Jin-Soo Kim
(jinsoo.kim@snu.ac.kr)

Systems Software &
Architecture Lab.

Seoul National University

Jan. 6 – 17, 2020

Python for Data Analytics

Data Preprocessing

Lab 1. titanic data preprocessing - step1

- titanic dataset의 데이터를 preprocessing 해서 decision tree를 만들어 본다.
  - age null값을 처리해준다.
  - Null 값이 있는 row를 제거하거나, 평균 값을 넣거나, 등등 …
  - 정확도(accuracy)를 확인해본다.

Lab 1. titanic data preprocessing – step2

- titanic dataset의 데이터를 outliers using IQR을 이용해 preprocessing 한다.
  - fare 값을 처리해준다.
  - 정확도(accuracy)를 확인해본다.

Lab 2. Distribution Transformation

- Boston 집값의 데이터셋으로 linear regression을 하고, distribution transformation을 한 결과를 확인한다.
  - x 값에는 LSTAT(인구 중 하위 계층비율)
  - x 값에 transformation 한 결과 중 normal distribution과 비슷한 값과 비교해본다.
  - y 값에는 MEDV(집값) 교안과 동일
  - 실제로 linear regression을 진행하고, sqrt, log로 데이터를 변형한다.
  - mean squared error 과 variance를 비교해본다.

Lab 3. one hot encoding – step 1

- Class : [“bad”, “not good”, “good”]
- [‘bad’, ‘bad’, ‘not good’, ‘good’, ‘not good’, ‘bad,’ ‘bad’, ‘good’, ‘good’, ‘not good’]
- 위 단어들을 integer encoding (such as 0,1,2) 한다.

```
>>> print(integer_encoded)
[0 0 2 1 2 0 0 1 1 2]
```

Lab 3. one hot encoding – step 2

- [0 0 2 1 2 0 0 1 1 2] 값은 one hot encoding 한다.

```
>>> print(onehot_encoded)
[[1 0 0],
 [1 0 0],
 [0 0 1]
…
]
```

Lab 3. one hot encoding – step 2

- 앞 페이지의 아우풋 값을 다시 단어로 바꾼다.

```
>>> print(inverted)
… ['good']
…
```