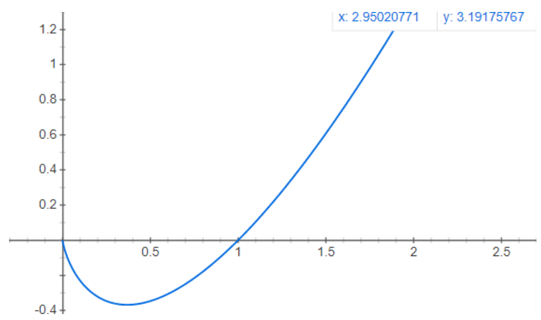
    $= -\int p(x) \ln \left\{\frac{q(x)}{p(x)}\right\} dx$
- KL divergence
  - 먼저, 등호가 성립할 조건은, $q(x)$ 와 $p(x)$ 가 같을 때 입니다.
  - $KL(p||q) \neq KL(q||p)$
  - KL divergence의 값은 0이상입니다.
  - 왜 그럴까요?

Source:

Convex

- $KL(p||q) = -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}dx \geq 0$을 위해서, Convex를 살펴보겠습니다.
- **Convexity**: $f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b)$ for $0 \leq \lambda \leq 1$
  - $\lambda f(a) + (1-\lambda)f(b)$: 현이라고 볼 수 있겠죠?
  - Example) $x^2, x\ln x\ (x > 0), -\ln x$
- NOTE)
  - $\lambda = 0,1$ 일때만 등호가 성립할 경우, strictly convex
  - 위와 반대되는 부등호를 가질 경우, concave
  - 만약 $f(x)$가 convex 일 경우, $-f(x)$는 concave
  - 다음과 같이 확장도 가능합니다.
    - $f\left(\sum_{i=1}^{M}\lambda_i x_i\right) \leq \sum_{i=1}^{M}\lambda_i f(x_i)$
- Jensens' inequality
  - 만약 위의 식에서, λ_i들이 probability 라고 해석한다면,
  - $\boldsymbol{f(E[x]) \leq E[f(x)]}$ (Convex 일 때!!!)





Source:

- $KL(p||q) = -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx$
- **Convexity**: $f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b)$ for $0 \leq \lambda \leq 1$
- $f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i)$
- Jensen’s inequality
  - 만약 위의 식에서, λ_i들이 probability 라고 해석한다면,
  - $\boldsymbol{f(E[x])} \leq \boldsymbol{E[f(x)]}$ (Convex 일 때!!!)
- Jensen‘s inequality를 활용하면,
  - $f\left(\int x p(x)\, dx\right) \leq \int f(x) p(x)\, dx$
- $KL(p||q) = -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx$

$$\geq -\ln \int p(x) \frac{q(x)}{p(x)} dx = -\ln \int q(x) dx = 0$$

Source:

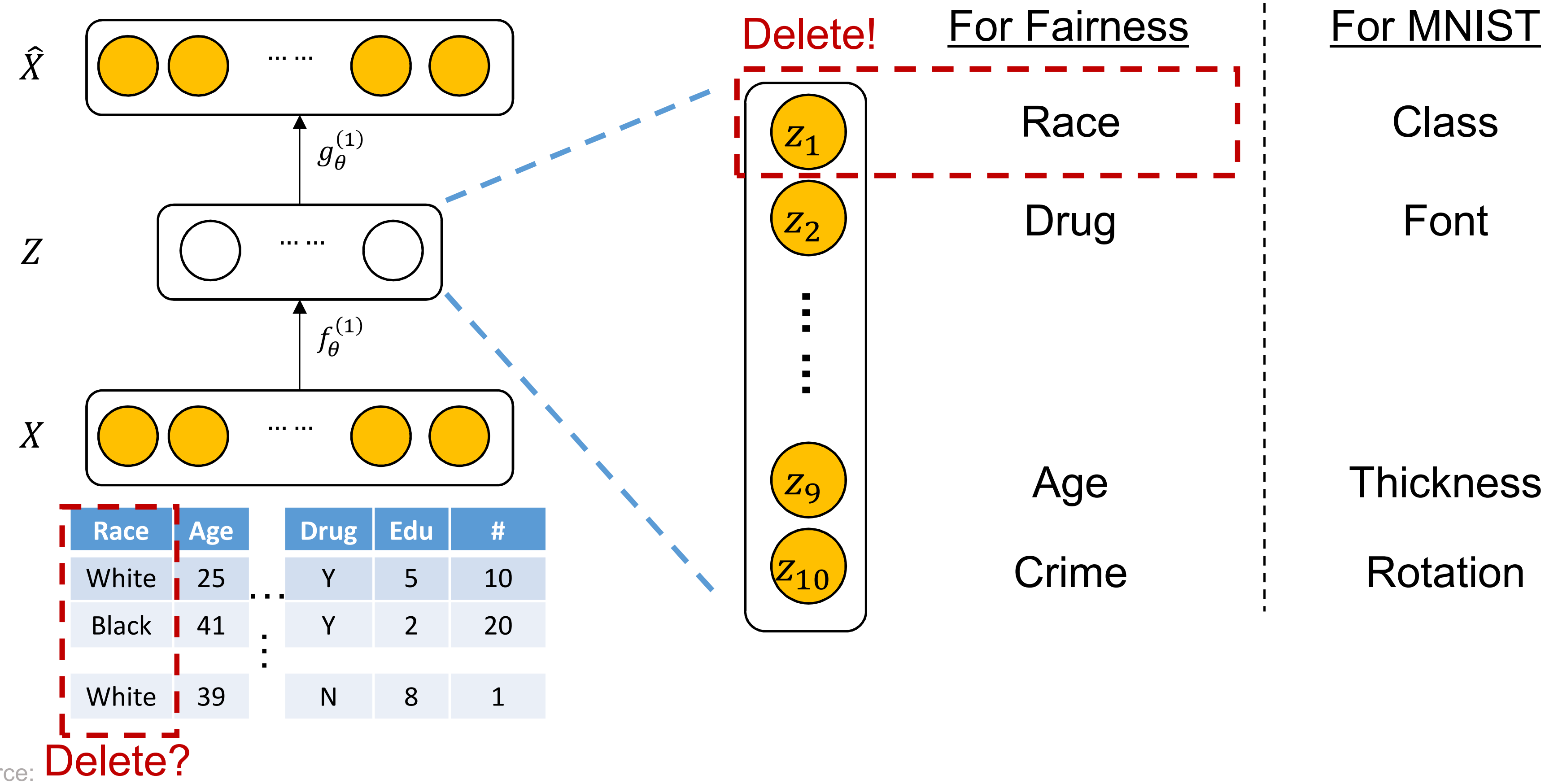
- 다시 돌아와서… KL Divergence가 왜 중요할까요? 어디에 쓰일까요?
- Maximum Likelihood Estimator
  - 우리는 unknown distribution $p(x)$에 관심이 많습니다.
  - 하지만, 모르기에, $q(x)$ 로 대신 모델링하고자 합니다.
  - 우리는 지금 AI/ML/DL 을 하고 있죠? Learnable parameter θ가 있습니다.
  - $q(x|\theta)$와 $p(x)$가 가까워지도록 θ를 학습하고 싶습니다.
  - $KL(p||q) = -\int p(x) \ln\left\{\frac{q(x)}{p(x)}\right\} dx \approx \sum_{n=1}^{N}\{-p(x_n) \boldsymbol{\ln q(x_n|\theta)} + p(x_n) \ln p(x_n)\}$
  - KL divergence 를 minimize 하는 것은, log-likelihood 를 maximize 하는 것
- Mutual Information
  - 만약 x 와 y가 독립이면 $p(x,y) = p(x)p(y)$ 이지만, 일반적으로는 성립하지 않습니다.
  - 그렇다면, 두개가 얼마나 독립에 가까운지는 어떻게 측정할 수 있을까요?
  - $I[x,y] = KL(p(x,y)||p(x)p(y)) = -\int\int p(x,y) \ln\left(\frac{p(x)p(y)}{p(x,y)}\right) dx\, dy$

Source:

- Mutual Information
  - 만약 x 와 y가 독립이면 $p(x,y) = p(x)p(y)$ 이지만, 일반적으로는 성립하지 않습니다.
  - 그렇다면, 두개가 얼마나 독립에 가까운지는 어떻게 측정할 수 있을까요?
  - $I[x,y] = KL(p(x,y)||p(x)p(y)) = -\int\int p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)dx\,dy$
- Properties
  - $I[x,y] \geq 0$
  - $I[x,y] = H(x) - H(x|y) = H(y) - H(y|x)$
    - ❖ $H(y) = -\int p(y)\ln p(y)\,dy = -\int\int p(x,y)\ln p(y)\,dydx$
    - ❖ $H(y|x) = -\int\int p(y,x)\ln p(y|x)\ dydx$
- Interpretation
  - Reduction in the uncertainty about x by virtue of being told the value of y
  - From a Bayesian perspective, the reduction in uncertainty about x as a consequence of the new observation y ($p(x)$: prior, $p(x|y)$: posterior)

Source:

- Disentangled Representation
  - Separate the role of hidden representation
  - Remove the sensitive related information (race, gender, …)



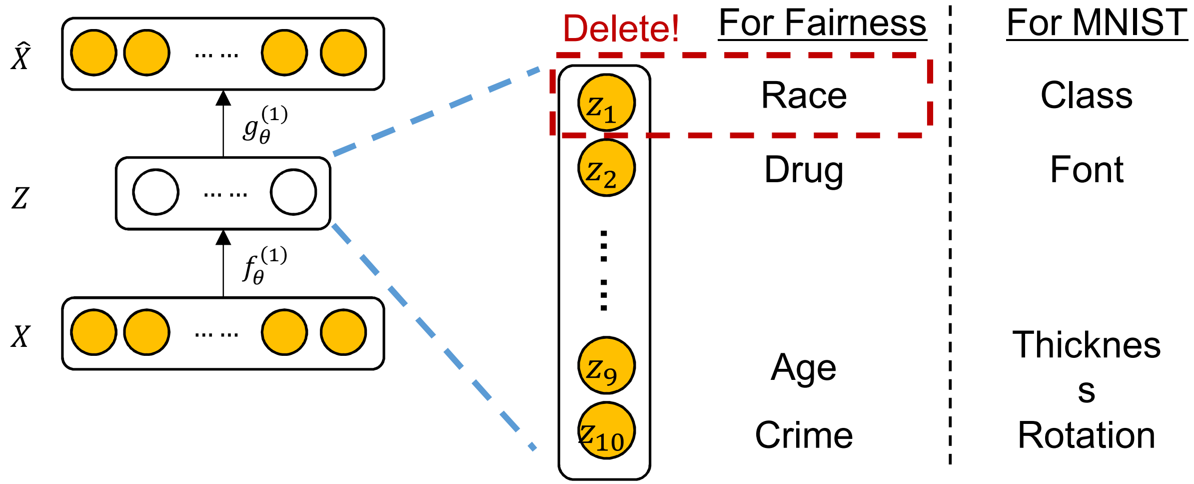
Source:

- Disentangled Representation
  - Separate the role of hidden representation
  - Remove the sensitive related information (race, gender, …)
- First approach) **Total Correlation (TC)**: popular measure of dependence
  - Minimize $\boldsymbol{KL(q(z)||\overline{q}(z))}$, where KL denotes the kl-divergence

❖ $\overline{q}(z) = \prod_{j=1}^{d} q(z_j)$

무슨 의미인지,
이해가 가시죠?

Note
X and Y are independent
iff $E[XY] = E[X]E[Y]$



Source: Disentangling by Factorising

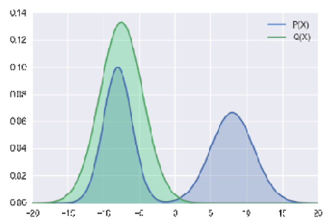
KL Divergence

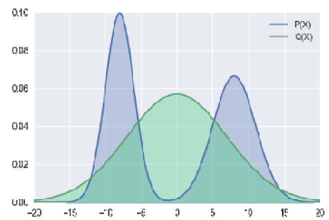
Information Theory

- KL Divergence 는 분포간의 거리를 잴 수 있는 한가지 도구 입니다.
  - 지금까지는, 두 vector 간의 similarity를 측정하는 방식에 대해서 살펴보았습니다.
  - 그렇다면, 두 분포간의 similarity 는 어떻게 측정할 수 있을까요?
  - 예를 들어, 다음 두 정규분포의 거리는 어떻게 잴 수 있을까요? $N(\mu_1, \sigma_1^2)$ 과 $N(\mu_2, \sigma_2^2)$
- KL Divergence (Kullback-Leibler)
  - Distribution $p(x)$와 $q(x)$
  - $KL(p||q) = -\int p(x) \ln \left\{\frac{q(x)}{p(x)}\right\} dx$
- Univariate Gaussian 에서의 KL-divergence
  - $KL(p||q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$
- Multivariate의 경우, 아래 링크 참조

NOTE)
$KL(p||q) \neq KL(q||p)$
⇒ metric 이라고 볼 순 없습니다.

p가 True, q가 approximation 이라고 할 때, $KL(p||q)$를 forward, $KL(q||p)$를 backward KL 이라고 부릅니다.





Source: https://mr-easy.github.io/2020-04-16-kl-divergence-between-2-gaussian-distributions/ https://agustinus.kristia.de/techblog/2016/12/21/forward-reverse-kl/