Information Theory

Jin Young Choi
Seoul National University

Outline

- Information
- Entropy
- Cross Entropy
- Error Backpropagation Learning
- Mutual Information
- Kullback Leibler Divergence
- Independent Component Analysis (ICA)
- Learning for ICA
- Blind Source Separation

Information

- Discrete random variable X is defined in the sample set Ψ

$$\Psi = \{x_k | k = 0,\ \pm 1,\ \ldots,\ \pm K\}$$

- Event $X = x_k$ occurs with probability $p_k = P(X = x_k)$

- **Information** ≡ surprise ≡ uncertainty

  The amount of information of the event is related to the *inverse* of the probability of occurrence. That is, the lower the probability p_k is, the more "surprise" there is, and the more "information".

$$I(x_k) = \log(\frac{1}{p_k}) = -\log p_k$$

내일도 지구가 회전한다 $p_k = 1$: 정보(×), surprise(×)

내일 미국이 북한을 공격한다 $p_k \ll 1$: 정보(O), surprise(O)

Information

- base=2 ⇒ 정보단위 bits
- base=e ⇒ 정보단위 nats
- 32 bit : 한 code의 정보는 $I(x_k) = -\log(\frac{1}{2^{32}}) = 32$

① $I(x_k) = 0$ for $p_k = 1$
② $I(x_k) \geq 0$ for $0 \leq p_k \leq 1$
③ $I(x_k) \geq I(x_i)$ for $p_k \leq p_i$

- Entropy : a measure of the *average amount of information conveyed per message*, i.e., expectation of Information

$$H(X) = E[I(X)] = \sum_{k=-K}^{K} p_k I(x_k) = -\sum_{k=-K}^{K} p_k \log p_k$$

Information

- Maximum entropy : when p_k is equiprobable.

$$0 \le H(X) \le -\sum_{k=-K}^{K} \frac{1}{2K+1} log \frac{1}{2K+1} = \log(2K+1)$$

- $H(X) = 0$ for an event that $p_k = 1$ o/w $p_k = 0$

- Theorem (Gray 1990): Relative entropy (or Kullback – Leibler divergence)

Discrete: $\sum_k p_k \log(\frac{p_k}{q_k}) \ge 0$

where p_k is probability mass ftn. (pmf), q_k is reference pmf

Continuous: $D_{p\|q} = \sum_{x \in \mathrm{X}} p_X(x) \log\left(\frac{p_X(x)}{q_X(x)}\right)$

where $p_X(x)$ is probability density ftn. (pdf), $q_X(x)$ is reference pdf.

Information

- Relative entropy (or Kullback – Leibler divergence) for neural networks

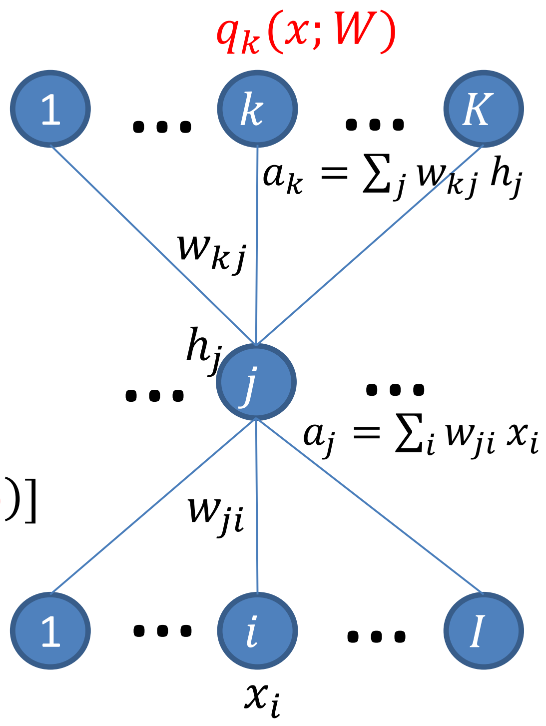
$$D_{p\|q}(W) = \sum_{x\in X} p_k(x) \log\left(\frac{p_k(x)}{q_k(x;W)}\right) = \sum_{x\in X} p_k(x) \log p_k(x) - \sum_{x\in X} p_k(x) \log q_k(x;W)$$

- Cross entropy for one-hot classification by deep learning

$$C_{p||q}(x;W) = -\sum_x \sum_k p_k(x)\ \log q_k(x;W)$$

- Cross entropy for multi-label classification by deep learning

$$C_{p||q}(X;\ W) = -\sum_x \sum_k [p_k(x)\ \log q_k(x;W) + (1-p_k(x)) \log(1-q_k(x;W))]$$



Backpropagation Learning Rule

- Empirical Risk Function:

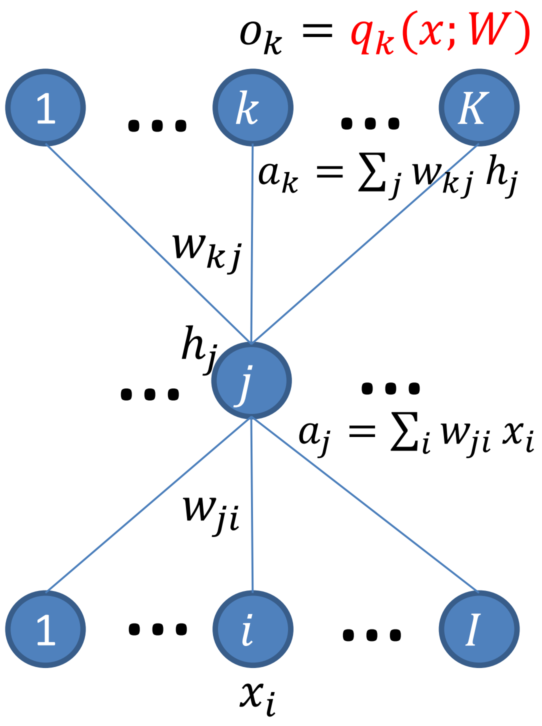
$E_d(w)$ Regression: L_2, linear
01001101: cross-entropy, sigmoid
00001000: cross-entropy, soft-max

- Gradient descent for output layer:

$$\Delta w_{kj} = -\eta \frac{\partial E_d}{\partial w_{kj}}$$

- Chain rule:

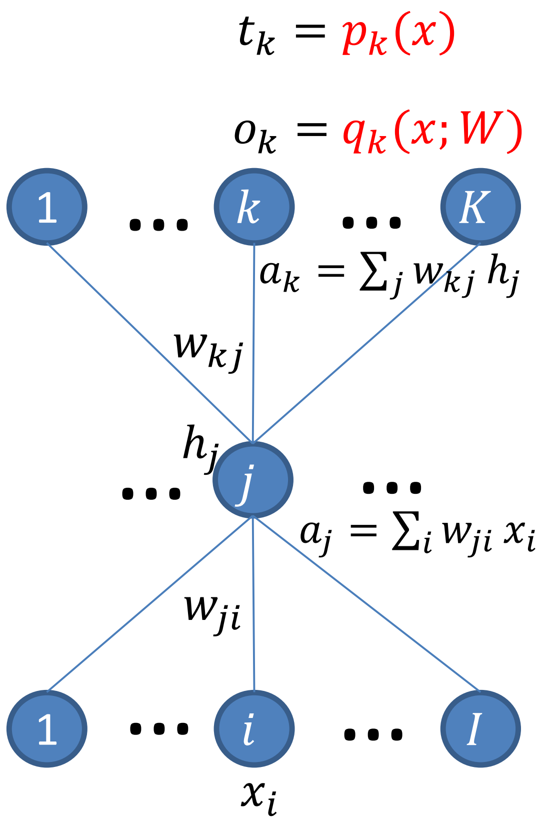
$$\frac{\partial E_d}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k} h_j$$



Backpropagation Learning Rule

- For multi-label classification (ex, output: 0110100), sigmoid activation function is used and the loss is defined by the cross entropy loss function:
  $E(w) = -\sum_k^K [t_k \log o_k(x,w) + (1-t_k)\log\big(1-o_k(x,w)\big)]$, where
  $o_k = \sigma(a_k) = \frac{1}{1+e^{-a_k}}$. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

  Sol.)



$$\frac{\partial E_d}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k} h_j$$

$$\Delta w_{kj} = -\eta \frac{\partial E_d}{\partial w_{kj}} = \eta \delta_k h_j$$

Backpropagation Learning Rule

- For multi-label classification (ex, output: 0110100), sigmoid activation function is used and the loss is defined by the cross entropy loss function:
  $E(w) = -\sum_k^K [t_k \log o_k(x,w) + (1-t_k)\log\big(1-o_k(x,w)\big)]$, where
  $o_k = \sigma(a_k) = \frac{1}{1+e^{-a_k}}$. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

  Sol.)
  $\frac{\partial \mathrm{E}}{\partial a_k} = \frac{\partial \mathrm{E}}{\partial o_k}\frac{\partial o_k}{\partial a_k}$.
  $\frac{\partial o_k}{\partial a_k} = \sigma(a_k)\big(1-\sigma(a_k)\big) = o_k(1-o_k)$,

Backpropagation Learning Rule

- For multi-label classification (ex, output: 0110100), sigmoid activation function is used and the loss is defined by the cross entropy loss function:
  $E(w) = -\sum_k^K [t_k \log o_k(x,w) + (1-t_k)\log(1-o_k(x,w))]$, where
  $o_k = \sigma(a_k) = \frac{1}{1+e^{-a_k}}$. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

Sol.)

$$\frac{\partial \mathrm{E}}{\partial a_k} = \frac{\partial \mathrm{E}}{\partial o_k}\frac{\partial o_k}{\partial a_k}.$$

$$\frac{\partial o_k}{\partial a_k} = \sigma(a_k)\big(1-\sigma(a_k)\big) = o_k(1-o_k).$$

$$\frac{\partial \mathrm{E}}{\partial a_k} = -t_k \frac{1}{o_k}\frac{\partial o_k}{\partial a_k} - (1-t_k)\frac{-1}{1-o_k}\frac{\partial o_k}{\partial a_k}$$

Backpropagation Learning Rule

- For multi-label classification (ex, output: 0110100), sigmoid activation function is used and the loss is defined by the cross entropy loss function:
  $E(w) = -\sum_k^K [t_k \log o_k(x,w) + (1-t_k)\log(1-o_k(x,w))]$, where
  $o_k = \sigma(a_k) = \frac{1}{1+e^{-a_k}}$. Then find $\frac{\partial E}{\partial a_k}$.

  Sol.)

  $$\frac{\partial E}{\partial a_k} = \frac{\partial E}{\partial o_k}\frac{\partial o_k}{\partial a_k}.$$

  $$\frac{\partial o_k}{\partial a_k} = \sigma(a_k)(1-\sigma(a_k)) = o_k(1-o_k).$$

  $$\frac{\partial E}{\partial a_k} = -t_k \frac{1}{o_k}\frac{\partial o_k}{\partial a_k} - (1-t_k)\frac{-1}{1-o_k}\frac{\partial o_k}{\partial a_k}$$

  $$= -t_k \frac{1}{o_k} o_k(1-o_k) - (1-t_k)\frac{-1}{1-o_k} o_k(1-o_k)$$

Backpropagation Learning Rule

- For multi-label classification (ex, output: 0110100), sigmoid activation function is used and the loss is defined by the cross entropy loss function: $E(w) = -\sum_k^K [t_k \log o_k(x,w) + (1-t_k)\log\big(1-o_k(x,w)\big)]$, where $o_k = \sigma(a_k) = \frac{1}{1+e^{-a_k}}$. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

Sol.)

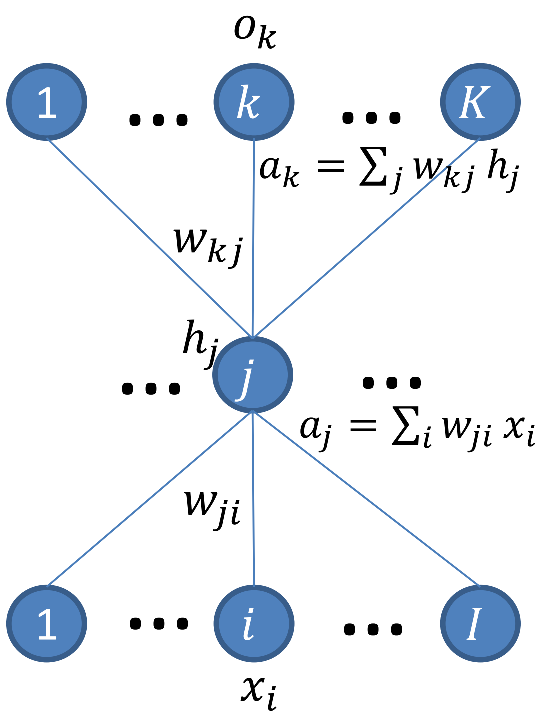
$$\frac{\partial \mathrm{E}}{\partial a_k} = \frac{\partial \mathrm{E}}{\partial o_k}\frac{\partial o_k}{\partial a_k}.$$

$$\frac{\partial o_k}{\partial a_k} = \sigma(a_k)\big(1-\sigma(a_k)\big) = o_k(1-o_k).$$

$$\begin{aligned}\frac{\partial \mathrm{E}}{\partial a_k} &= -t_k \frac{1}{o_k}\frac{\partial o_k}{\partial a_k} - (1-t_k)\frac{-1}{1-o_k}\frac{\partial o_k}{\partial a_k} \\ &= -t_k \frac{1}{o_k} o_k(1-o_k) - (1-t_k)\frac{-1}{1-o_k} o_k(1-o_k) \\ &= -t_k(1-o_k) + (1-t_k)o_k = o_k - t_k = -(t_k - o_k) = -\delta_k\end{aligned}$$

Backpropagation Learning Rule

- For multi-label classification (ex, output: 0110100), sigmoid activation function is used and the loss is defined by the cross entropy loss function:
  $E(w) = -\sum_k^K [t_k \log o_k(x,w) + (1-t_k)\log(1-o_k(x,w))]$, where
  $o_k = \sigma(a_k) = \frac{1}{1+e^{-a_k}}$. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.



Sol.)

$$\frac{\partial \mathrm{E}}{\partial a_k} = \frac{\partial \mathrm{E}}{\partial o_k}\frac{\partial o_k}{\partial a_k}.$$

$$\frac{\partial o_k}{\partial a_k} = \sigma(a_k)(1-\sigma(a_k)) = o_k(1-o_k).$$

$$\begin{aligned}\frac{\partial \mathrm{E}}{\partial a_k} &= -t_k \frac{1}{o_k}\frac{\partial o_k}{\partial a_k} - (1-t_k)\frac{-1}{1-o_k}\frac{\partial o_k}{\partial a_k} \\ &= -t_k \frac{1}{o_k} o_k(1-o_k) - (1-t_k)\frac{-1}{1-o_k} o_k(1-o_k) \\ &= -t_k(1-o_k) + (1-t_k)o_k = o_k - t_k = -(t_k - o_k) = -\delta_k\end{aligned}$$

$$\frac{\partial E_d}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k} h_j$$

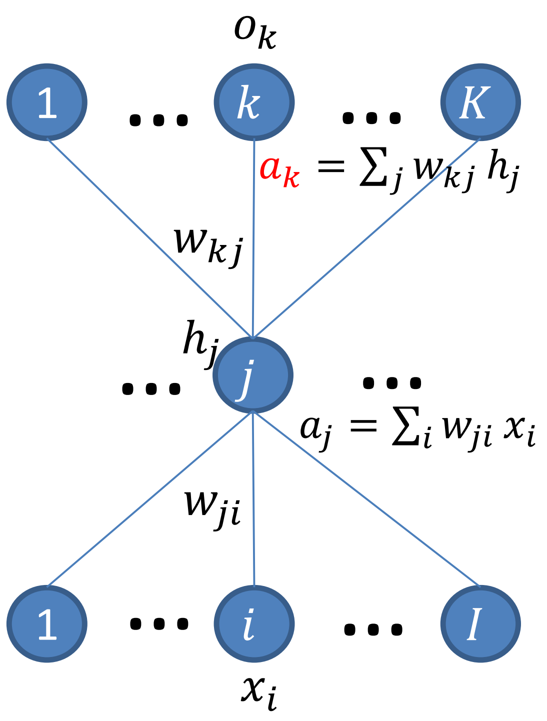
$$\Delta w_{kj} = -\eta \frac{\partial E_d}{\partial w_{kj}} = \eta \delta_k h_j$$

Backpropagation Learning Rule

- For multi-class classification (ex, [0 0 0 1 0 0]), the softmax activation function is used and the loss is defined by the cross entropy loss function: $E(w) = -\sum_i^K t_i \log(o_i(x,w))$, where $o_k(x,w) = \frac{e^{a_k}}{\sum_j e^{a_j}}$. The target value $t_k \in \{0,1\}$ is labelled by 1 hot vector. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

Sol.)

$$\frac{\partial E_n}{\partial a_k} = \frac{\partial}{\partial a_k}\left(-\sum_i^K t_i \log\left(\frac{e^{a_i}}{\sum_j e^{a_j}}\right)\right)$$



Backpropagation Learning Rule

- For multi-class classification (ex, [0 0 0 1 0 0]), the softmax activation function is used and the loss is defined by the cross entropy loss function: $E(w) = -\sum_i^K t_i \log(o_i(x,w))$, where $o_k(x,w) = \frac{e^{a_k}}{\sum_j e^{a_j}}$. The target value $t_k \in \{0,1\}$ is labelled by 1 hot vector. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

Sol.)

$$\frac{\partial E_n}{\partial a_k} = \frac{\partial}{\partial a_k}\left(-\sum_i^K t_i \log\left(\frac{e^{a_i}}{\sum_j e^{a_j}}\right)\right)$$

$$= \frac{\partial}{\partial a_k}\left(-\sum_i^K [t_i \log(e^{a_i}) - t_i \log(\sum_j e^{a_j})]\right)$$

Backpropagation Learning Rule

- For multi-class classification (ex, [0 0 0 1 0 0]), the softmax activation function is used and the loss is defined by the cross entropy loss function: $E(w) = -\sum_i^K t_i \log(o_i(x,w))$, where $o_k(x,w) = \frac{e^{a_k}}{\sum_j e^{a_j}}$. The target value $t_k \in \{0,1\}$ is labelled by 1 hot vector. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

Sol.)

$$\frac{\partial E_n}{\partial a_k} = \frac{\partial}{\partial a_k}\left(-\sum_i^K t_i \log\left(\frac{e^{a_i}}{\sum_j e^{a_j}}\right)\right)$$

$$= \frac{\partial}{\partial a_k}\left(-\sum_i^K [t_i \log(e^{a_i}) - t_i \log(\sum_j e^{a_j})]\right)$$

$$= \frac{\partial}{\partial a_k}\left(-\sum_i^K [t_i a_i - t_i \log(\sum_j e^{a_j})]\right) = -t_k + \sum_i t_i \frac{e^{a_k}}{\sum_j e^{a_j}}$$

Backpropagation Learning Rule

- For multi-class classification (ex, [0 0 0 1 0 0]), the softmax activation function is used and the loss is defined by the cross entropy loss function: $E(w) = -\sum_i^K t_i \log(o_i(x,w))$, where $o_k(x,w) = \frac{e^{a_k}}{\sum_j e^{a_j}}$. The target value $t_k \in \{0,1\}$ is labelled by 1 hot vector. Then find $\frac{\partial \mathrm{E}}{\partial a_k}$.

Sol.)

$$\frac{\partial E_n}{\partial a_k} = \frac{\partial}{\partial a_k}\left(-\sum_i^K t_i \log(\frac{e^{a_i}}{\sum_j e^{a_j}})\right)$$

$$= \frac{\partial}{\partial a_k}(-\sum_i^K [t_i \log(e^{a_i}) - t_i \log(\sum_j e^{a_j})])$$

$$= \frac{\partial}{\partial a_k}(-\sum_i^K [t_i a_i - t_i \log(\sum_j e^{a_j})]) = -t_k + \sum_i t_i \frac{e^{a_k}}{\sum_j e^{a_j}}$$

$$= -t_k + \frac{e^{a_k}}{\sum_j e^{a_j}}\sum_i t_i = o_k - t_k = -(t_k - o_k) = -\delta_k$$

Backpropagation Learning Rule

- For multi-class classification (ex, [0 0 0 1 0 0]), the softmax activation function is used and the loss is defined by the cross entropy loss function: $E(w) = -\sum_i^K t_i \log(o_i(x,w))$, where $o_k(x,w) = \frac{e^{a_k}}{\sum_j e^{a_j}}$. The target value $t_k \in \{0,1\}$ is labelled by 1 hot vector. Then find $\frac{\partial E}{\partial a_k}$.

Sol.)

$$\frac{\partial E_n}{\partial a_k} = \frac{\partial}{\partial a_k}\left(-\sum_i^K t_i \log\left(\frac{e^{a_i}}{\sum_j e^{a_j}}\right)\right)$$

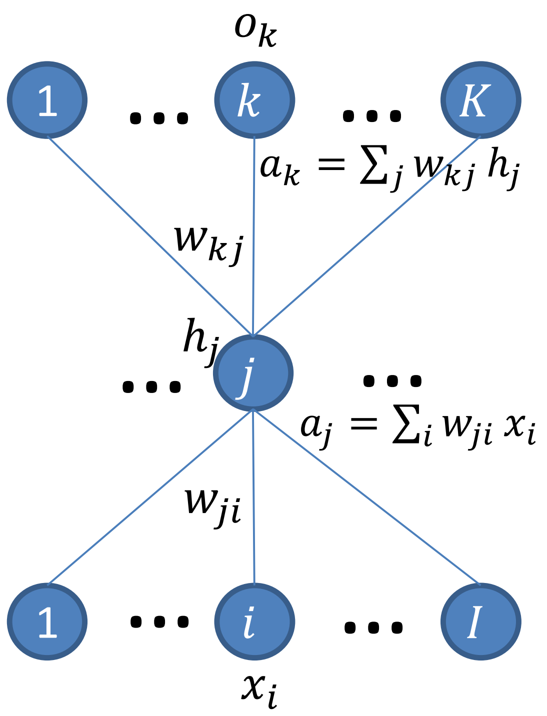
$$= \frac{\partial}{\partial a_k}\left(-\sum_i^K [t_i \log(e^{a_i}) - t_i \log(\sum_j e^{a_j})]\right)$$

$$= \frac{\partial}{\partial a_k}\left(-\sum_i^K [t_i a_i - t_i \log(\sum_j e^{a_j})]\right) = -t_k + \sum_i t_i \frac{e^{a_k}}{\sum_j e^{a_j}}$$

$$= -t_k + \frac{e^{a_k}}{\sum_j e^{a_j}}\sum_i t_i = o_k - t_k = -(t_k - o_k) = -\delta_k$$

$$\frac{\partial E_d}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}} = \frac{\partial E_d}{\partial a_k} h_j$$

$$\Delta w_{kj} = -\eta \frac{\partial E_d}{\partial w_{kj}} = \eta \delta_k h_j$$



Backpropagation Learning Rule

- Empirical Risk Function:

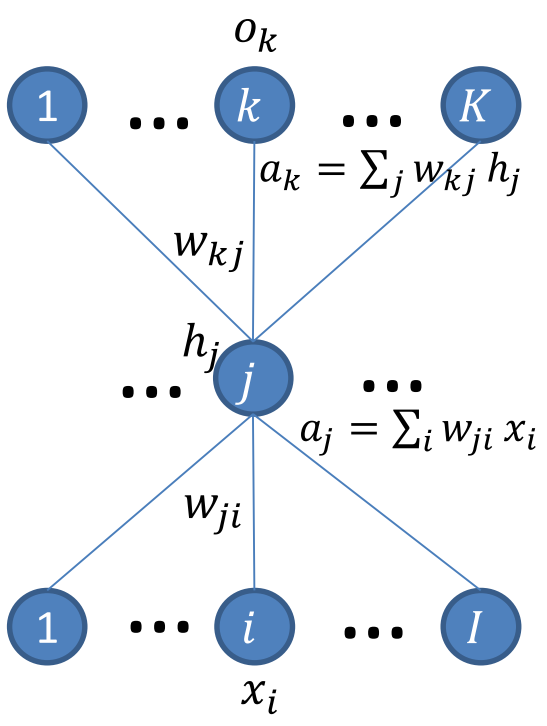
$E_d(w)$ Regression: L_2, linear
01001101: cross-entropy, sigmoid
00001000: cross-entropy, soft-max

- Gradient descent for hidden layer:

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}$$

- Chain rule:

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = \frac{\partial E_d}{\partial a_j} x_i$$

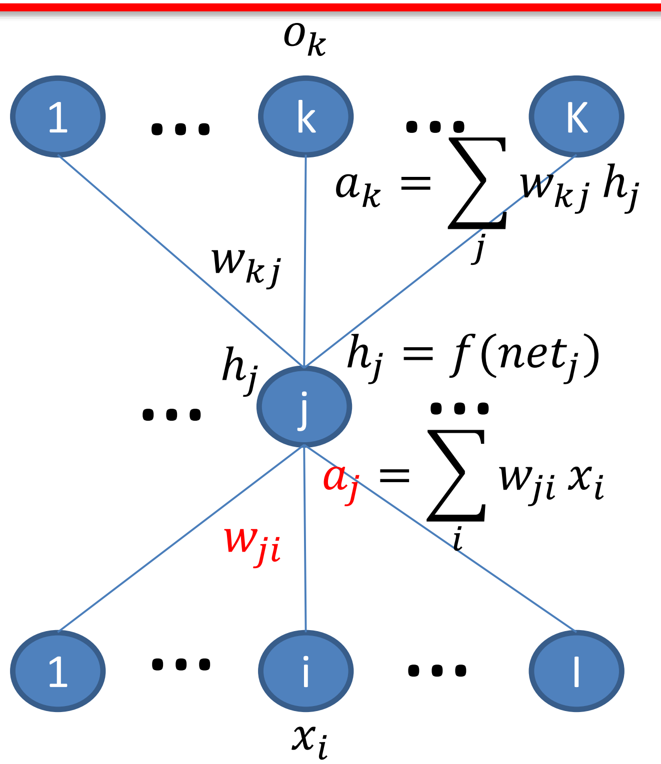


Backpropagation Learning Rule

- Chain rule:

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial a_j} x_i, \qquad \frac{\partial E_d}{\partial a_k} = -\delta_k$$

$$\begin{aligned}
\frac{\partial E_d}{\partial a_j} &= \sum_{k \in outputs} \frac{\partial E_d}{\partial a_k} \frac{\partial a_k}{\partial h_j} \frac{\partial h_j}{\partial a_j} \\
&= \sum_{k \in outputs} -\delta_k \frac{\partial a_k}{\partial h_j} \frac{\partial h_j}{\partial a_j} \\
&= \sum_{k \in outputs} -\delta_k w_{kj} \frac{\partial h_j}{\partial a_j} \\
&= \sum_{k \in outputs} -\delta_k w_{kj} f'(a_j) \\
&= \boxed{-\delta_j}
\end{aligned}$$



$$\Delta w_{ji} = \eta \delta_j x_i,$$

$$\delta_j = f'(a_j) \sum_{k \in outputs} \delta_k w_{kj}$$

Backpropagation Learning Rule

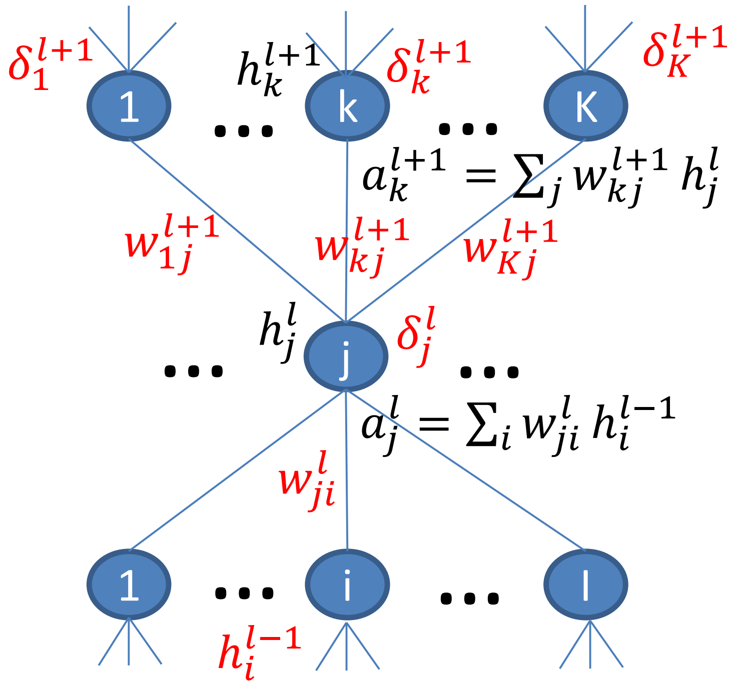
$$\Delta w_{ji}^{l} = \eta \delta_j^l h_i^{l-1},$$

$$\delta_j^l = f'(a_j^{l+1}) \sum_{k \in l+1 layer} \delta_k^{l+1} w_{kj}^{l+1}$$

$$\delta_k^L = -\frac{\partial E_d}{\partial a_k} = t_k - o_k$$

Regression: L_2, linear
01001101: cross-entropy, sigmoid
00001000: cross-entropy, soft-max

$\delta_k^L = t_k - o_k$



Backpropagation Learning Rule

Matrix Form (Forward)

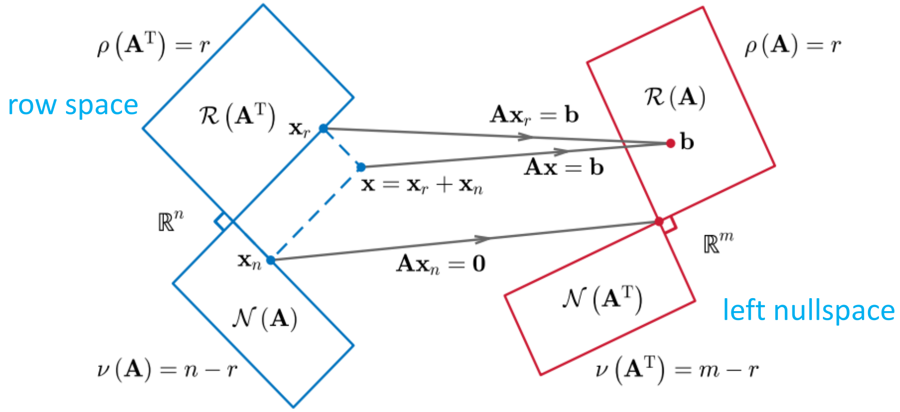
$$h^0 = x$$

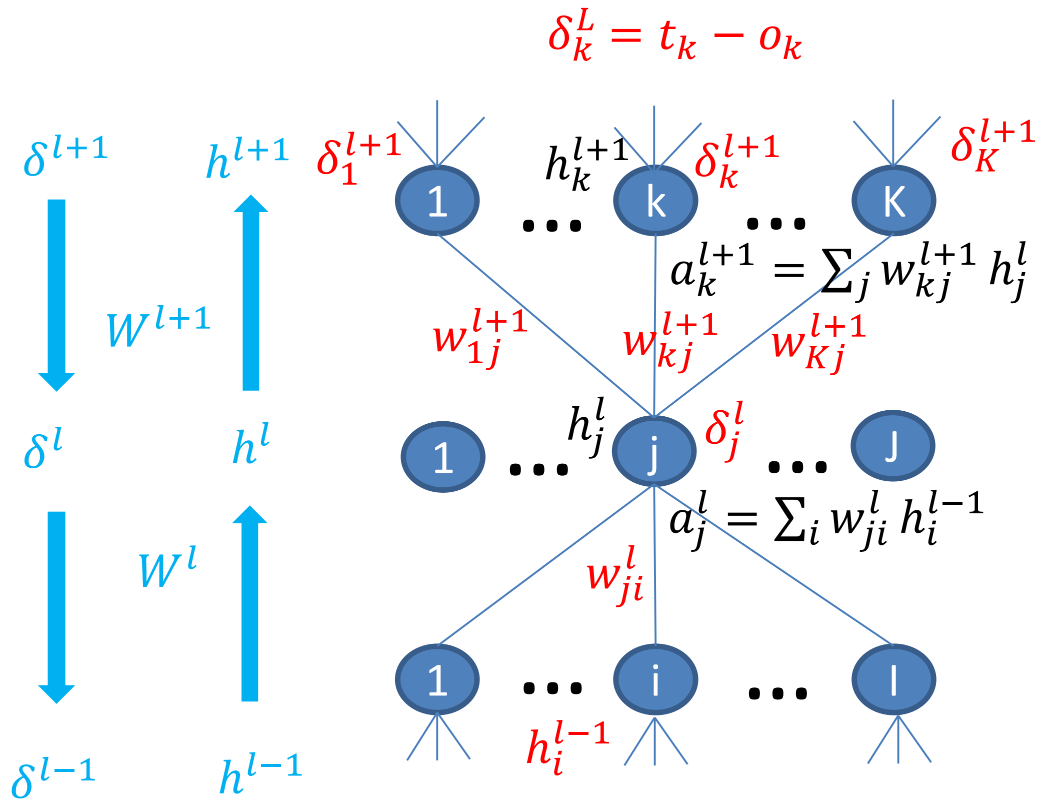
$$h^{l+1} = Diag[f]^{\circ} W^{l+1} h^l$$

Matrix Form (Backward. EBP)

$$\Delta W^l = \eta \delta^l {h^{l-1}}^T + \rho \Delta W^{l(old)}$$

$$\delta^l = Diag[f'(a^l)] {W^{l+1}}^T \delta^{l+1}$$





Error Back Propagation rule

2-layer Neural Network:

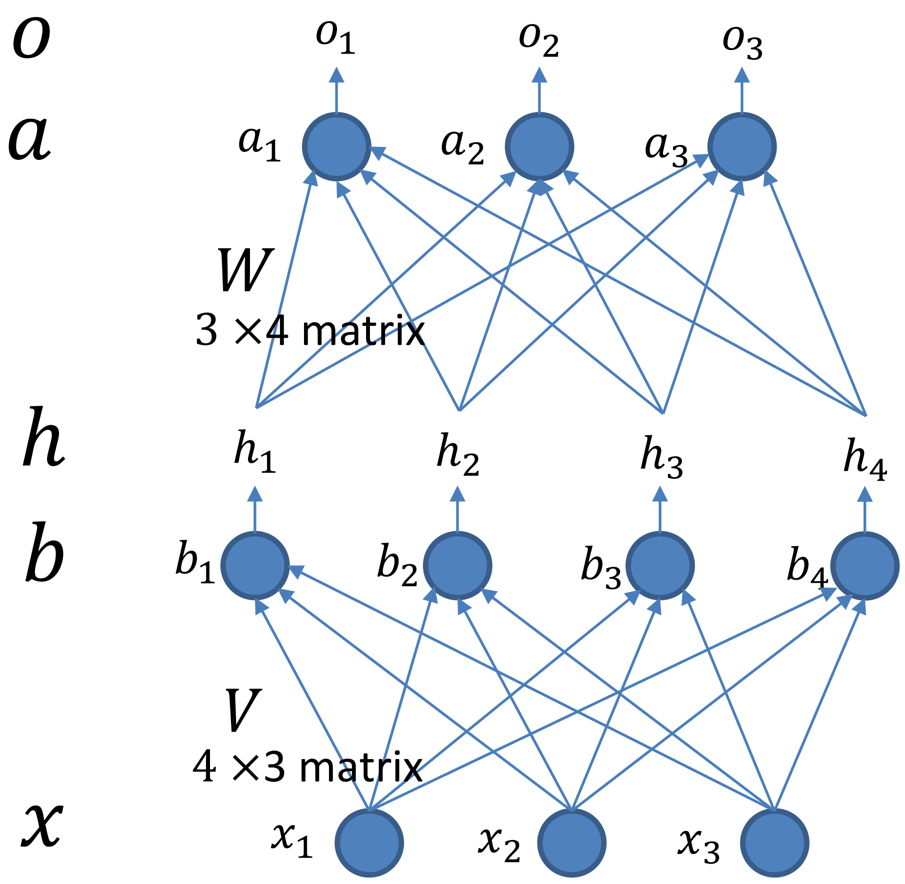
x: 3 ×1 input vector

V: 4 ×3 weight matrix

h: 4 ×1 hidden feature

W: 3 ×4 weight matrix

o: 3 ×1 output vector



Error Back Propagation rule

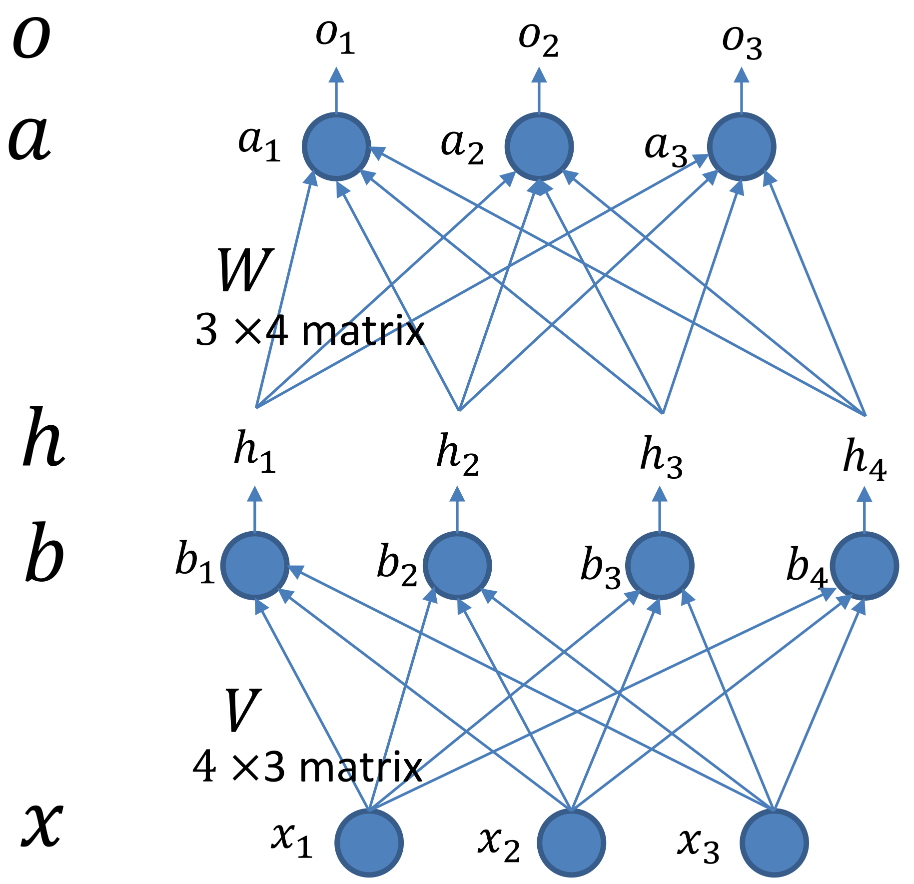
Forward Pass (first layer):

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \\ v_{41} & v_{42} & v_{43} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = Vx$$

$$h = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} = \begin{bmatrix} r(b_1) \\ r(b_2) \\ r(b_3) \\ r(b_4) \end{bmatrix}$$

$$a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ w_{31} & w_{32} & w_{33} & w_{34} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} = Wh$$

$$o = \begin{bmatrix} o_1 \\ o_2 \\ o_3 \end{bmatrix} = \begin{bmatrix} s(a_1) \\ s(a_2) \\ s(a_3) \end{bmatrix}$$



Error Back Propagation rule

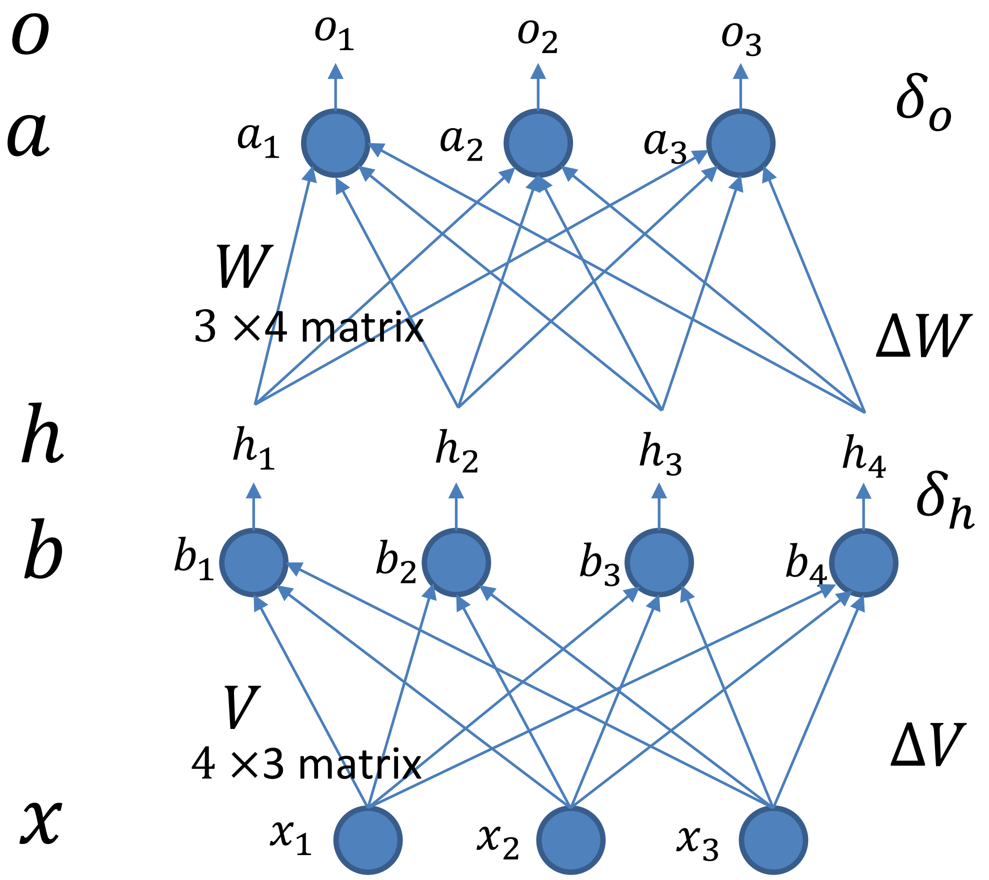
Backward Pass (second layer):

$$\delta_o = \begin{bmatrix} t_1 - o_1 \\ t_2 - o_2 \\ t_3 - o_3 \end{bmatrix} = \begin{bmatrix} \delta_{o_1} \\ \delta_{o_2} \\ \delta_{o_3} \end{bmatrix}$$

$$\Delta W \quad = \eta \delta_o h^T = \eta \begin{bmatrix} \delta_{o_1} \\ \delta_{o_2} \\ \delta_{o_3} \end{bmatrix} [h_1 \quad h_2 \quad h_3 \quad h_4]$$

3 ×4 matrix   3 ×1   1 ×4

$$W^{new} = W^{old} + \Delta W$$



Error Back Propagation rule

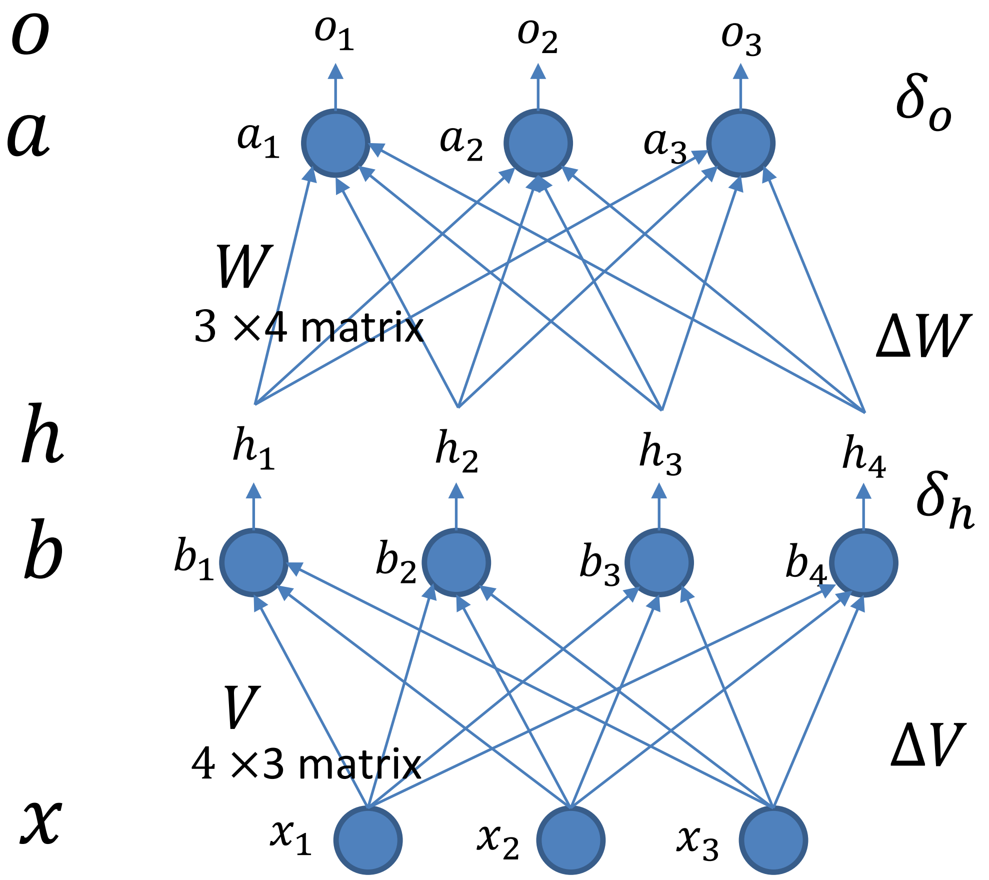
Backward Pass (first layer):

$$\begin{bmatrix}\delta_{h_1}\\ \delta_{h_2}\\ \delta_{h_3}\\ \delta_{h_4}\end{bmatrix} = \begin{bmatrix} r'(b_1)\bar{\delta}_{h_1}\\ r'(b_2)\bar{\delta}_{h_2}\\ r'(b_2)\bar{\delta}_{h_3}\\ r'(b_2)\bar{\delta}_{h_3}\end{bmatrix}, \quad \begin{bmatrix}\bar{\delta}_{h_1}\\ \bar{\delta}_{h_2}\\ \bar{\delta}_{h_3}\\ \bar{\delta}_{h_3}\end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} & w_{31}\\ w_{12} & w_{22} & w_{32}\\ w_{13} & w_{23} & w_{33}\\ w_{14} & w_{24} & w_{34}\end{bmatrix} \begin{bmatrix}\delta_{o_1}\\ \delta_{o_2}\\ \delta_{o_3}\end{bmatrix} = W^T\delta_o$$

$$\Delta V \;=\eta\delta_h x^T = \eta \begin{bmatrix}\delta_{h_1}\\ \delta_{h_2}\\ \delta_{h_3}\\ \delta_{h_4}\end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3\end{bmatrix}$$

4 ×3 matrix 4 ×1 1 ×3

$$V^{new} = V^{old} + \Delta V$$



Error Back Propagation rule

Forward Pass :

$$b = Vx, \;\; h = r(b)$$

$$a = Wh, \;\; o = \sigma(a)$$

Backward Pass :

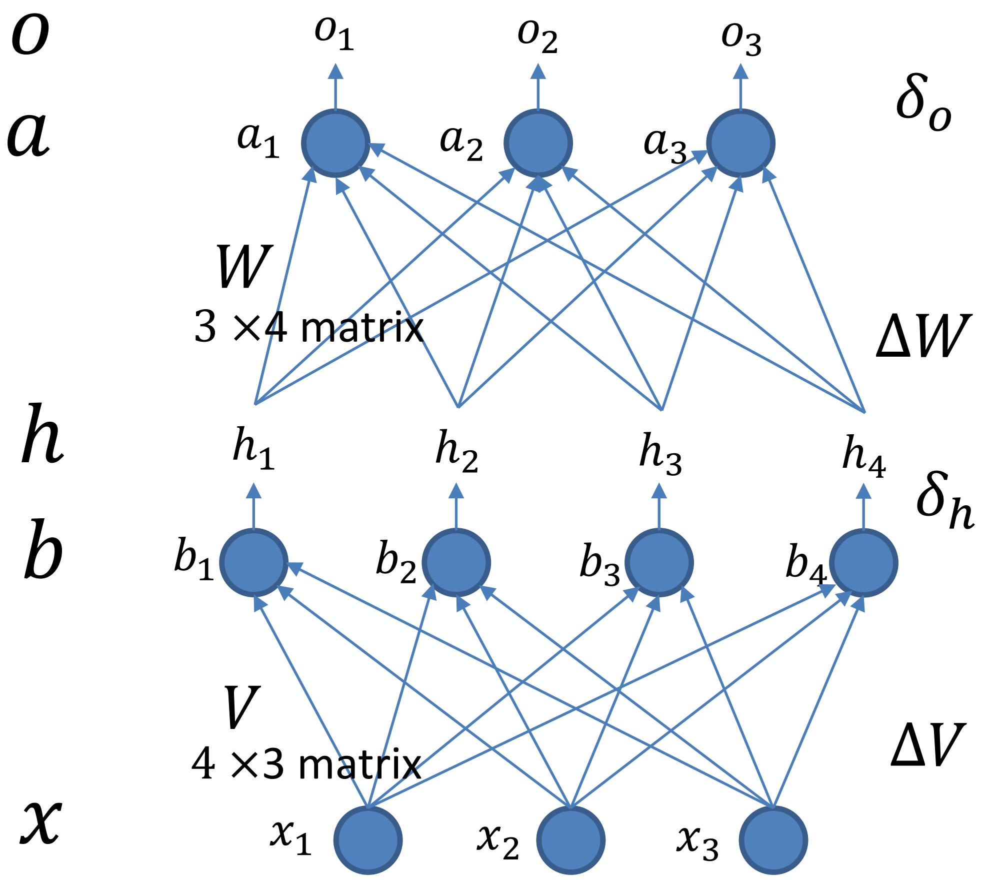
$$E = \frac{1}{2}\|t - o\|_2^2$$

$$\nabla_W E = (t - o)(-h^T) = -\,\delta_o h^T$$

$$W^{new} = W^{old} + \eta \delta_o h^T$$

$$\begin{aligned}\nabla_V E &= -W^T \delta_o (\nabla_V h) \\ &= -(Diag(r'(b) W^T \delta_o x^T) \\ &= -\delta_h x^T\end{aligned}$$

$$V^{new} = V^{old} + \eta \delta_h x^T$$

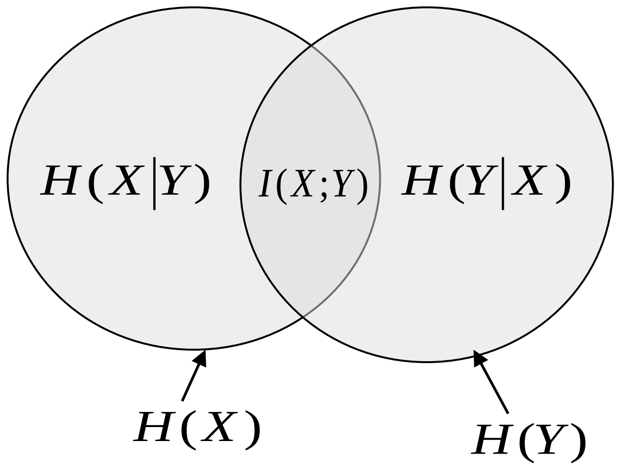


Mutual Information

- Conditional Entropy (조건부 불확실성의 량)
  Y 가 관측되고 난 후의 X 의 정보기대치 (Entropy)
  Y 와 연관이 있는 X 의 정보는 제외

- Theorem (Gray 1990)
  $H(X|Y) = H(X,Y) - H(Y)$
  $0 \leq H(X|Y) \leq H(X)$

$\leftarrow p(x|y) = \dfrac{p(x,y)}{p(y)}$



- Joint Entropy
  $H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log \underline{p(x,y)}$

Joint probability mass(or density) function

Mutual Information

- Mutual Information: Output Y 의 관측에 의해 알 수 있는 X 의 uncertainty (정보)

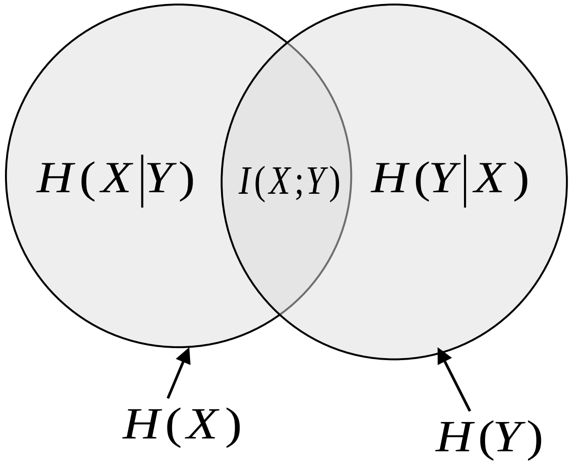
$$I(X;Y) = H(X) - H(X|Y)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= -\sum_{x\in X} p(x) \log (p(x)) - \sum_{y\in Y} p(y) \log (p(y))$$
$$+ \sum_{x\in X}\sum_{y\in Y} p(x,y) \log (p(x,y))$$
$$= \sum_{x\in X}\sum_{y\in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

$$p(x) = \sum_{y\in Y} p(x,y)$$

$$p(y) = \sum_{x\in X} p(x,y)$$

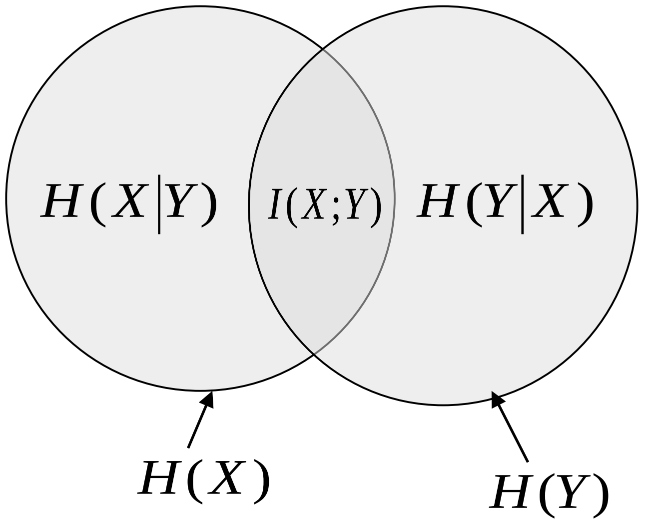
- KL-divergence & Independence ?

$$H(X) = I(X,X)$$



Mutual Information

- Properties of $I(X,Y)$
  ① $I(Y;X) = I(X;Y)$
  ② $I(X;Y) \geq 0$
  ③ $I(X;Y) = H(Y) - H(Y|X)$



Mutual Information

- Mutual Information for Continuous Random Variables

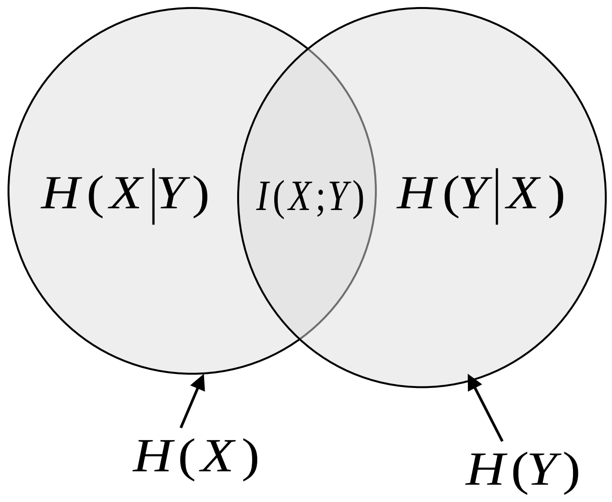
$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right)dxdy$$

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$$

$$= h(X) + h(Y) - h(X,Y)$$

$$I(X;Y) = I(Y;X)$$

$$I(X;Y) \geq 0$$



Exercise

- 프리미어리그에서 아스널과 토트넘이 경기를 하고 있다. TV를 보며 마음 껏 떠들 수 있도록 자리가 마련된 치킨집의 식객 30명과 바로 옆 삼겹살 집 식객 60명이 응원전을 펼치고 있다. 치킨집 사람들에게 어느 팀을 응원하는지 물었을 때 토트넘 10명, 아스널을 20명이 응원한다고 답했다.삼겹살 집에서는 각 팀을 몇 명이 응원하고 있는지 확인하지 못했다.
- 치킨 집에서 '토트넘'을 응원한다는 답변에 담긴 정보량(Information Gain)은?

Exercise

- 프리미어리그에서 아스널과 토트넘이 경기를 하고 있다. TV를 보며 마음 껏 떠들 수 있도록 자리가 마련된 치킨집의 식객 30명과 바로 옆 삼겹살 집 식객 60명이 응원전을 펼치고 있다. 치킨집 사람들에게 어느 팀을 응원하는지 물었을 때 토트넘 10명, 아스널을 20명이 응원한다고 답했다.삼겹살 집에서는 각 팀을 몇 명이 응원하고 있는지 확인하지 못했다.
- 치킨 집에서 ‘토트넘’을 응원한다는 답변에 담긴 정보량(Information Gain)은?

- 일목요연하게 내용 정리.

Exercise

- 프리미어리그에서 아스널과 토트넘이 경기를 하고 있다. TV를 보며 마음 껏 떠들 수 있도록 자리가 마련된 치킨집의 식객 30명과 바로 옆 삼겹살 집 식객 60명이 응원전을 펼치고 있다. 치킨집 사람들에게 어느 팀을 응원하는지 물었을 때 토트넘 10명, 아스널을 20명이 응원한다고 답했다.삼겹살 집에서는 각 팀을 몇 명이 응원하고 있는지 확인하지 못했다.
- 치킨 집에서 '토트넘'을 응원한다는 답변에 담긴 정보량(Information Gain)은?

- 일목요연하게 내용 정리.
- $Information = -\log p(X = x)$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60 - n)$명

Exercise

- 프리미어리그에서 아스널과 토트넘이 경기를 하고 있다. TV를 보며 마음 껏 떠들 수 있도록 자리가 마련된 치킨집의 식객 30명과 바로 옆 삽겹살 집 식객 60명이 응원전을 펼치고 있다. 치킨집 사람들에게 어느 팀을 응원하는지 물었을 때 토트넘 10명, 아스널을 20명이 응원한다고 답했다.삼겹살 집에서는 각 팀을 몇 명이 응원하고 있는지 확인하지 못했다.
- 치킨 집에서 ‘토트넘’을 응원한다는 답변에 담긴 정보량(Information Gain)은?

- 일목요연하게 내용 정리.
- $Information = -\log P(x)$

$P(X = \text{토트넘} | Y = \text{치킨집}) = 1/3$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

Exercise

- 프리미어리그에서 아스널과 토트넘이 경기를 하고 있다. TV를 보며 마음 껏 떠들 수 있도록 자리가 마련된 치킨집의 식객 30명과 바로 옆 삼겹살 집 식객 60명이 응원전을 펼치고 있다. 치킨집 사람들에게 어느 팀을 응원하는지 물었을 때 토트넘 10명, 아스널을 20명이 응원한다고 답했다.삼겹살 집에서는 각 팀을 몇 명이 응원하고 있는지 확인하지 못했다.
- 치킨 집에서 '토트넘'을 응원한다는 답변에 담긴 정보량(Information Gain)은?

- 일목요연하게 내용 정리.
- $Information\colon\ I(x) = -\log P(x)$

$$P(X = \text{토트넘} | Y = \text{치킨집}) = 1/3$$

$$I(X = \text{토트넘} | Y = \text{치킨집}) = \log 3$$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60 - n)$명

Exercise

- 치킨집에서 토트넘 응원하는 경우를 $X = 0$, 아스널 응원하는 경우를 $X = 1$이라 할 때 우측 표가 지닌 X의 엔트로피는?

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60 - n)$명

Exercise

- 치킨집에서 토트넘 응원하는 경우를 $X = 0$, 아스널 응원하는 경우를 $X = 1$이라 할 때 우측 표가 지닌 X 의 엔트로피는?

- $Entropy$: $H(X) = -\sum_{x \in X} p(x) \log p(x)$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60 - n)$명

Exercise

- 치킨집($Y = 0$)에서 토트넘 응원하는 경우를 $X = 0$, 아스널 응원하는 경우를 $X = 1$이라 할 때 우측 표가 지닌 X 의 엔트로피는?

- $Entropy\colon\ H(X) = -\sum_x p(x)\log p(x)$

- $H(x|Y = 0) = -\sum_x p(x|Y = 0)\log p(x|Y = 0)$

- $H(x|Y = 0) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60 - n)$명

Exercise

- KL-Divergence의 의미를 생각할 때 각 음식점 에서 두팀을 응원할 확률분포간의 KL-divergence, 즉 $D_{P(X|Y=0)||P(X|Y=1)}$을 최소로 하는 n 값을 구하시오.

$D_{P(Y=0)||P(Y=1)}$을 최소로한다는 것은 각 음식점에서 두팀을 응원할 확률 분포가 같게 된다는 의미이다.

즉, $P(X|Y = 0) = P(X|Y = 1)$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60 - n)$명

Exercise

- KL-Divergence의 의미를 생각할 때 각 음식점 에서 두 팀을 응원할 확률분포 간의 KL-divergence, 즉 $D_{P(X|Y=0)||P(X|Y=1)}$을 최소로 하는 n 값을 구하시오.

$D_{P(Y=0)||P(Y=1)}$을 최소로한다는 것은 각 음식점에서 두팀을 응원할 확률 분포가 같게 된다는 의미이다.

즉, $P(X|Y=0) = P(X|Y=1)$

$$\frac{1}{3} = \frac{n}{60}, \qquad \frac{2}{3} = \frac{60-n}{60} \rightarrow n = 20$$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

Exercise

- $D_{P(X|Y=0)||P(X|Y=1)}$을 최소로 하는 n 값을 최적화 방법으로 구하시오.

$$n^* = \underset{n}{\operatorname{argmin}}\, D_{P(X|Y=0)||P(X|Y=1)}$$

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

Exercise

- $D_{P(X|Y=0)||P(X|Y=1)}$을 최소로 하는 n 값을 최적화 방법으로 구하시오.

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

$$n^* = \underset{n}{\operatorname{argmin}}\, D_{P(X|Y=0)||P(X|Y=1)}$$

$$D_{P(X|Y=0)||P(X|Y=1)} = \sum_x P(X=x|Y=0)\log\frac{P(X=x|Y=0)}{P(X=x|Y=1)}$$

Exercise

- $D_{P(X|Y=0)||P(X|Y=1)}$을 최소로 하는 n 값을 최적화 방법으로 구하시오.

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

$$n^* = \operatorname*{argmin}_n D_{P(X|Y=0)||P(X|Y=1)}$$

$$D_{P(X|Y=0)||P(X|Y=1)} = \sum_x P(X=x|Y=0)\log\frac{P(X=x|Y=0)}{P(X=x|Y=1)}$$

$$= 1/3\log\frac{\frac{1}{3}}{\frac{n}{60}} + 2/3\log\frac{\frac{2}{3}}{\frac{60-n}{60}}$$

Exercise

- $D_{P(X|Y=0)||P(X|Y=1)}$을 최소로 하는 n 값을 최적화 방법으로 구하시오.

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

$$n^* = \underset{n}{\operatorname{argmin}}\, D_{P(X|Y=0)||P(X|Y=1)}$$

$$D_{P(X|Y=0)||P(X|Y=1)} = \sum_x P(X=x|Y=0)\log\frac{P(X=x|Y=0)}{P(X=x|Y=1)}$$
$$= 1/3\log\frac{\frac{1}{3}}{\frac{n}{60}} + 2/3\log\frac{\frac{2}{3}}{\frac{60-n}{60}}$$

$$\frac{d}{dn}D_{P(X|Y=0)||P(X|Y=1)} = \frac{n}{60}\left(-\frac{20}{n^2}\right) + \frac{60-n}{60}\left(\frac{40}{(60-n)^2}\right) = -\frac{1}{3n} + \frac{2}{3(60-n)} = \frac{-60+3n}{3n(60-n)} = 0 \rightarrow n = 20$$

Exercise

- $D_{P(X|Y=0)||P(X|Y=1)}$을 이용하여 구한 n이 참값이라고 할 때, 위 표가 지닌 응원팀(X)과 음식점(Y)에 관한 Mutual Information I(X, Y)을 수식을 사용하지 않고 개념적으로 구하시오. 그리고 수식을 사용하여 구하여 개념적으로 구한 경우와 비교하시오.

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

Exercise

- $D_{P(X|Y=0)||P(X|Y=1)}$을 이용하여 구한 n이 참값이라고 할 때, 위 표가 지닌 응원팀(X)과 음식점(Y)에 관한 Mutual Information I(X, Y)을 수식을 사용하지 않고 개념적으로 구하시오. 그리고 수식을 사용하여 구해보고 개념적으로 구한 경우와 비교하시오.

응원팀과 음식점은 서로 독립이다. 그 이유는 음식점에 따라 두 팀을 응원하는 확률 분포가 달라지지 않기 때문이다. 따라서 Mutual Information은 0 이다.

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

Exercise

- $D_{P(X|Y=0)||P(X|Y=1)}$을 이용하여 구한 n이 참값이라고 할 때, 위 표가 지닌 응원팀(X)과 음식점(Y)에 관한 Mutual Information $\mathrm{I}(X,Y)$을 수식을 사용하지 않고 개념적으로 구하시오. 그리고 수식을 사용하여 구해보고 개념적으로 구한 경우와 비교하시오.

응원팀과 음식점은 서로 독립이다. 그 이유는 음식점에 따라 두 팀을 응원하는 확률 분포가 달라지지 않기 때문이다. 따라서 Mutual Information은 0 이다.

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

$$\mathrm{I}(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x|y)p(y)\log\frac{p(x|y)p(y)}{p(x)p(y)}$$

$$= \frac{1}{3}\frac{1}{3}\log\frac{\frac{11}{33}}{\frac{11}{33}} + \frac{2}{3}\frac{1}{3}\log\frac{\frac{21}{33}}{\frac{21}{33}} + \frac{1}{3}\frac{2}{3}\log\frac{\frac{12}{33}}{\frac{12}{33}} + \frac{2}{3}\frac{2}{3}\log\frac{\frac{22}{33}}{\frac{22}{33}} = 0\,.$$

Exercise

- Mutual Information과 Conditional Entropy의 관계에 의하여 $H(X|Y)$을 구하시오.

	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

Exercise

- Mutual Information과 Conditional Entropy의 관계에 의하여 $H(X|Y)$을 구하시오.

$X \quad \backslash \quad Y$	치킨집	삼겹살집
토트넘 응원자	10	n명
아스널 응원자	20	$(60-n)$명

$$\mathrm{I}(X,Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) = 0$$

$$\mathrm{H}(\mathrm{X}|Y) = \mathrm{H}(X) = -\sum_x p(x)\log\, p(x) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$$

Evaluation

Metrics

- Precision
- Recall
- Accuracy
- F1 score
- ROC(Receiver Operating Characteristic) curve
- AUC(Area Under Curve)

Evaluation

Measures (Classification or Hypothesis Test)

		Actual Labels	
		Positive(1)	Negative(0)
Prediction Results	Positive(1)	**True Positive(TP)**	**False Positive(FP)**
	Negative(0)	**False Negative(FN)**	True Negative(TN)

Precision $= {}^{TP}/_{TP+FP}$: Positive 로 예측 한 것 중에 제대로 맞춘 비율

Recall $= {}^{TP}/_{TP+FN}$: 실제 Positive 중에서 예측을 맞춘 비율

Recall =Sensitivity, Specificity $= {}^{TN}/_{TN+FP}$

Evaluation

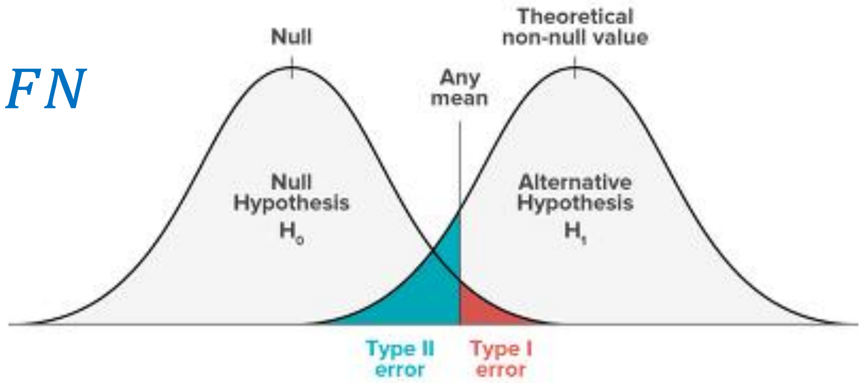
Precision-Recall Trade-off (ex, Hypothesis Test, 가설 검정)

		H_0	
		True	False
Test Results	Accept	True Positive(TP)	Type 2 error(FP)
	Reject	Type 1 error(FN)	True Negative(TN)

Precision $= {}^{TP}/_{TP+FP}$, Recall $= {}^{TP}/_{TP+FN}$

Type 1 error =P(reject $\boldsymbol{H_0}$ | $\boldsymbol{H_0}$ is true)

Type 2 error =P(accept $\boldsymbol{H_0}$ | $\boldsymbol{H_0}$ is not true)



Evaluation

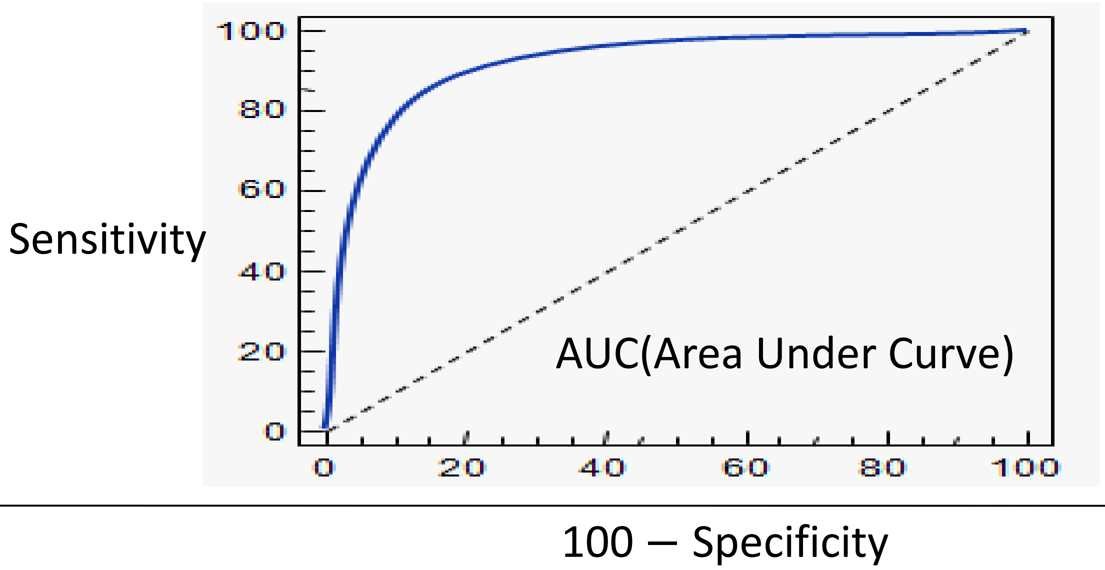
ROC(Receiver Operating Characteristic) curve

		Actual Labels	
		Positive(1)	Negative(0)
Prediction Results	Positive(1)	**True Positive(TP)**	**False Positive(FP)**
	Negative(0)	**False Negative(FN)**	True Negative(TN)

Sensitivity= $TP/_{TP+FN}$

Specificity = $TN/_{TN+FP}$

AUC(Area Under Curve)



Evaluation

Accuracy

		Actual Labels	
		Positive(1)	Negative(0)
Prediction Results	Positive(1)	**True Positive(TP)**	**False Positive(FP)**
	Negative(0)	**False Negative(FN)**	True Negative(TN)

$$\text{Specificity} = {}^{TN}/_{TN+FP}$$

$$\text{Recall} = {}^{TP}/_{TP+FN}$$

$$\text{Accuracy} = {}^{TP+TN}/_{TP+FN+TN+FP}$$

Evaluation

F1 score

		Actual Labels	
		Positive(1)	Negative(0)
Prediction Results	Positive(1)	**True Positive(TP)**	**False Positive(FP)**
	Negative(0)	**False Negative(FN)**	True Negative(TN)

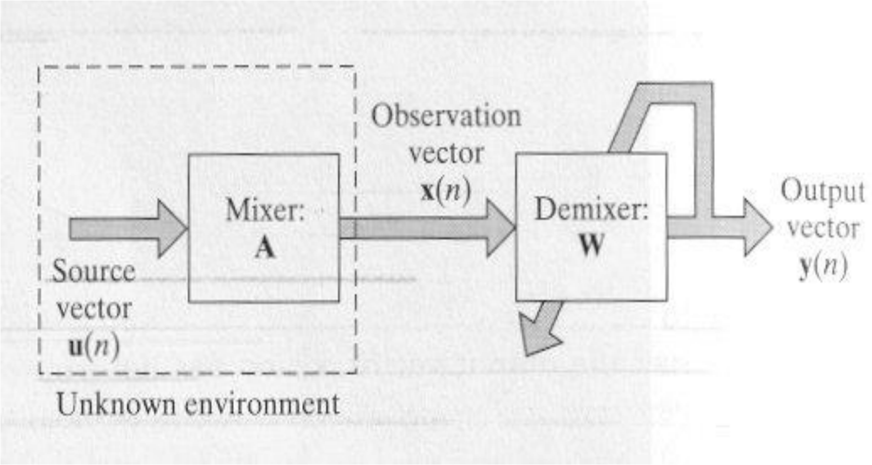
Precision $= {}^{TP}/_{TP+FP}$

Recall $= {}^{TP}/_{TP+FN}$

F1 score $= 2 \times \dfrac{Precision \times Recall}{Precision + Recall}$ (Precision과 Recall의 조화평균)

ICA(Independent Component Analysis)

- Blind source separation problem:
  Given N independent realizations of the observation vector X , find an estimate of the inverse of the mixing matrix A



- Algorithm of ICA:

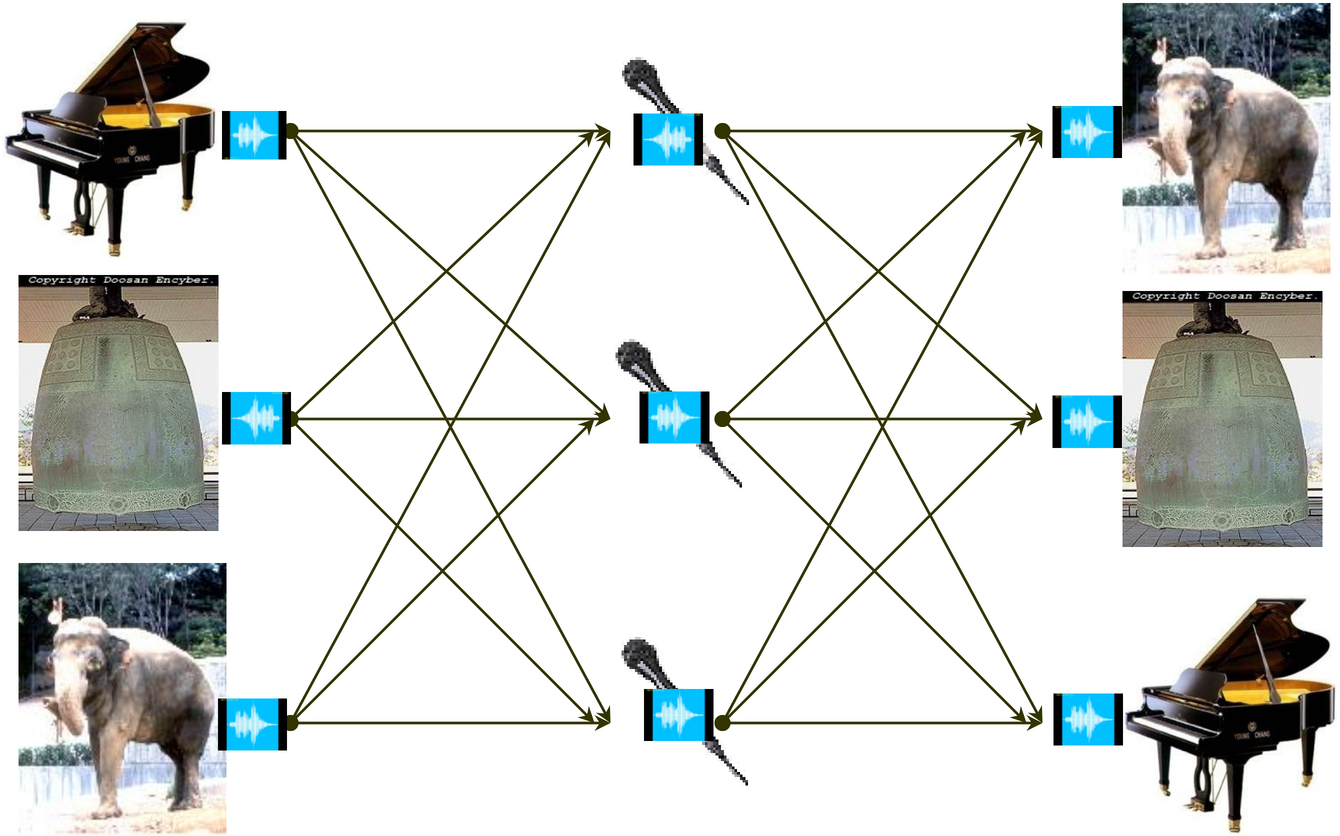
  → "as statistically independent as possible"

  → minimizing the mutual information between the each components

  of the output vector .

ICA(Independent Component Analysis)

- **ICA** Example



ICA(Independent Component Analysis)

- blind source separation problem

$U = [u_1, u_2, \dots, u_m]^T$: Independent Sources

$X = AU$, A: Mixing Matrix

$X = [x_1, x_2, \dots, x_m]^T$: Observations

$Y = WX$, : W: Demixing Matrix

U, X, Y : Zero mean Signals

→ $Y = WX = WAU = DPU$,
where D: Diagonal matrix, P: Permutation matrix

→ How to find W?

ICA(Independent Component Analysis)

- ICA : statistical independence
- Applications
  - Speech separation : teleconference
  - Array antenna processing
  - Multisensor biomedical records

    (태아의 심장박동을 어머니 심장박동과 분리)
  - Financial market data (Dominant data 추출)
  - Feature Extraction

ICA(Independent Component Analysis)

- Criterion for Statistical Independence

  Goal : Y_i, Y_j 간 mutual information을 최소화

  $$\min I\,(Y_i\,;Y_j) \quad i,j = 1,\cdots,m$$

  $$I(Y_1, Y_2, \cdots, Y_m) = D_{f_Y \| \tilde{f}_Y} = \int_{-\infty}^{\infty} f_Y(y) \log\left(\frac{f_Y(y)}{\prod_{i=1}^{m} \tilde{f}_{Y_i}(y_i)}\right) dY$$

  $$\tilde{f}_Y(y) = \prod_{i=1}^{m} \tilde{f}_{Y_i}(y_i),\ \ \tilde{f}_{Y_i}(y_i\ )\text{: Marginal p.d.f}$$

- Learning Rule for ICA

  $$\Delta w_{ik} = -\eta \frac{\partial}{\partial w_{ik}} D_{f\|\tilde{f}}$$

ICA(Independent Component Analysis)

- Kullback-Leibler Divergence

$$D_{f_Y \| \tilde{f}_Y} = \int_{-\infty}^{\infty} f_Y(\mathrm{y}) \log\left(\frac{f_Y(\mathrm{y})}{\prod_{i=1}^{m} \tilde{f}_{Y_i}(\mathrm{y}_i)}\right) dy$$

$$D_{f_Y \| \tilde{f}_Y} = \int_{-\infty}^{\infty} f_Y(\mathrm{y}) \log f_Y(\mathrm{y})\, dy - \sum_{i=1}^{m} \int_{-\infty}^{\infty} f_Y(\mathrm{y}) \log \tilde{f}_{Y_i}(\mathrm{y}_i)\, dy$$

- The second term is

$$\int_{-\infty}^{\infty} \log \tilde{f}_{Y_i}(\mathrm{y}_i)\ \left[\int_{-\infty}^{\infty} f_Y(\mathrm{y})\, dy^{(i)}\right] dy_i = \int_{-\infty}^{\infty} \tilde{f}_{Y_i}(\mathrm{y}_i) \log \tilde{f}_{Y_i}(\mathrm{y}_i)\, dy_i$$

$$= -\tilde{h}(Y_i) \quad \text{:marginal entropy}$$

- Kullback-Leibler Divergence

$$D_{f_Y \| \tilde{f}_Y} = -h(Y) + \sum_{i=1}^{m} \tilde{h}(Y_i)$$

ICA(Independent Component Analysis)

- Entropy $h(Y)$

$$h(Y) = h(\mathrm{W}X) = h(X) + \log\left|\det(\mathrm{W})\right|,$$

$$(f_Y(y) = |\det(\mathrm{W})|^{-1}\ f_X(x),\ \ dy = |\det(\mathrm{W})|\, dx)$$

- Marginal entropy $h(Yi)$

Pdf of Y_i is obtained using truncate of Gram-Charlier series

$$\tilde{f}_{Y_i}(y_i(\mathrm{W})) = \alpha(y_i)[1 + \textstyle\sum_{k=3}^{\infty} c_{ik}\, H_k(y_i)]$$

where

$\alpha(y_i) = 1/\sqrt{2\pi}\exp(-yi^2)$

$H_k(y_i)$: Hermite polynomials

Cumulants $\{c_{ik} : k = 3, 4, \ldots,\}$ is obtained from k-th order moment of Y_i

Hermite polynomials: $H_3(y) = y^3 - 3x, H_4(y) = y^4 - 6y^2 + 3, \ldots$

ICA(Independent Component Analysis)

- $\tilde{f}_{Y_i}(y_i(\mathrm{W})) = \alpha(y_i)[1 + \sum_{k=3}^{\infty} c_{ik}\, H_k(y_i)]$

- The index grouping is done as $k = (0), (3), (4,6), (5,7,9), \dots$
- By choosing by $k = (4,6)$

$$\tilde{f}_{Y_i}(y_i) = \alpha(y_i)\left(1 + \frac{k_{i,3}}{3!} H_3(y_i) + \frac{k_{i,4}^2}{4!} H_4(y_i) + \frac{(k_{i,6} + 10k_{i,3}^2)}{6!} H_6(y_i)\right)$$

- c_{ik} and k-th order moment of Y_i

$$k_{i,3} = m_{i,3},\ k_{i,4} = m_{i,4} - 3m_{i,2}^2$$

$$k_{i,6} = m_{i,6} - 10m_{i,3}^2 - 15m_{i,2}m_{i,4} + 30m_{i,2}^3$$

$$m_{i,k} = E\left[Y_i^k\right] = E\left[\left(\sum_{j=1}^{m} \boxed{w_{ij}} X_j\right)^k\right]$$

ICA(Independent Component Analysis)

- The cumulants are functions of W.
- Gradient of K-L divergence

1) $$\frac{\partial}{\partial w_{ij}}\log(\det(\mathrm{W})) = \frac{1}{\det(\mathrm{W})}\frac{\partial}{\partial w_{ij}}\det(\mathrm{W})$$
$$= \frac{A_{ij}}{\det(\mathrm{W})} \quad = (\mathrm{W}^{-T})_{ij}$$

2) $$\frac{\partial \kappa_{i,3}}{\partial w_{ij}} \approx 3\mathrm{y}_i{}^2 x_j, \qquad \frac{\partial \kappa_{i,4}}{\partial w_{ij}} \approx -8y_i{}^3 x_j \quad \ldots\ldots$$

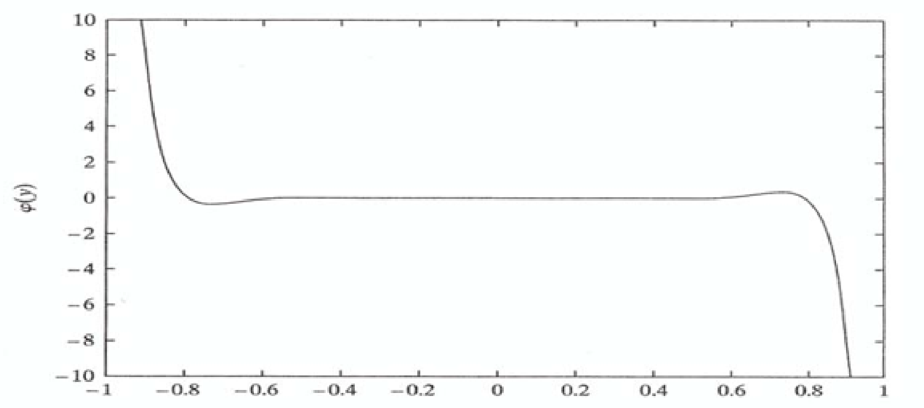
ICA(Independent Component Analysis)

- Minimization of Kullback-Leibler Divergence

$$D_{f_{\mathrm{Y}}\|\tilde{f}_{\mathrm{Y}}} = -h(\mathrm{Y}) + \sum_{i=1}^{m} \tilde{h}(\mathrm{Y}_i)$$

$$\frac{\partial}{\partial w_{ij}} D_{f\|\tilde{f}}(\mathrm{W}) \approx -(\mathrm{W}^{-T})_{ij} + \varphi(y_i)x_j$$

$$\varphi(y_i) = \frac{1}{2}y_i^5 + \frac{2}{3}y_i^7 + \frac{15}{2}y_i^9 + \frac{2}{15}y_i^{11} - \frac{112}{3}y_i^{13} + 128y_i^{15} - \frac{512}{3}y_i^{17}$$



ICA(Independent Component Analysis)

- Learning algorithm for ICA

$$\Delta w_{ij} = -\eta \frac{\partial}{\partial w_{ij}} D_{f \| \tilde{f}}$$

$$= \eta \left((\mathrm{W}^{-T})_{ij} - \phi(y_i) x_j \right)$$

$$\Delta \mathrm{W} = \eta(\mathrm{W}^{-T} - \phi(y)\mathrm{x}^T)$$

$$\Delta \mathrm{W} = \eta[\mathrm{I} - \phi(y)\mathrm{x}^T\mathrm{W}^T]\mathrm{W}^{-T}$$

$$= \eta[\mathrm{I} - \phi(y)\mathrm{y}^T]\mathrm{W}^{-T}$$

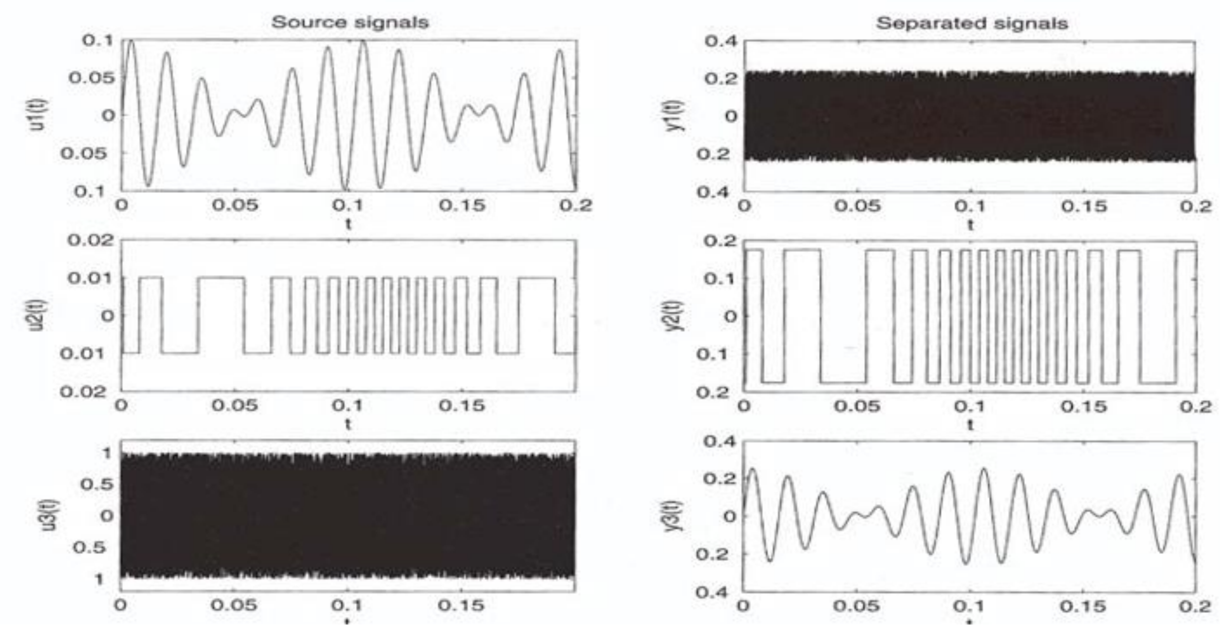
$$\mathrm{W}(n+1) = \mathrm{W}(n) + \eta(n)\left[\mathrm{I} - \phi\left(y(n)\right)y^T(n)\right]\mathrm{W}^{-T}(n)$$

ICA(Independent Component Analysis)

- Experiments

$$u_1(n) = 0.1\sin(400n)\cos(30n)$$
$$u_2(n) = 0.01\, sgn(\sin(500n + 9\cos(40n)))$$
$$u_3(n) = noise\ uniformly\ distributed\ in\ [-1, 1]$$

$$A = \begin{bmatrix} 0.56 & 0.79 & -0.37 \\ -0.75 & 0.65 & 0.86 \\ 0.17 & 0.32 & -0.48 \end{bmatrix}$$



Exercise

- In computer science(CS) department, the probability of dropping the machine learning(ML) course in March is 1/6, that in April is 1/3, and the probability of taking ML course to the end without dropping is 1/2, whereas those in Electrical engineering(EE) department are 1/8, 1/8, and 3/4, respectively. Meanwhile, the portions of CS & EE students in ML course are 1/5 & 4/5, respectively. Letting X be the random variable on dropping or not of a student, and Y be the random variable on the department of a student, find the followings.

  1. Conditional entropy $H(X|Y)$.
  2. Mutual information $I(X;Y)$.

Exercise

- In computer science(CS) department, the probability of dropping the machine learning(ML) course in March is 1/6, that in April is 1/3, and the probability of taking ML course to the end without dropping is 1/2, whereas those in Electrical engineering(EE) department are 1/8, 1/8, and 3/4, respectively. Meanwhile, the portions of CS & EE students in ML course are 1/5 & 4/5, respectively. Letting X be the random variable on dropping or not of a student, and Y be the random variable on the department of a student, find $H(X|Y)$, $I(X;Y)$.
- 서술식을 수식으로 변경:

Exercise

- In computer science(CS) department, the probability of dropping the machine learning(ML) course in March is 1/6, that in April is 1/3, and the probability of taking ML course to the end without dropping is 1/2, whereas those in Electrical engineering(EE) department are 1/8, 1/8, and 3/4, respectively. Meanwhile, the portions of CS & EE students in ML course are 1/5 & 4/5, respectively. Letting X be the random variable on dropping or not of a student, and Y be the random variable on the department of a student, find $H(X|Y)$, $I(X;Y)$.
- 서술식을 수식으로 변경:

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- $H(X|Y)=?,\ I(X;Y)=?.$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X = 0|Y = 0) = 1/6, P(X = 1|Y = 0) = 1/3, P(X = 2|Y = 0) = 1/2$
- $P(X = 0|Y = 1) = 1/8, P(X = 1|Y = 1) = 1/8, P(X = 2|Y = 1) = 3/4$
- $P(Y = 0) = 1/5,\ P(Y = 1) = 4/5$
- $H(X|Y) =?,\ I(X;Y) =?.$

- Sol. $H(X|Y) =?,\ I(X;Y) =?.$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- $H(X|Y)=?,\ I(X;Y)=?.$

- Sol. $H(X|Y)=?,\ I(X;Y)=?.$

$$H(X|Y)=H(X,Y)-H(Y).$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- $H(X|Y)=?,\ I(X;Y)=?.$

- Sol. $H(X|Y)=?,\ I(X;Y)=?.$

$$H(X|Y)=H(X,Y)-H(Y).$$

$$H(Y)=-\Sigma_{y\in Y}p(y)logp(y)=-\frac{1}{5}\log\frac{1}{5}-\frac{4}{5}\log\frac{4}{5}=0.7219$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X = 0|Y = 0) = 1/6, P(X = 1|Y = 0) = 1/3, P(X = 2|Y = 0) = 1/2$
- $P(X = 0|Y = 1) = 1/8, P(X = 1|Y = 1) = 1/8, P(X = 2|Y = 1) = 3/4$
- $P(Y = 0) = 1/5,\ P(Y = 1) = 4/5$
- $H(X|Y) =?,\ I(X;Y) =?.$

- Sol. $H(X|Y) =?,\ I(X;Y) =?.$

$$H(X|Y) = H(X,Y) - H(Y).$$

$$H(Y) = -\Sigma_{y\in Y} p(y) log p(y) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} = 0.7219$$

$$H(X,Y) = \ -\Sigma_{x\in X}\Sigma_{y\in Y} p(x,y) log p(x,y)$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- $H(X|Y)=?,\ I(X;Y)=?.$

- Sol. $H(X|Y)=?,\ I(X;Y)=?.$

$$H(X|Y)=H(X,Y)-H(Y).$$

$$H(Y)=-\Sigma_{y\in Y}p(y)logp(y)=-\frac{1}{5}\log\frac{1}{5}-\frac{4}{5}\log\frac{4}{5}=0.7219$$

$$H(X,Y)=\ -\Sigma_{x\in X}\Sigma_{y\in Y}p(x,y)logp(x,y)$$

$$H(X,Y)=\ -\Sigma_{x\in X}\Sigma_{y\in Y}p(x|y)\times p(y)logp(x|y)\times p(y)$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- $H(X|Y)=?,\ I(X;Y)=?.$
- Sol. $H(X|Y)=?,\ I(X;Y)=?.$

$$H(X|Y)=H(X,Y)-H(Y).$$

$$H(Y)=-\Sigma_{y\in Y}p(y)logp(y)=-\frac{1}{5}\log\frac{1}{5}-\frac{4}{5}\log\frac{4}{5}=0.7219$$

$$H(X,Y)=\ -\Sigma_{x\in X}\Sigma_{y\in Y}p(x,y)logp(x,y)$$

$$H(X,Y)=\ -\Sigma_{x\in X}\Sigma_{y\in Y}p(x|y)\times p(y)logp(x|y)\times p(y)$$

$$=-\frac{1}{6}*\frac{1}{5}\log\left(\frac{1}{6}*\frac{1}{5}\right)-\frac{1}{3}*\frac{1}{5}\log\left(\frac{1}{3}*\frac{1}{5}\right)-\frac{1}{2}*\frac{1}{5}\log\left(\frac{1}{2}*\frac{1}{5}\right)$$

$$-\frac{1}{8}*\frac{4}{5}\log\left(\frac{1}{8}*\frac{4}{5}\right)-\frac{1}{8}*\frac{4}{5}\log\left(\frac{1}{8}*\frac{4}{5}\right)-\frac{3}{4}*\frac{4}{5}\log\left(\frac{3}{4}*\frac{4}{5}\right)\ =\ 1.8628$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X = 0|Y = 0) = 1/6, P(X = 1|Y = 0) = 1/3, P(X = 2|Y = 0) = 1/2$
- $P(X = 0|Y = 1) = 1/8, P(X = 1|Y = 1) = 1/8, P(X = 2|Y = 1) = 3/4$
- $P(Y = 0) = 1/5,\ P(Y = 1) = 4/5$
- $H(X|Y) = ?,\ I(X;Y) = ?.$
- Sol. $H(X|Y) = ?,\ I(X;Y) = ?.$

$$H(X|Y) = H(X,Y) - H(Y).$$

$$H(Y) = -\Sigma_{y\in Y} p(y) logp(y) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} = 0.7219$$

$$H(X,Y) = -\Sigma_{x\in X}\Sigma_{y\in Y} p(x,y) logp(x,y)$$

$$H(X,Y) = -\Sigma_{x\in X}\Sigma_{y\in Y} p(x|y) \times p(y) logp(x|y) \times p(y)$$

$$= -\frac{1}{6}*\frac{1}{5}\log\left(\frac{1}{6}*\frac{1}{5}\right) - \frac{1}{3}*\frac{1}{5}\log\left(\frac{1}{3}*\frac{1}{5}\right) - \frac{1}{2}*\frac{1}{5}\log\left(\frac{1}{2}*\frac{1}{5}\right)$$

$$-\frac{1}{8}*\frac{4}{5}\log\left(\frac{1}{8}*\frac{4}{5}\right) - \frac{1}{8}*\frac{4}{5}\log\left(\frac{1}{8}*\frac{4}{5}\right) - \frac{3}{4}*\frac{4}{5}\log\left(\frac{3}{4}*\frac{4}{5}\right) = 1.8628$$

$$\boxed{\begin{aligned} H(X|Y) &= 1.8628 - 0.7219 \\ &= 1.1409 \end{aligned}}$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- Sol. $H(X|Y)=?,\ I(X;Y)=?.$

  $I(X;Y)=H(X)+H(Y)-\ H(X,Y)=?$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- Sol. $H(X|Y)=?,\ I(X;Y)=?.$

  $I(X;Y)=H(X)+H(Y)-\ H(X,Y)=?$

  $H(X,Y)=1.8628, H(Y)=0.7219$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X = 0|Y = 0) = 1/6, P(X = 1|Y = 0) = 1/3, P(X = 2|Y = 0) = 1/2$
- $P(X = 0|Y = 1) = 1/8, P(X = 1|Y = 1) = 1/8, P(X = 2|Y = 1) = 3/4$
- $P(Y = 0) = 1/5,\ P(Y = 1) = 4/5$
- Sol. $H(X|Y) =?,\ I(X;Y) =?.$

  $I(X;Y) = H(X) + H(Y) -\ H(X,Y) =?$

  $H(X,Y) = 1.8628, H(Y) = 0.7219$

  $H(X) = -\Sigma_{x\in X}p(x)logp(x)$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X = 0|Y = 0) = 1/6, P(X = 1|Y = 0) = 1/3, P(X = 2|Y = 0) = 1/2$
- $P(X = 0|Y = 1) = 1/8, P(X = 1|Y = 1) = 1/8, P(X = 2|Y = 1) = 3/4$
- $P(Y = 0) = 1/5,\ P(Y = 1) = 4/5$
- Sol. $H(X|Y) = ?,\ I(X;Y) = ?.$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = ?$$

$$H(X,Y) = 1.8628, H(Y) = 0.7219$$

$$H(X) = -\Sigma_{x \in X} p(x) log p(x)$$

By total probability,

$$P(X = x) = \Sigma_{y \in Y} P(X = x|Y = y)P(Y = y)$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- Sol. $H(X|Y)=?,\ \ I(X;Y)=?.$

$$I(X;Y)=H(X)+H(Y)-\ H(X,Y)=?$$

$$H(X,Y)=1.8628, H(Y)=0.7219$$

$$H(X)=-\Sigma_{x\in X}p(x)logp(x)$$

By total probability,

$$P(X=x)=\Sigma_{y\in Y}P(X=x|Y=y)P(Y=y)$$

$$P(X=0)=\frac{1}{6}*\frac{1}{5}+\frac{1}{8}*\frac{4}{5}=\frac{2}{15}, P(X=1)=\frac{1}{3}*\frac{1}{5}+\frac{1}{8}*\frac{4}{5}=\frac{1}{6}, P(X=3)=\frac{1}{2}*\frac{1}{5}+\frac{3}{4}*\frac{4}{5}=\frac{7}{10}$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X = 0|Y = 0) = 1/6, P(X = 1|Y = 0) = 1/3, P(X = 2|Y = 0) = 1/2$
- $P(X = 0|Y = 1) = 1/8, P(X = 1|Y = 1) = 1/8, P(X = 2|Y = 1) = 3/4$
- $P(Y = 0) = 1/5,\ P(Y = 1) = 4/5$
- Sol. $H(X|Y) =?,\ \ I(X;Y) =?.$

$$I(X;Y) = H(X) + H(Y) - \ H(X,Y) =?$$

$$H(X,Y) = 1.8628, H(Y) = 0.7219$$

$$H(X) = -\Sigma_{x \in X} p(x) logp(x)$$

By total probability,

$$P(X = x) = \Sigma_{y \in Y} P(X = x|Y = y)P(Y = y)$$

$$P(X = 0) = \frac{1}{6} * \frac{1}{5} + \frac{1}{8} * \frac{4}{5} = \frac{2}{15}, P(X = 1) = \frac{1}{3} * \frac{1}{5} + \frac{1}{8} * \frac{4}{5} = \frac{1}{6}, P(X = 3) = \frac{1}{2} * \frac{1}{5} + \frac{3}{4} * \frac{4}{5} = \frac{7}{10}$$

$$H(X) = \ -\left[\frac{2}{15}\log\left(\frac{2}{15}\right) + \frac{1}{6}\log\left(\frac{1}{6}\right) + \frac{7}{10}\log\left(\frac{7}{10}\right) +\right] = 1.1786$$

Exercise

- X: random variable on dropping or not of a student
- Y: random variable on the department of a student
- X=0: Mar. drop, X=1: Apr. drop, X=2: No drop
- Y=0: CS, Y=1: EE
- $P(X=0|Y=0)=1/6, P(X=1|Y=0)=1/3, P(X=2|Y=0)=1/2$
- $P(X=0|Y=1)=1/8, P(X=1|Y=1)=1/8, P(X=2|Y=1)=3/4$
- $P(Y=0)=1/5,\ P(Y=1)=4/5$
- Sol. $H(X|Y)=?,\ I(X;Y)=?.$

$I(X;Y)=H(X)+H(Y)-\ H(X,Y)=?$

$H(X,Y)=1.8628, H(Y)=0.7219$

$H(X)=-\Sigma_{x\in X}p(x)logp(x)$

$$\boxed{\begin{aligned} I(X,Y) &= 1.1786+0.7219-1.8628 \\ &= 0.038 \end{aligned}}$$

By total probability,

$P(X=x)=\Sigma_{y\in Y}P(X=x|Y=y)P(Y=y)$

$P(X=0)=\frac{1}{6}*\frac{1}{5}+\frac{1}{8}*\frac{4}{5}=\frac{2}{15}, P(X=1)=\frac{1}{3}*\frac{1}{5}+\frac{1}{8}*\frac{4}{5}=\frac{1}{6}, P(X=3)=\frac{1}{2}*\frac{1}{5}+\frac{3}{4}*\frac{4}{5}=\frac{7}{10}$

$H(X)=\ -\left[\frac{2}{15}\log\left(\frac{2}{15}\right)+\frac{1}{6}\log\left(\frac{1}{6}\right)+\frac{7}{10}\log\left(\frac{7}{10}\right)+\right]=1.1786$

Summary

- Information
- Entropy
- Cross Entropy
- Error Backpropagation Learning
- Mutual Information
- Kullback Leibler Divergence
- Independent Component Analysis (ICA)
- Learning for ICA
- Blind Source Separation

Reference: Simon Haykin, *Neural Networks: A Comprehensive Foundation,* Prentice Hall