1-1 ✍️: 주사위를 던졌을 때 나오는 눈의 사건을 X 라 하고, 각 눈이 나올 확률을 p_i $(0 \le p_i \le 1,$ $\sum_{i=1}^{6} p_i = 1)$ 이라 할 때, 아래의 조건에서 주사위의 엔트로피 $H(X)$를 각각 구하세요. $H(X)$는 언제 최대가 되는지 간단한 이유 (내지는 직관)와 함께 서술하세요.

💡 엔트로피의 정의: $H(X) = \sum_{x \in X} p(x) \log_2\{1/p(x)\} = \sum_{x \in X} -p(x) \log_2\{p(x)\}$

1) 공정한 주사위 (모든 $p_i = 1/6$)

2) 모든 눈이 6 인 주사위

$2x$ x $\rightarrow$ $9x = 1$. $x = \frac{1}{9}$

3) 짝수 눈이 나올 확률이 홀수 눈이 나올 확률의 2 배인 주사위(짝수 각각 및 홀수 각각의 확률은 동일)

4) 1 이 나올 확률이 0.5, 나머지 눈이 나올 확률이 각각 0.1 인 주사위

1) $H(X) = 6 \times \frac{1}{6} \log_2 6 = \log_2 6 \approx 2.6$

2) $H(X) = 1 \log_2 1 = 0$

3) $H(X) = -3 \times \frac{1}{9} \log_2 \frac{1}{9} - 3 \times \frac{2}{9} \log_2 \frac{2}{9}$

$= -\frac{1}{3} \log_2 \frac{1}{9} - \frac{2}{3} \log_2 \frac{2}{9}$

$= -\frac{2}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{9}$

$= -\frac{2}{3} \log_2 \frac{1}{3} \cdot \frac{2}{9}$

$= -\frac{2}{3} \log_2 \frac{2}{27} \approx 2.5$

4) $H(X) = -0.5 \log_2 0.5 - 5 \times 0.1 \log_2 0.1$

$= \frac{1}{2} \log_2 2 - \frac{1}{2} \log_2 \frac{1}{10}$

$= \frac{1}{2} \log_2 2 \times 10$

$= \frac{1}{2} \log_2 20 \approx 2.2$

엔트로피는 불확실성의 척도.

가능한 결과들의 확률이 고르게 분포할수록 불확실성이 크고, 엔트로피가 크다.

1-2 ✍️: 임의의 이산확률변수 X 에 대하여 X 가 가질 수 있는 값의 집합을 $\mathcal{X}$라 합시다. 예를 들어, 동전을 1 회 던졌을 때의 $\mathcal{X} = \{Head, Tail\}$ 로 $|\mathcal{X}| = 2$입니다. 이때 엔트로피 $H(X)$가 가질 수 있는 값의 범위가 $0 \le H(X) \le \log_2 |\mathcal{X}|$임을 증명하세요.

> 💡 좌측 부등식의 경우, 확률 값의 성질을 활용해 증명 가능합니다.
> 💡 우측 부등식의 경우, 아래의 Jensen's Inequality 를 활용해 증명 가능합니다.
> - Jensen's Inequality: $\mathbb{E}[f(X)] \ge f(\mathbb{E}[X])$ for a convex f, and $\mathbb{E}[f(X)] \le f(\mathbb{E}[X])$ for a concave f
> (Bonus Question: Jensen's Inequality 는 어떻게 증명할까요? Hint: 수학적 귀납법으로 n=2 부터...!)

1) $0 \le H(X)$

$0 \le p(x) \le 1$. $\log_2 p(x) \le 0$

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

$H(X) \ge 0$

2) $H(X) \le \log_2 |X|$

$$H(X) = -\sum p(x)\log_2 p(x)$$
$$= \sum p(x)\log_2 \frac{1}{p(x)}$$
$$= E\left[\log_2 \frac{1}{p(x)}\right] \le \log_2\left(E\left[\frac{1}{p(x)}\right]\right)$$
$$= \log_2 n$$
$$= \log_2 |\mathcal{X}|$$

$\therefore\ 0 \le H(X) \le \log_2 |\mathcal{X}|$

1-3 ✍️: 어떤 지역 주민들의 흡연 여부 (X)와 폐암 발병 여부 (X)에 대한 설문 조사를 실시하여 다음과 같은 결합 확률 분포를 얻었다고 가정합니다. 여기서 X = 1 은 폐암 발병, X = 0 은 폐암 미발명을 의미하며, Y = 1 은 흡연, Y = 0 은 비흡연을 의미합니다. 아래 질문에 답해주세요.

P(X,Y)	Y = 0 (비흡연)	Y = 1 (흡연)
X = 0 (폐암 미발병)	0.45	0.15
X = 1 (폐암 발병)	0.1	0.3

1) 폐암 발병 여부 X 와 흡연 여부 Y 의 엔트로피 H(X), H(Y)를 계산하시오.

2) X 와 Y 의 결합 엔트로피 H(X, Y)를 계산하시오.

3) 흡연 여부 Y 에 대한 정보가 주어졌을 때, 폐암 발병 여부 X 의 조건부 엔트로피 H(X | Y)를 계산하시오.

4) 폐암 발병 여부 X 와 흡연 여부 Y 의 Mutual Information I(X ; Y)을 계산하시오.

5) 4)의 결과로 나온 I(X ; Y)의 의미를 통계적 관점에서 간략하게 해석하고,
I(X ; Y)값이 0 일 경우와, H(X)와 같을 경우 각각 어떤 의미를 갖는지 설명하시오.

💡 엔트로피 계산 시 log 의 밑은 2 로 계산하면 됩니다.
💡 단위는 비트(bits)를 이용하며, log 계산 값은 근사치로만 이용하시면 됩니다.

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

$I(X;Y) > 0$ → Y는 X에 대한 정보를 제공

$I(X;Y) = 0$ → X와 Y는 독립

$I(X;Y) = H(X)$ → Y를 알면 X는 예측 가능

1) $H(X) = -\Sigma P(X) \log_2 P(X)$

$= -0.6 \log_2 0.6 - 0.4 \log_2 0.4$

≈ 0.97

$H(Y) = -\Sigma P(y) \log_2 P(y)$

$= -0.55 \log_2 0.55 - 0.45 \log_2 0.45$

≈ 0.99

2) $H(X.Y) = -\Sigma P(x,y) \log_2 P(x.y)$

$= -0.45 \log_2 0.45 - 0.15 \log_2 0.15 - 0.1 \log_2 0.1 - 0.3 \log_2 0.3$

≈ 1.8

3) $H(X|Y) = \Sigma P(y) H(X|y)$

$H(X|y=0) = -0.818 \log_2 0.818 - 0.182 \log_2 0.182$

≈ 0.68

$H(X|y=1) = -0.333 \log_2 0.333 - 0.667 \log_2 0.667$

≈ 0.92

$H(X|Y) \approx 0.55 \cdot 0.68 + 0.45 \cdot 0.92$

≈ 0.79

4) $I(X;Y) = H(X) - H(X|Y) \approx 0.97 - 0.79 \approx 0.18$

2-1 ✍️: 대각화 (Diagonalization)의 정의와 계산 과정은 다음과 같습니다.

💡 정의: 임의의 정방행렬 A 를 고유값과 고유벡터를 통해 대각행렬의 형태로 선형 변환하는 과정이며, 이를 수식으로 표현하면 $D = P^{-1}AP$를 만족하는 대각행렬 D 를 찾는 것입니다. 만약 A 가 n 개의 독립인 고유벡터를 가지고 있다면 (= 가역이라면) A 는 대각화 가능하며, 이때 대각행렬의 원소는 A 의 고유값으로 구성됩니다.

💡 과정:

1) A 에 대한 고유값과 이에 대응하는 고유벡터를 찾은 후 각각 $p_1, p_2, \dots, p_n$으로 놓습니다.
2) 고유벡터로 구성된 행렬 $P = [p_1, p_2, \dots, p_n]$을 만들고 $D = P^{-1}AP$을 구합니다.

다음 정방행렬을 대각화했을 때 나오는 대각행렬 D 와 고유벡터 행렬 P 를 각각 구하세요.

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

$$\begin{bmatrix} 2-\lambda & 0 & 0 \\ 1 & 2-\lambda & 1 \\ -1 & 0 & 1-\lambda \end{bmatrix}$$

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

$$\begin{matrix} 0 & 0 & -1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{matrix}$$

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

2-3 ✍: 특이값 분해 (SVD, Singular Value Decomposition)은 대각화와 달리 모든 크기의 행렬에 대해 적용이 가능하며, 그 계산 과정은 다음과 같습니다.

> 💡 1) 항상 대칭 행렬을 이루는 두 행렬 B^TB, BB^T를 계산하고, 이들의 고유값을 계산해 특이값 σ와 고유벡터를 통해 직교대각화하여 $B^TB = VDV^T, BB^T = UDU^T$를 구합니다.
> 2) 0 이 아닌 특이값들을 내림차순으로 나열하여 Σ를 구성하고, 이들을 모두 활용해 $A = U\Sigma V^T$를 구합니다.
>
> 계산 방법에 대한 구체적인 과정은 다음의 영상 자료 (링크)를 반드시 참고하여 문제를 풀어주세요.

다음 행렬에 특이값 분해를 적용하였을 때 나오는 행렬 U, Σ, V^T를 각각 구하세요.

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

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

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

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & 0 \end{bmatrix}$$

3-1 ✍: 정규세션에서 살펴본 것과 같이, Logistic Regression 의 목적함수는 아래와 같습니다.

$$\min_{w} \sum_{i=1}^{m} \{-y^{(i)} \log(\hat{y}^{(i)}) - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})\}$$

최적해를 구하기 위해서는 먼저 내부의 목적 함수가 convex function 인지 여부를 판단해야 합니다. 내부 목적 함수 $-y^{(i)} \log(\hat{y}^{(i)}) - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$가 convex function 임을 보이세요.

- 💡 $y^{(i)}$ 을 Logistic Function 식을 활용하면 $\hat{y}^{(i)} = \frac{1}{1+e^{-w^T x^{(i)}}}$과 같이 표현할 수 있습니다.
- 💡 목적 함수를 두 부분으로 나누어서 각각 Convex 하다는 것을 증명하면 됩니다.
- 💡 convexity 를 증명하기 위해서는 w 에 대해 2 번 미분하여, 0 보다 크다는 것을 보이면 됩니다.

$*\quad -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \,.\quad \hat{y} = \sigma(W^T x) = \frac{1}{1+e^{-W^T x}}$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$\frac{\partial \sigma}{dz} = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z)(1-\sigma(z)) \geq 0$$

$$l(z) = -y\log\sigma(z) - (1-y)\log(1-\sigma(z))$$

$$\frac{dl}{dz} = -y \cdot \frac{1}{\sigma(z)} \cdot \frac{\partial\sigma}{dz} + (1-y) \cdot \frac{1}{1-\sigma(z)} \cdot \frac{\partial\sigma}{dz}$$

$$= -y \cdot (1-\sigma(z)) + (1-y) \cdot \sigma(z)$$

$$= -y + \sigma(z)$$

$*\quad \nabla_W l(W) = \frac{dl}{dz} \cdot \frac{dz}{dW} = (\sigma(z) - y) \cdot x$

$$\nabla^2_W l(W) = \frac{d}{dW}((\sigma(z) - y) \cdot x)$$

$$= \sigma(z)(1-\sigma(z)) \cdot xx^T \geq 0$$

$\therefore$ convex.

3-2 ✍️: Minimization 에 대한 Gradient Descent Algorithm 은 아래와 같습니다.

$$\min_{x \in X} f(x)$$

$$x_{t+1} = x_t - \gamma \nabla_x f(x_t)$$

$\nabla_x f(x_t)$: $f(x_t)$ 에 대해서 $\mathrm{x_t}$ 지점에서 미분한 값

다음과 같은 Loss function 과 시작점에서 t = 2 일 때 나오는 값, 즉 $\mathrm{x_2}$의 값을 구하세요.

$$f(x) = x^4 - 2x^3 - 3x^2 + x$$

$$x_0 = 1, \qquad \gamma = 0.2$$

$f'(x) = 4x^3 - 6x^2 - 6x + 1$

$x_1 = x_0 - \gamma \cdot (4x_0^3 - 6x_0^2 - 6x_0 + 1)$

$= 1 - 0.2(4 - 6 - 6 + 1)$

$= 1 + 0.2 \times 7$

$= 2.4$

$x_2 = x_1 - \gamma(4x_1^3 - 6x_1^2 - 6x_1 + 1)$

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

$= 0.93$

$\therefore 0.93$

3-3 ✍️: 라그랑주 승수법은 아래와 같습니다.

💡 라그랑주 승수(Lagrange Multiplier)는 제약 조건이 있는 최적화 문제를 제약 조건이 없는 문제로 변환하여 최적해를 찾기 위한 보조 변수를 말합니다. 제약 조건이 없을 때는 단순히 ▽f(x)=0 으로 최적점을 찾을 수 있지만, 제약조건 g(x)가 있을 때는 최적해는 제약을 만족하면서 f(x)를 최소화하는 지점이 되어야 합니다. 이를 위해 제약 조건과 목적함수를 결합한 새로운 함수 (lagrangian)을 만들고, 그 stationary point 를 찾아서 목적함수 최적화와 제약 만족을 동시에 달성할 수 있습니다.

제약 조건이 있는 최적화 문제:

$$\min_x f(x) \quad \text{subject to } g(x) = 0$$

→ 이를 다음과 같은 **Lagrangian 함수**로 바꿉니다:

$$L(x, \lambda) = f(x) + \lambda \cdot g(x)$$

25-2 Mathematics for ML #2 세션 자료 中

사진의 λ가 라그랑주 승수입니다. 제약조건을 penalty 처럼 반영하여 이를 만족시키는 해를 찾아낼 수 있습니다.

다음 조건을 만족하는 $f(x,y)$의 최솟값을 라그랑주 승수법을 이용해 구하세요.

$$\min f(x,y) = x^2 + y^2 \quad s.t. \ \ x + y \geq 2 \rightarrow 2 - x - y \leq 0$$

1) 주어진 목적 함수와 제약 조건을 이용하여 lagrangian $L(x,y,\lambda)$을 설정하세요.

2) $L(x,y,\lambda)$을 각 변수에 대해 편미분하고, 그 값이 0 이 되는 stationary point 조건을 제시하세요.

3) 위에서 얻은 조건들을 만족하는 x, y, λ 의 값을 계산하세요.

4) 구한 $f(x,y)$값을 목적함수 $f(x,y)$에 대입하여 최솟값을 계산하세요.

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

2) $\nabla_x L(x.y.\lambda) = 2x - \lambda$

$\nabla_y L(x.y.\lambda) = 2y - \lambda$

$\nabla_\lambda L(x.y.\lambda) = 2 - x - y$

$x = \frac{\lambda}{2}. \ y = \frac{\lambda}{2} \rightarrow 2 - \lambda = 0 \quad \therefore \lambda = 2$

3) $x = 1. \ y = 1. \ \lambda = 2$

4) $L(x.y.\lambda) = 1 + 1 = 2$