기수 : 11 이름 : 윤정수

24-1 DSL 정규 세션

기초과제 2 머신러닝을 위한 수학 + 전처리 기초

- ☑ 본 과제는「수리통계학(1)」,「수리통계학(2)」,「머신러닝을 위한 수학 (세션)」 일부에 상응하는 내용과 Python 의 Numpy, Pandas, Matplotlib 패키지 활용법의 복습을 돕기 위해 기획되었습니다. 평가를 위한 것이 아니므로, 주어진 힌트(👌)를 적극 활용하시고 학회원 간 토론, Slack 의 질의응답을 활용하시어 해결해주십시오. 단, 답안 표절은 금지합니다.
- ☑ 서술형 문제는 ✎ , 코딩 문제는 © 으로 표기가 되어 있습니다. 각 문제에서 요구하는 방법에 맞게 해결하며, 서술형 문제들은 따로 작성하시어 pdf 로 제출해주시고 코드 문제들은 ipynb 파일에 답안을 작성하시어 제출해주십시오.
- ☑ **1/28 (일) 23 시 59 분까지** Github 에 PDF 파일과 ipynb 파일을 모두 제출해주십시오. Github 에 제출하는 방법을 모른다면 학술부장 혹은 Slack 의 질의응답을 활용해주시오.
- ☑ 모든 파일은 이름을 [0128]elementary_2_JaeWooShin 형태의 이름으로 제출해주시길 바랍니다.

문제 1 전처리 기초

데이터 사이언스에서 전처리는 데이터를 분석이 가능한 형태 혹은 모델링에 적합한 형태로 변형을 시켜줍니다. 이 문제에서는 전처리를 직접 해볼 것이며, 전처리를 한 데이터를 바탕으로 시각화까지 할 것입니다.

1-1 © : .ipynb 파일을 참조해서 state.name 컬럼에서 결측값들을 모두 삭제해주시오. ✓

1-2 © : .ipynb 파일을 참조해서 가장 많은 투표 수를 얻은 5 명의 후보자들을 구하시오. ✓

1-3 © : .ipynb 파일을 참조해서 가장 많은 투표 수를 얻은 5 명의 후보자들에 대한 데이터를 시각화 해주시오. ✓

문제 2 최적화개론

최적화는 머신러닝 분야 이외에서도 많이 쓰이는 개념입니다. 하지만 머신러닝에서는 특히 많이 쓰이고 있으며, 최적의 정답을 찾기 위해서 필수적으로 사용되고 있습니다. 첫 문제의 경우 어려울 것으로 예상되기에 Optional 으로 두었으며 필수적으로 풀 필요는 없습니다. 하지만 GDA 의 경우 딥러닝에서 거의 항상 사용되기에 꼭 풀어보시길 바랍니다.

2-1 ✎ : (Optional) Logistic Regression 의 Objective Function 은 다음과 같습니다.

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

이것의 Convex Optimization 과정을 서술하시오.

- Objective Function 에서의 함수인 Loss 함수가 Convex 하다는 것을 먼저 보여야 합니다.
- 위의 Convexity 를 증명했다면 이러한 Optimization 을 어떤 방법을 사용해서 풀 수 있는지 알아야합니다.

2-2 ✎ : Minimization 에 대한 Gradient Descent Algorithm (GDA) 는 다음과 같습니다

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

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

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

다음과 같은 로스 함수와 시작점에서 $t = 2$ 일 때에 나오는 값, 즉 x_2 의 값을 구하시오.

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

$$x_0 = 1, \qquad \eta = 0.1$$

- x_1 의 값은 다음과 같이 구할 수 있습니다.

$$f(x) = x^4 - 2x^3 - 3x^2 + x, \quad \nabla_x f(x) = 4x^3 - 6x^2 - 6x + 1$$

$$x_0 = 1, \quad \eta = 0.1$$

$$x_1 = x_0 - \eta \cdot \nabla_x f(x_0) = 1 - 0.1 \cdot (4x_0^3 - 6x_0^2 - 6x_0 + 1)$$

$$= 1 - 0.1 \cdot (4 \cdot 1^3 - 6 \cdot 1^2 - 6 \cdot 1 + 1) = 1 - 0.1 \cdot (4 - 6 - 6 + 1) = 1 - 0.1 \cdot (-7)$$

$$= 1 + 0.7 = 1.7 = x_1$$

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

$\nabla_x f(x) = 4x^3 - 6x^2 - 6x + 1.$ $\quad x_0 = 1,\ \eta = 0.1$

$x_2 = x_1 - \eta \cdot \nabla_x f(x_1) = 1.7 - 0.1 \cdot (4x_1^3 - 6x_1^2 - 6x_1 + 1)$

$1.7 - 0.1(4 \cdot 1.7^3 - 6 \cdot 1.7^2 - 6 \cdot 1.7 + 1) = 2.3888 = x_2$

문제 3 | 정보이론 기초

정보이론은 데이터의 전송, 압축, 저장, 해석 등과 관련된 기본적인 원리와 개념을 제공합니다. 해당 문제들은 머신러닝을 위한 수학 세션에서 배웠던 Entropy 와 딥러닝을 배우면서 자주보게 될 KL-Divergence 를 활용한 문제들입니다. 난이도가 어려울 것으로 예상돼서 틀려도 괜찮다는 말씀을 드리고 싶습니다.

3-1 ✎ **:** Entropy 의 식은 다음과 같습니다.

$$H(x) = -\sum_{x \in X} P(x) \log_2 P(x)$$

$P(x)$: Probability of Event X

spam.csv 데이터를 .ipynb 파일에서 관찰하고 데이터의 Entropy 를 구하시오.

- 데이터를 살펴보면 이항분포를 따른다는 것을 알 수가 있습니다. 시그마에서 $x \in X$, 즉 모든 X 에 대해서 더해야하기 때문에 $X = Spam$, $X = Not\ Spam$ 일 때를 둘다 포함시켜야 합니다.

✓

3-2 ✎ **:** KL-Divergence 의 정의는 다음과 같습니다.

$$KL(P \,||\, Q) = \sum_{x \in X} P(x)\ \log\left(\frac{P(x)}{Q(x)}\right)$$

spam_other.csv 데이터가 있을 때에, 이것의 분포와 spam.csv 의 분포를 비교하고 싶습니다. P 를 spam.csv 으로, Q 를 spam_other.csv 으로 두며 두 분포의 차이를 KL-Divergence 를 이용해서 차이를 손으로 계산해주세요.

- 주의점으로 여기에서의 로그는 베이스가 2 가 아닌 e 입니다.
- 최종적으로 $KL(P||Q)$ 를 구하면 됩니다.

✓

3-3 ✎ **:** KL-Divergence 은 Asymmetric 한 특징을 가지고 있습니다. 이것을 증명해주시오.

- KL-Divergence 가 Symmetric 하다는 말은 $KL(P \,||\, Q)$ 와 $KL(Q \,||\, P)$ 가 같은 값을 가져야 합니다. 즉
$$KL(P \,||\, Q) = KL(Q \,||\, P)$$
가 맞는지 아닌지를 확인해주시면 됩니다.
- Asymmetric 하다는 것은 거리 (Distance Metric) 이 아니라는 뜻입니다. 그렇기 때문에 보통 KL-Divergence 를 바로 거리 (Distance Metric) 을 사용하지는 않습니다. (예: JSD, Wasserstein Distance)

$KL(P||Q) = H(P||Q) - H(P) \neq H(P||Q) - H(Q) = KL(Q||P)$

문제 4 수리통계학

통계학에 있어서 수리통계학은 굉장히 중요합니다. 데이터를 해석하고 예측하는데 필수적 수학적 기반들을 제공해줍니다. 해당 문제들은 응용통계학과의 수리통계학 (1), (2) 강의에 자주 나오는 문제들입니다. 첫번째 문제는 생성에서도 굉장히 자주 사용되는 베이즈 정리이며, 두번째 문제는 우도함수를 최대화시키는 추정량을 찾는 문제입니다.

4-1 ✎ : 베이즈 정리는 다음과 같이 정의가 됩니다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

다음과 같은 정보들이 주어졌습니다.

- 연세대학교 학생이 감기에 걸릴 확률이 10% 입니다.
- DSL 학회원 중에서 감기에 걸릴 확률은 1% 입니다.
- DSL 학회원은 연세대학교 학생 중에서 0.1% 입니다.
- 모든 DSL 학회원은 연세대학교 학생입니다.

베이즈 정리를 활용해서 DSL 학회원이 아닌 교내생이 감기에 걸렸을 확률을 구하시오.

4-2 ✎ : 최대 우도 추정법 (MLE) 에 대한 예시 풀이를 보이겠습니다.

Q:

$X_1, \cdots, X_n$은 i.i.d 하며 다음과 같은 pdf 를 가집니다.

$$f(x;\, \theta) = \frac{1}{2\pi} e^{-\frac{1}{2}(x-\theta)^2}, \qquad -\infty < x < \infty$$

이때 θ 의 MLE 인 $\hat{\theta}$ 를 구하시오.

A:

$$f(x;\theta) = \frac{1}{2\pi} e^{-\frac{1}{2}(x-\theta)^2}$$

우선은 우도함수를 먼저 구해야 하며, 이것은 다음과 같이 구하면 됩니다.

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{1}{2\pi} e^{-\frac{1}{2}(x_i-\theta)^2} = \frac{1}{2\pi} \prod_{i=1}^{n} e^{-\frac{1}{2}(x_i-\theta)^2}$$

우도함수가 너무 큰 값이 나오는 것을 방지하기 위해서 log 함수를 씌어줍니다.

$$l(\theta) = \ln\left(L\left(\theta\right)\right) = \ln\left(\frac{1}{2\pi}\prod_{i=1}^{n} e^{-\frac{1}{2}(x_i-\theta)^2}\right) = \ln\left(\frac{1}{2\pi}\right) + \ln\left(\prod_{i=1}^{n} e^{-\frac{1}{2}(x_i-\theta)^2}\right)$$

$$= -\ln\left(2\pi\right) + \sum_{i=1}^{n} \ln\left(e^{-\frac{1}{2}(x_i-\theta)^2}\right) = -\ln\left(2\pi\right) + \sum_{i=1}^{n}\left(-\frac{1}{2}(x_i-\theta)^2\right)$$

$t = (x_i - \theta)$

$2t \cdot -1$

$2(x_i - \theta) \cdot -1$

이때 θ 에 대한 미분계수가 0 인 지점에서 최댓값을 가지기 때문에

θ 로 미분을 시켜서 이것이 0 이 되는 값을 찾습니다.

$$\frac{\partial}{\partial\theta} l(\theta) = -\sum_{i=1}^{n} -\frac{1}{2} \cdot 2 \cdot (-1) \cdot (x_i - \theta) = -\sum_{i=1}^{n}(x_i - \theta) = n\cdot\theta - \sum_{i=1}^{n} x_i = 0$$

$$n \cdot \hat{\theta} = \sum_{i=1}^{n} x_i, \quad \hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$\sum_{i=1}^{n}(x_i + \theta)$

MLE 는 $\sum_{i=1}^{n} x_i/n$, 혹은 $\bar{x}$ 가 나오게 됩니다.

위를 참조하면서 다음 문제를 직접 푸시오.

$X_1, \cdots, X_n$은 i.i.d 하며 다음과 같은 pdf 를 가집니다.

$$f(x;\ \theta) = \ \theta x^{\theta-1}, \qquad 0 < x < 1, \qquad 0 < \theta < \infty$$

이때 θ 의 MLE 인 $\hat{\theta}$ 를 구하시오.

Reference

- Introduction to Mathematical Statistics(8 판, Hogg et.al)
- 23-2 기초과제 2 (9 기 이성균)
- 23-2 머신러닝을 위한 수학 과제 (9 기 김서진, 서연우)

Data Science Lab

담당자 : 학술부 (신재우)

jaewoo356@gmail.com

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad 0 < \theta < \infty$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta x^{\theta-1}$$

$$\ln(L(\theta)) = \ln(\theta) + \ln\left(\prod_{i=1}^{n} x^{\theta-1}\right)$$

$l(\theta)$

$$x^{\theta-1} = \left(e^{\ln x}\right)^{\theta-1} = e^{(\theta-1)\ln x}$$

$$= \ln(\theta) + \sum_{i=1}^{n} ((\theta-1)\ln x)$$

$$\frac{\partial}{\partial\theta} l(\theta) = \frac{1}{\theta} + \sum_{i=1}^{n} \ln x = 0$$