





기초인공지능 과제 설명

과제4. Probability

Directory

-  `data_load.py` 텍스트 데이터들을 불러오는데 사용하는 코드
-  `hw4.py` 작성해야 하는 함수 정의 (5문제, 7개의 함수)
-  `reader.py` 데이터를 읽고 console로 output을 출력하여 결과값을 확인할 수 있는 코드
-  `data` 사용할 500건의 이메일 데이터 (.txt)

문제1. Joint distribution구하기

이번 과제에서 함수의 parameter로 사용할 두 개의 random variable은 다음과 같다.

- X_0 = text에서 word0의 등장 횟수
- X_1 = text에서 word1의 등장 횟수

위의 X_0 과 X_1 의 joint distribution을 구하는 함수, `joint_distribution_of_word_counts`를 구현해야한다.

수식)

Joint distribution:
$$P(X_0 = x_0, X_1 = x_1) = \frac{N(X_0 = x_0, X_1 = x_1)}{\sum_{x_0} \sum_{x_1} N(X_0 = x_0, X_1 = x_1)}$$

함수 설명) 함수 `joint_distribution_of_word_counts`는 주어진 텍스트에서 두 단어의 joint distribution을 계산하는 함수이다. 리스트로 구성된 texts와 첫 번째 단어인 word0, 두 번째 단어인 word1이 입력으로 주어진다.

주어진 texts에서 word0과 word1의 joint probability를 계산하고 그 값을 Pjoint에 할당하고 return한다.

출력 예시)

```
문제1. Joint distribution:
[[0.964 0.024 0.002 0.      0.002]
 [0.006 0.      0.      0.      0.    ]
 [0.      0.      0.      0.      0.    ]
 [0.      0.      0.      0.      0.    ]
 [0.002 0.      0.      0.      0.    ]]
```

문제2. Marginal distribution 구하기

수식)

Marginal distribution: $P(X_0 = x_0) = \sum_{x_1} P(X_0 = x_0, X_1 = x_1)$,

$$P(X_1 = x_1) = \sum_{x_0} P(X_0 = x_0, X_1 = x_1)$$

함수 설명) 본 문제에서 구현해야하는 함수 `marginal_distribution_of_word_counts`는 문제1에서 구한 `Pjoint`를 이용하여 `marginal distribution`을 구하는 함수이다. 입력으로는 `Pjoint`와 `index`가 주어지고, 여기서 `index`는 어떤 변수를 유지할지 나타낸다. `Pjoint`에서 주어진 `index`를 사용하여 `marginal distribution`을 계산하여 그 결과를 `Pmarginal`에 저장하여 `return`한다.

출력 예시)

```
문제2. Marginal distribution:
P0: [0.992 0.006 0.    0.    0.002]
P1: [0.972 0.024 0.002 0.    0.002]
```

문제3. Conditional distribution 구하기

수식)

Conditional distribution:
$$P(X_1 = x_1 | X_0 = x_0) = \frac{P(X_0 = x_0, X_1 = x_1)}{P(X_0 = x_0)}$$

함수 설명) 본 문제에서 구현해야하는 함수 `conditional_distribution_of_word_counts`는 문제1과 2에서 구한 `Pjoint`와 `Pmarginal`을 입력으로 받아 conditional distribution을 구하기 위한 함수이다. 결과는 `Pcond`에 저장하여 return한다.

출력 예시)

```
문제3. Conditional distribution:
[[0.97177419 0.02419355 0.00201613 0.          0.00201613]
 [1.          0.          0.          0.          0.          ]
 [          nan          nan          nan          nan          nan]
 [          nan          nan          nan          nan          nan]
 [1.          0.          0.          0.          0.          ]]
```

문제4. Mean, Variance, Covariance 구하기

영어 문장에서 자주 등장하는 a, the를 이용하여 joint distribution(Pathe), marginal distribution(Pthe)를 구한다. Pathe와 Pthe는 reader.py에 이미 정의되어있다. 본 문제에서는 이 확률 분포들을 이용하여 (4-1) Mean, (4-2) Variance, (4-3) Covariance를 계산하는 함수를 구현해야한다.

수식)

$$\text{mean: } \mu = \sum_x x \cdot P(X=x) \quad \text{variance: } \sigma^2 = \sum_x (x - \mu)^2 \cdot P(X=x)$$

$$\text{Covariance: } \text{Cov}(X_0, X_1) = \sum_{x_0, x_1} (x_0 - \mu_{X_0})(x_1 - \mu_{X_1}) \cdot P(X_0 = x_0, X_1 = x_1)$$

4-1, 4-2 함수 설명) `mean_from_distribution`, `variance_from_distribution`는 각각 입력으로 주어진 확률 분포 P(Pthe)에서 확률 변수 X의 평균과 분산을 계산해서 return하는 함수이다. Mean, Variance 모두 반올림하여 소수 셋째 자리까지 구한다.

4-3 함수 설명) `covariance_from_distribution`는 주어진 확률 분포 P(Pathe)에서 확률 변수 X0과 X1의 Covariance를 return하는 함수이다. Covariance 역시 반올림하여 소수 셋째 자리까지 출력한다.

출력 예시)

```
문제4-1. Mean from distribution:
4.432
문제4-2. Variance from distribution:
41.601
문제4-3. Covariance from distribution:
9.245
```

문제5. Expected Value of a Function 구하기

함수 설명) 본 문제에서 구현해야하는 함수 `expectation_of_a_function`는 두 확률변수 X_0, X_1 의 $E[f(X_0, X_1)]$ 을 계산하는 함수이다. 입력으로 받는 P 는 joint distribution이고, f 는 두 개의 실수값을 입력으로 받는 함수로 $f(x_0, x_1)$ 형태로 호출된다. 함수 f 는 `reader.py`에 정의되어있다. $E[f(X_0, X_1)]$ 은 반올림하여 소수 셋째 자리까지 구하여 return한다.

```
def f(x0, x1):  
    return np.log(x0 + 1) + np.log(x1 + 1)
```

수식

If $f(x_0, x_1)$ is some real-valued function of variables x_0 and x_1 then its expected value is:

$$E[f(X_0, X_1)] = \sum f(x_0, x_1)P(X_0 = x_0, X_1 = x_1)$$

출력 예시) 문제5. Expectation of a function:
1.772

최종 출력

```
문제1. Joint distribution:
[[0.964 0.024 0.002 0.    0.002]
 [0.006 0.    0.    0.    0.   ]
 [0.    0.    0.    0.    0.   ]
 [0.    0.    0.    0.    0.   ]
 [0.002 0.    0.    0.    0.   ]]

-----

문제2. Marginal distribution:
P0: [0.992 0.006 0.    0.    0.002]
P1: [0.972 0.024 0.002 0.    0.002]

-----

문제3. Conditional distribution:
[[0.97177419 0.02419355 0.00201613 0.    0.00201613]
 [1.        0.        0.        0.        0.        ]
 [         nan         nan         nan         nan         nan]
 [         nan         nan         nan         nan         nan]
 [1.        0.        0.        0.        0.        ]]

-----

문제4-1. Mean from distribution:
4.432
문제4-2. Variance from distribution:
41.601
문제4-3. Covariance from distribution:
9.245

-----

문제5. Expectation of a function:
1.772
```


주의사항 및 제출

- 코드와 보고서 압축하여 zip파일로 제출 (**2023. 11. 16 (목) 18:00 마감**)
python file: hw4.py
report: AI분반_학번_이름.pdf
- **Copy 금지**