기초 인공지능

2023 Second Semester CSE 4185
Assignment#04

1. Requirements

- python version >= 3.6
- numpy >= 1.15
- nltk >= 3.4
- tqdm >= 4.24.0
- scikit-learn >= 0.22

2. 문제 설명

주어진 파일에는 데이터 파일(data 폴더), data_load.py, reader.py, hw4.py이 있다. 데이터 파일은 이번 과제에서 사용할 500건의 이메일 데이터가 각각 txt파일 형식으로 저장되어 있다. data_load.py는 텍스트 데이터들을 불러오는데 사용되며, reader.py는 데이터를 읽고 console로 output을 출력하여 결과값을 확인할 수 있는 코드들이 작성되어있다. 작성해야하는 함수는 모두 hw4.py에 정의되어 있다. hw4.py의 주석을 참고하여 주어진 함수를 모두 작성해야 한다.

결과값은 다음과 같은 command를 콘솔에 입력하여 확인한다.

python reader.py or py reader.py

문제마다 출력 예시가 주어지며, 출력 예시와 본인이 작성한 코드의 결과가 같은지 확인하고 다음 문제로 넘어가기를 권장한다. 문제마다 주어진 수식을 참고하면 문제 풀이에 도움이 될 것이다.

문제는 총 5개로 구성되어 있으며, 작성해야하는 함수는 총 7개로 이루어져있다. 오류를 전파하는 기본 코드(raise RuntimeError("You need to write this part!"))를 지우고 주어진 주석에 맞게 함수를 구현해야한다.

각 문제와 문제별로 구현해야할 각 함수에 대한 설명은 아래와 같다. input과 output에 대한 상세한 설명은 hw4.py의 주석을 참고하면 된다.

▶ 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_0과 X_1의 joint distribution을 구하는 함수, **joint_distribution_of_word_counts**를 구현해야한다.

수식)

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

기초 인공지능

2023 Second Semester CSE 4185
Assignment#04

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 Pjoint를 이용하여 marginal distribution을 구하는 함수이다. 입력으로는 Pjoint와 index가 주어지고, 여기서 index는 어떤 변수를 유지할지 나타낸다. Pjoint에서 주어진 index를 사용하여 marginal distribution을 계산하여 그 결과를 Pmarginal에 저장하여 return한다.

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: $P(X_1 = x_1 | X_0 = x_0) = \dfrac{P(X_0 = x_0, X_1 = x_1)}{P(X_0 = x_0)}$

함수 설명) 본 문제에서 구현해야하는 함수 **conditional_distribution_of_word_counts**는 문제1과 2에서 구한 Pjoint와 Pmarginal을 입력으로 받아 conditional distribution을 구하기 위한 함수이다. 결과는 Pcond에 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.        ]]
```

기초 인공지능

2023 Second Semester CSE 4185
Assignment#04

▶ 문제4. Mean, Variance, Covariance 구하기

주어진 확률 분포에서 확률 변수의 Mean, Variance, Covariance를 계산하고자 한다.
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)

mean: $\mu = \sum_{x} x \cdot P(X = x)$

variance: $\sigma^2 = \sum_{x} (x - \mu)^2 \cdot P(X = x)$

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

4-1, 4-2 함수 설명) mean_from_distribution, variance_from_distribution는 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_distribution는 주어진 확률 분포 P(Pathe)에서 확률 변수 X0과 X1의 Covariance을 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_1의 $E[f(X_0, X_1)]$을 계산하는 함수이다. 입력으로 받는 P는 joint distribution이고, f는 두 개의 실수값을 입력으로 받는 함수로 f(x0, x1) 형태로 호출된다. 함수 f는 reader.py에 정의되어있다. $E[f(X_0, X_1)]$은 반올림하여 소수 셋째 자리까지 구하여 return한다.

수식)

If $f(x_0, x_1)$ is some real-valued function of variables x_0 and x_1 then its expected value is: $E[f(X_0, X_1)] = \sum_{x_0, x_1} f(x_0, x_1) P(X_0 = x_0, X_1 = x_1)$

출력 예시)

```
문제5. Expectation of a function:
1.772
```

기초 인공지능

2023 Second Semester CSE 4185
Assignment#04

3. 보고서

보고서 분량 제한은 없으나, 반드시 다음과 같은 내용이 포함되어야한다.

1. 각 함수마다 구현한 방법에 대한 간략한 설명
2. 실행 결과 캡처 화면

4. 주의사항

- 코드 실행시 출력 화면과 보고서에 첨부된 화면 캡처 내용이 반드시 동일해야 한다. (다를 경우 코드 실행 시의 결과를 기준으로 점수를 산정할 것)
- 라이브러리는 자유롭게 사용 가능하며 추가적인 test case가 있을 수 있음.
- 본인이 **작성한 코드에 대하여 annotation**을 작성할 것. (미 작성 시 감점)
- **copy check 적발시 0점 처리.**

5. 제출

아래 두 가지 파일만 압축하여 AI분반_학번_이름.zip으로 사이버 캠퍼스에 업로드 한다.
- python file: hw4.py
- report: AI분반_학번_이름.pdf