

25-2 DSL 정규 세션

Mathematics for ML 과제

- 본 과제는 학회 정규 세션 「Mathematics for Machine Learning」의 내용을 다루며, 개념의 적용과 실제 활용 사례에 대한 이해를 돋기 위해 기획되었습니다. 해당 과제는 평가를 위한 것이 아니므로, 주어진 힌트(💡)를 적극 활용하시고 학회원 간 토론 및 Slack 질의응답을 적극 활용하여 해결해주십시오. 단, 답안 표절이나 LLM 의 남용은 금지합니다.
- 각 문제에서 요구하는 방법에 맞게 해결하며, 서술형 문제들은 따로 작성하시어 .pdf 파일로 답안을 작성하여 제출해 주십시오.
- 7/30 (수) 23 시 59 분까지 Github 에 .pdf 파일을 제출해 주십시오. Github 에 제출하는 방법을 모른다면 학술부 (강승우, 조지성, 한연주)로 문의 주시기 바랍니다.

문제 1 Entropy and Mutual Information

다음은 어느 드라마의 한 장면 중 일부입니다. [연재기자2](#)

“…지는 분은 물론 벌칙이 있습니다. 아마 영화에서 보신 적이 있으실 겁니다, 러시안룰렛이라고. 이 권총에 총알 하나를 넣고 닫습니다. 그리고 지는 분 쪽으로 방아쇠를 당길 겁니다. 죽을 확률이 1/6, 살 확률은 5/6. 생각보다 나쁘지 않죠? (중략) …조금 지루한 감이 없지 않아 있네요. 그럼 이제 확률을 뒤집어 볼까요? 권총에 총알 다섯 개를 넣고 닫습니다. 확률은 1/6, 죽을 확률은 5/6. 자, 다시 시작하겠습니다.”

그러나 사실은 두 상황 모두 엔트로피의 측면에서의 지루함, 즉 불확실성은 같습니다! 총알이 발사될 확률이 각각 1/6, 5/6 인 베르누이 분포를 따르기 때문에, 엔트로피의 정의를 따라 계산하면 같은 값이 나오는 것이죠.

이 문제에서는 이처럼 주어진 사건들이 이산적인 경우 엔트로피의 성질을 알아보고, 직접 계산해보며 엔트로피와 Mutual Information 이 가진 의미를 알아보자 합니다.

1-1 💡: 주사위를 던졌을 때 나오는 눈의 사건을 X 라 하고, 각 눈이 나올 확률을 p_i ($0 \leq p_i \leq 1$, $\sum_{i=1}^6 p_i = 1$) 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

3) 짹수 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

1-1

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

$$2) H(x) = -1 \cdot \log_2 1 = 0$$

$$3) 3p + 3q = 1, p = 2q, p = \frac{2}{3}, q = \frac{1}{3}$$

$$H(x) = -\frac{2}{3} \cdot \log_2 \frac{2}{3} \times 3 + -\frac{1}{3} \cdot \log_2 \frac{1}{3} \times 3$$

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

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

$$4) H(x) = -0.5 \cdot \log_2 0.5 - 0.5 \times 0.1 \cdot \log_2 0.1$$

$$= 0.5 + 0.5 \log_2 10$$

$p = \frac{1}{n}$ 일 때 uniform 분포에서 엔트로피는 최대이며 $p = \frac{1}{2}$ 일 때 가장 크다.

1-2 🔥 : 임의의 이산확률변수 X 에 대하여 X 가 가질 수 있는 값의 집합을 \mathcal{X} 라 합시다. 예를 들어, 동전을 1회 던졌을 때의 $\mathcal{X} = \{\text{Head}, \text{Tail}\}$ 로 $|\mathcal{X}| = 2$ 입니다. 이때 엔트로피 $H(X)$ 가 가질 수 있는 값의 범위가 $0 \leq H(X) \leq \log_2 |\mathcal{X}|$ 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: $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ for a convex f , and $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$ for a concave f
 - (Bonus Question: Jensen's Inequality 는 어떻게 증명할까요? Hint: 수학적 귀납법으로 n=2 부터...!)

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 미발병을 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 2로 계산하면 됩니다.
- 💡 단위는 비트(bits)를 이용하며, \log 계산 값은 근사치로만 이용하시면 됩니다.

1-2

1) 좌측 부등식

$H(X) = E[-\log p(x)]$, $X \sim P$ 이고, P 는 확률 분포이면

$$0 \leq p(x) \leq 1 \Leftrightarrow -\infty < \log p(x) \leq 0$$

$$\Leftrightarrow 0 \leq E(-\log p(x)) \leq \infty$$

2) 우측 부등식

엔트로피는 어떤 확률 변수에 대한 불확실성이 수치이므로 모든 사건의 확률이 같은 uniform 분포 환경에서 최대화된다.

$X \sim \text{Unif}$ 면, $p(x) = \frac{1}{n}$ 이므로 X 가 같은 분포를 따를 때 $H(X)$ 값은

$$H(X) = -\sum_{k=1}^n p(x_k) \cdot \log p(x_k) = n \cdot \frac{1}{n} \cdot -\log \frac{1}{n} = \log_2 n = \log_2 |X|$$

$$\therefore 0 \leq H(X) \leq \log_2 |X|$$

I-3

1) $P(X=0)=0.6, P(X=1)=0.4, P(Y=0)=0.55, P(Y=1)=0.45$

$$H(X) = -P(X=0)\log_2 0.6 - P(X=1)\log_2 0.4$$

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

$$\approx 0.917$$

$$H(Y) = -P(Y=0)\log_2 0.55 - P(Y=1)\log_2 0.45$$

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

$$\approx 0.99$$

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

$$= -(0.45 \cdot \log_2 0.45 + 0.15 \cdot \log_2 0.25 \\ + 0.1 \cdot \log_2 0.1 + 0.3 \cdot \log_2 0.3)$$

$$\approx 1.178$$

$$3) H(X|\Gamma) = H(X) - H(\Gamma) \approx 1.78 - 0.99 \\ = 0.79$$

$$4) I(X;Y) = H(X) - H(Y|X) \approx 0.97 - 0.59 \\ = 0.38$$

5) $I(X;Y)=0$ 이면 X 와 Y 가 선형적, 비선형적으로 독립이고,

$I(X;Y)=H(X)$ 면 $X \subseteq Y$ 이다.

문제 2 SVD and Data Compression

선형대수학은 머신러닝과 딥러닝의 원리를 수학적으로 분석하는 과정에서 사용되는 핵심적인 내용 중 하나로, 그 중에서도 가장 중요한 개념은 바로 SVD (Singular Value Decomposition, 특이값 분해)입니다. SVD를 이해하기 위해 우선 고유값과 고유벡터를 활용한 대각화 (Diagonalization)에 대해 알아보겠습니다.

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.

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

2-2 🔥 : (Optional) 대칭행렬이 직교대각화(Orthogonal Diagonalization)가 가능하다는 것은 필요충분조건이며, 이는 대칭행렬의 중요한 성질 중 하나입니다. 이를 증명하시오.

💡 직교대각화의 정의는 다음과 같습니다.

'정사각행렬 A 가 주어져있을 때, $P^{-1}AP = P^TAP$ 가 대각행렬이 되게 만드는 P 가 존재한다면, A 를 직교대각화 가능하다고 하고, P 가 A 를 직교대각화 한다고 한다.'

💡 직교대각화 -> 대칭행렬을 증명하는 과정은 위의 정의를 이용하면 됩니다.

💡 대칭행렬 -> 직교대각화를 증명하기 위해서는 아래의 내용들을 증명하면 됩니다.

- 1) 모든 성분이 실수인 대칭행렬은 실수 고유값만을 갖는다. (해당 내용은 그냥 이용해도 무방)
- 2) $n \times n$ 행렬 A 가 직교대각화 가능하다는 말은 A 가 서로 정규직교하는 n 개의 고유벡터들을 가진다는 말과 동치이다. -> 즉, 서로 다른 고유값에 대응하는 고유벡터들은 항상 서로 직교한다.
- 3) 각 고유값에 대해 대수적 중복도와 기하적 중복도가 항상 일치한다.

2-1

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

A I

$$\begin{vmatrix} 2-\lambda & 0 & 0 \\ 1 & 2-\lambda & 1 \\ -1 & 0 & 1-\lambda \end{vmatrix} = (2-\lambda)^2(1-\lambda) \quad \lambda=1 \text{ or } 2$$

i) $\lambda=2$ 일 때,

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad x_2 + x_3 = 0, x_2 \text{는 자유변수}$$

$$p_1 = [1, 0, -1]', p_2 = [0, 1, 0]'$$

ii) $\lambda=1$ 일 때

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad x_1=0, x_2+x_3=0, p_3 = [0, 1, -1]'$$

$$p = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \end{bmatrix}, b = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

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 $B^T B, BB^T$ 를 계산하고, 이들의 고유값을 계산해 특이값 σ 와 고유벡터를 통해 직교대각화하여 $B^T B = V D V^T, BB^T = U D U^T$ 를 구합니다.
- 2) 0 이 아닌 특이값들을 내림차순으로 나열하여 Σ 를 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^T 를 각각 구하세요.

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

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

$$\det(BB^T - \lambda I) = \begin{vmatrix} 1-\lambda & 1 & 0 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{vmatrix} = (1-\lambda)^2(2-\lambda)$$

$$\lambda = 1 \text{ or } 2, \quad \sigma = 1 \text{ or } \sqrt{2} \quad v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$B^T B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\text{대칭행렬이므로 } \lambda = 2 \text{ or } 1, \quad u_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

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

문제 3 Optimization Theory

최적화는 다양한 분야에서 많이 쓰이는 개념이지만, 특히 머신러닝에서는 최적의 정답을 찾기 위해 필수적으로 사용되는 개념입니다. 정규세션으로 이를 간단하게 배워보았고, 직접 수식으로 문제를 풀어보며 해당 개념을 익히고자 합니다.

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 function임을 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 w에 대해 2 번 미분하여, 0 보다 크다는 것을 보이면 됩니다.

3-2 🔥: Minimization에 대한 Gradient Descent Algorithm은 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 x_t 지점에서 미분한 값

다음과 같은 Loss function과 시작점에서 t = 2 일 때 나오는 값, 즉 x_2 의 값을 구하세요.

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

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

3-1

$$L(w) = -y \log y - (1-y) \log(1-y).$$

$$\frac{\partial L(w)}{\partial w} = \frac{\partial L(w)}{\partial y} \cdot \frac{\partial y}{\partial w}$$

$$\frac{\partial L(w)}{\partial y} = -\frac{y}{y} + \frac{1-y}{(1-y)}, \quad \frac{\partial y}{\partial w} = y(1-y)x$$

$$\frac{\partial^2 L(w)}{\partial w^2} = (y - y)x, \quad \frac{\partial^2 L(w)}{\partial w^2} = y(1-y)x^2$$

$0 < y, 1-y < 1$ 이고, $x^2 > 0$ 이고 $\frac{\partial^2 L(w)}{\partial w^2} > 0, L(w)$ 은 convex

3-2

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

$$x_1 = x_0 - r f'(x_0) = 1 - 0.2 \cdot (-7) = 2.4$$

$$x_2 = x_1 - r f'(x_1) = 2.4 - 0.2 \cdot (1.236) = 0.9328$$

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 $\nabla f(x)=0$ 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 함수로 바꿉니다:

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

25-2 Mathematics for ML #2 세션 자료 中

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

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

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

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.

Reference

- 24-1 기초과제 2 (10 기 신재우)
- 24-2 기초과제 (11 기 김현진, 김정우)
- 25-1 Mathematics for ML 과제 (12 기 김민규)
- Information Theory for Data Science, Prof. Son, Fall 2024, Yonsei Univ.

Data Science Lab

담당자: 13 기 조지성
xiwang01@yonsei.ac.kr

3-3

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

$$2) \frac{\partial L}{\partial x} = 2x - \lambda, \quad \frac{\partial L}{\partial y} = 2y - \lambda, \quad \frac{\partial L}{\partial \lambda} = 2-x-y$$

$$\lambda = \frac{1}{2}, \quad y = \frac{1}{2}, \quad x+y=2$$

$$3) x^* = 1, \quad y^* = 1, \quad \lambda^* = 2$$

$$4) \min f(x,y) = 1^2 + 1^2 = 2$$