[0730] Math for ML

141 이건일

≠ 1-1

1) $H(x) = 6 \cdot (-\frac{1}{6}) \cdot \log_2 \frac{1}{6}$

$= \log_2 6$

≈ 2.585

2) $p_6 = 1$, 나머지 0

$H(x) = -1 \cdot \log_2 1 = 0$

3) $p_{짝수} = \frac{2}{9}$, $p_{홀수} = \frac{1}{9}$

$H(x) = 3\left(-\frac{1}{9}\log_2\frac{1}{9}\right) + 3\left(-\frac{2}{9}\log_2\frac{2}{9}\right)$

≈ 2.502

4) $H(x) = -0.5\log_2 0.5 + 5(-0.1\log_2 0.1)$

≈ 2.16

$H(x)$는 1)의 경우 최대가 된다.

모든 사건의 확률이 균등하면 예측이 어려워지기 때문에 엔트로피가 최대가 된다.

#1-2

1) $H(x) \geq 0$

$p(x) \in [0,1]$

$\log_2 1 = 0$, $p(x)$가 1보다 작을 때에는 음수값.

$-p(x)\log_2 p(x) \geq 0$

모든 항이 0 이상이므로 총합인 $H(x)$도 0 이상.

2) $H(x) \leq \log_2 |x|$

$|x| = n$ (가능한 경우의 수)

$H(x) = \mathbb{E}[-\log_2 p(x)]$

$f(p) = -\log p$ → convex function

by Jensen's inequality

$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$

평균이 아닌 가중합 적용

$H(x) = \sum p_i \cdot \log \frac{1}{p_i} \leq \log n$

$H(x) \leq \log |x|$

* Bonus : Jensen's inequality 증명

$n=2$, 가정 : f는 convex, $\lambda \in [0,1]$

$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ ⇒ convex function의 정의, $n=2$ 일때 성립

$n=k$일때, Jensen's inequality 성립한다는 가정

$f\left(\sum_{i=1}^{k} \alpha_i x_i\right) \leq \sum_{i=1}^{k} \alpha_i f(x_i)$, where $\sum_{i=1}^{k} \alpha_i = 1$

$\alpha_1 + \alpha_2 + \cdots + \alpha_{k+1} = 1$ 이고 $\alpha_i \geq 0$ 일 때,

$f\left(\sum_{i=1}^{k+1} \alpha_i x_i\right) \leq \sum_{i=1}^{k+1} \alpha_i f(x_i)$ 를 보이자. ⊥

$S = \sum_{i=1}^{k} \alpha_i$, $\beta_i = \frac{\alpha_i}{S}$ $(i = 1 \sim k)$ (앞의 k개 하나로 묶기)

$S \in [0,1)$, $\sum_{i=1}^{k} \beta_i = 1$

이때, $\sum_{i=1}^{k+1} \alpha_i x_i = S \cdot \sum_{i=1}^{k} \beta_i x_i + \alpha_{k+1} x_{k+1}$

⇒ $S + \alpha_{k+1} = 1$ convex 결합

1-3

1) $H(X) = -\sum p(x)\log_2 p(x)$

$= -[0.6\log_2 0.6 + 0.4\log_2 0.4] \approx 0.971$

$H(Y) = -\sum p(y)\log_2 p(y)$

$= -[0.55\log_2 0.55 + 0.45\log_2 0.45] \approx 0.992$

2) $H(X,Y) = -\sum p(x,y)\log_2 p(x,y)$

$= -[0.45\log_2 0.45 + 0.15\log_2 0.15 + 0.1\log_2 0.1 + 0.3\log_2 0.3] \approx 1.846$

3) $H(X|Y) = \sum_y p(y) H(X|Y=y)$

$P(X=0|Y=0) = \frac{0.45}{0.55} \approx 0.8182$

$P(X=1|Y=0) = \frac{0.1}{0.55} \approx 0.1818$

$P(X=0|Y=1) = \frac{0.15}{0.45} \approx 0.333$

$P(X=1|Y=1) = \frac{0.3}{0.45} \approx 0.666$

$H(X|Y=0) = -[0.8182\log_2 0.8182 + 0.1818\log_2 0.1818] \approx 0.683$

$H(X|Y=1) = -[0.333\log_2 0.333 + 0.666\log_2 0.666] \approx 0.918$

$H(X|Y) = 0.55 \cdot 0.683 + 0.45 \cdot 0.918 \approx \underline{0.786}$

4) $I(X;Y) = H(X) - H(X|Y)$

$= 0.971 - 0.786 = \underline{0.185}$

5) $I(X;Y)$는 Y를 알게 되었을 때 X에 대한 불확실성이 얼마나 줄어들었는지에 대한 값이다. 4)의 경우, 동전 여부를 알면 파업 발령에 대해 약 0.185 비트만큼 더 잘 알 수 있음을 의미한다.

$I(X;Y) = 0$ 이라면, X와 Y는 서로 독립이며 Y를 알아도 X를 예측하는데 아무 도움이 되지 않는다.

$I(X;Y) = H(X)$라면, Y만 알게되면 X를 완벽히 예측할 수 있다는 뜻이다.

귀납가정 : $\sum_{i=1}^{k} \beta_i f(x_i) \geq f\left(\sum_{i=1}^{k} \beta_i x_i\right)$

Convexity 정의

$$\begin{aligned} f\left(\sum_{i=1}^{k+1} \alpha_i x_i\right) &= f\left(S \sum_{i=1}^{k} \beta_i x_i + \alpha_{k+1} x_{k+1}\right) \\ &\leq S f\left(\sum_{i=1}^{k} \beta_i x_i\right) + \alpha_{k+1} f(x_{k+1}) \\ &\leq S \sum_{i=1}^{k} \beta_i f(x_i) + \alpha_{k+1} f(x_{k+1}) \\ &= \sum_{i=1}^{k} \alpha_i f(x_i) + \alpha_{k+1} f(x_{k+1}) \end{aligned}$$

$\Rightarrow f\left(\sum_{i=1}^{k+1} \alpha_i x_i\right) \leq \sum_{i=1}^{k+1} \alpha_i f(x_i)$

$\therefore f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \leq \sum_{i=1}^{n} \alpha_i f(x_i)$

$n \in \mathbb{N}$, $\alpha_i \geq 0$, $\sum \alpha_i = 1$ 에서 성립.

#2-1

$A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 1 \\ -1 & 0 & 1 \end{bmatrix}$ $\quad A-\lambda I = \begin{bmatrix} 2-\lambda & 0 & 0 \\ 1 & 2-\lambda & 1 \\ -1 & 0 & 1-\lambda \end{bmatrix}$

$\det(A-\lambda I) = 0$

$= (2-\lambda) \cdot \det \begin{bmatrix} 2-\lambda & 1 \\ 0 & 1-\lambda \end{bmatrix}$

$= (2-\lambda)\left[(2-\lambda)(1-\lambda) - 0\right]$

$= (2-\lambda)^2(1-\lambda)$

i). $\lambda = 1$

$A-I = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{bmatrix}$ $\quad$ RREF $\Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

x_3 = free variable t

$x_2 = -t$ $\quad \rightarrow \quad V_1 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} t$

$x_1 = 0$

ii). $\lambda = 2$

$A-2I = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}$ $\quad$ RREF $\Rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

x_2 = free variable s

x_3 = free variable t $\quad \rightarrow \quad V_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} t + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} s$

$x_1 = -t$

$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$, $\quad P = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

#2-2

직교대각화 → 대칭행렬

$\exists$ 직교행렬 P s.t. $P^TAP = D$ ($\exists$ 대각행렬 D)

$\Rightarrow A = PDP^T$

$\Rightarrow A^T = (P^T)^T D^T P^T = PDP^T = A$

대칭행렬 → 직교대각화

- 실수 대칭행렬 A의 모든 고유값은 실수.
- 서로 다른 고유값 $\lambda_i \neq \lambda_j$에 대응하는 고유벡터 v_i, v_j는 항상 서로 직교.
  $v_i^T v_j = 0$
- 대칭행렬 A는 대각화 가능
  → n개의 고유벡터 가짐
  고유벡터를 통해 GS 정규직교화로 직교 기저 만들 수 있음.

$P = [v_1, v_2, \cdots, v_n] \in \mathbb{R}^{n \times n}$

P는 직교행렬, $P^TP = I$

$Av_i = \lambda_i v_i = AP = PD \Rightarrow \underline{P^TAP = D}$ $\Rightarrow A = PDP^T$

(D는 고유값들이 대각인 대각행렬)

If $A = A^T$, then $\exists$ orthogonal P, diagonal D such that $A = PDP^T$

2-3

$$B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$M = B^T B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\lambda I - M = \begin{bmatrix} \lambda-1 & -1 & 0 \\ -1 & \lambda-1 & 0 \\ 0 & 0 & \lambda-1 \end{bmatrix}$$

$$\det(\lambda I - M) = 0$$
$$= (\lambda-1)\left[(\lambda-1)^2 - 1\right]$$
$$= (\lambda-1)\left[(\lambda-1+1)(\lambda-1-1)\right]$$
$$= (\lambda-1)(\lambda)(\lambda-2)$$

$$\lambda = 2, 1, 0$$

i) $\lambda_1 = 2$, $\sigma_1 = \sqrt{2}$

$$2I - M = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{RREF} \Rightarrow \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$x_2 = t$
$x_1 = t$
$x_3 = 0$

$$\vec{w}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} t \qquad \vec{v}_1 = \frac{w_1}{\|w_1\|} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

ii) $\lambda_2 = 1$, $\sigma_2 = 1$

$$I - M = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{RREF} \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$x_3 = t$
$x_2 = 0$
$x_1 = 0$

$$\vec{w}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} t \qquad \vec{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

iii) $\lambda_3 = 0$, $\sigma_3 = 0$

$$-M = \begin{bmatrix} -1 & -1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \text{RREF} \Rightarrow \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$x_2 = t$
$x_3 = 0$
$x_1 = -t$

$$\vec{w}_3 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} t \qquad \vec{v}_3 = \frac{\vec{w}_3}{\|\vec{w}_3\|} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$V^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}$$

$$U = BB^T = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

3-1

Sample 하나에 대한 $l(w)$

$$l(w) = -y \log \hat{y} - (1-y)\log(1-\hat{y}) \quad \text{where} \quad \hat{y} = \frac{1}{1+e^{-w^T x}} = \sigma(w^T x)$$

$$z = w^T x \rightarrow \frac{dz}{dw} = x$$

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \rightarrow \frac{d\hat{y}}{dz} = \hat{y}(1-\hat{y})$$

$$l(w) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$$

$$\rightarrow \frac{dl}{d\hat{y}} = -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} = \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$\frac{dl}{dw} = \frac{dl}{d\hat{y}} \cdot \frac{d\hat{y}}{dz} \cdot \frac{dz}{dw} = \left(\frac{\hat{y}-y}{\hat{y}(1-\hat{y})}\right) \cdot (\hat{y}(1-\hat{y})) \cdot x = (\hat{y}-y)x$$

$$\nabla_w^2 l(w) = \hat{y}(1-\hat{y}) x x^T$$

logistic function의 출력은 (0,1) 사이이므로 항상 $\hat{y}(1-\hat{y}) > 0$

$x x^T$ 는 항상 양의 준정부호

$\nabla_w^2 l(w) \geq 0$ ⇒ Convex function

전체 목적함수는 이들의 합이기 때문에 convex하다.

#3-2

$x_{t+1} = x_t - \gamma \cdot f'(x_t)$

$x_1 = x_0 - \gamma \cdot f'(x_0)$

$= 1 - 0.2\{4(1)^3 - 6(1)^2 - 6(1) + 1\}$

$= 2.4$

$x_2 = x_1 - \gamma \cdot f'(x_1)$

$= 2.4 - 0.2\{4(2.4)^3 - 6(2.4)^2 - 6(2.4) + 1\}$

$= 0.9328$

$x_2 = 0.9328$

#3-3

1) 제약조건 $g(x,y) = x+y-2 = 0$ 설정

$L(x, y, \lambda) = x^2 + y^2 + \lambda(x+y-2)$

2) $\frac{\partial L}{\partial x} = 2x + \lambda = 0$

$\frac{\partial L}{\partial y} = 2y + \lambda = 0$

$\frac{\partial L}{\partial \lambda} = x + y - 2 = 0$

3) $2x + \lambda = 0 \cdots ①$

$2y + \lambda = 0 \cdots ②$

$x + y - 2 = 0 \cdots ③$

①, ②

$2x = 2y$

$x = y$

→ ③

$2x - 2 = 0$

$x = 1, y = 1$

→ ①

$2 + \lambda = 0$

$\lambda = -2$

$(x, y, \lambda) = (1, 1, -2)$

4) $f(x,y) = x^2 + y^2$

$f(1,1) = 1 + 1 = 2$