Python & Azure Machine Learning

Python을 이용한 머신러닝 시작부터 예측모델 배포까지
https://github.com/CloudBreadPaPa/azure-ml-python-seminar

한국마이크로소프트 | 김대우

2017-01-21

Python Machine Learning

Azure Machine Learning	01
Iris 분석 모델 workflow 구축	02
Data partition	03
Task flow	04
Accuracy check	05
Python code execution	06
Deploy to Web API and Excel	07

머신러닝 데모

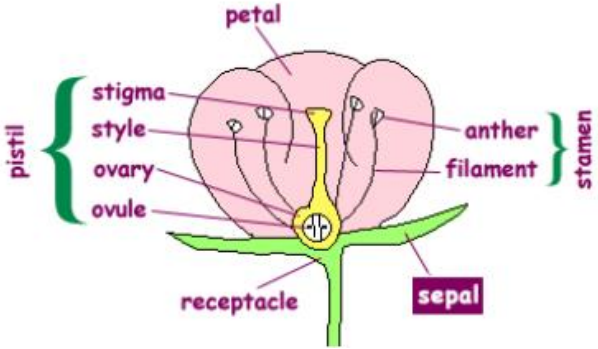
- IRIS Data

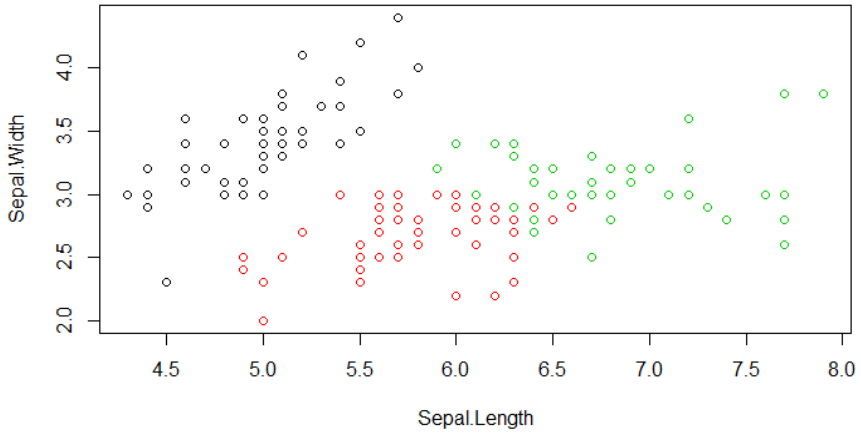
(뵨사마와 무관한 통계업계의 “Hello World”)

Iris setosa

Iris versicolor

Iris virginica





출처 : 디에스이트레이드 이성희

Iris 붓꽃 데이터 현황

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

- **Column : 4개**
- **Row : 150개**

- Sepal.Length : num
- Sepal.Width : num
- Petal.Length : num
- Petal.Width : num
- Species : Factor

Iris모델 구축

데모 + 코드

예측모델 생성 데모

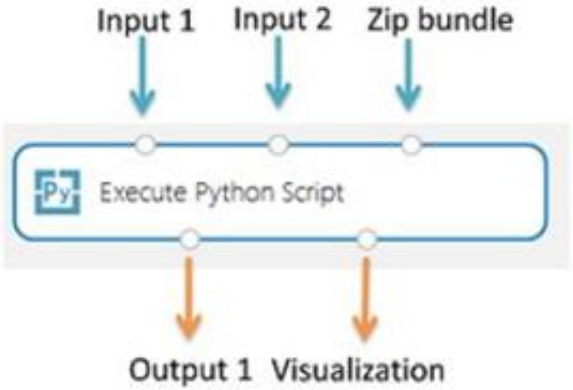
pydata & pandas

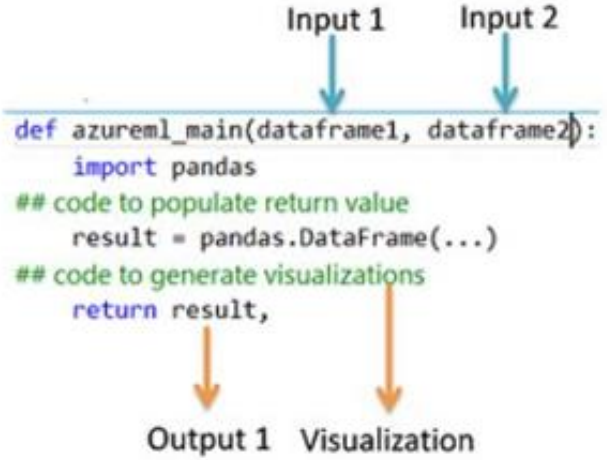
https://github.com/wesm/pydata-book

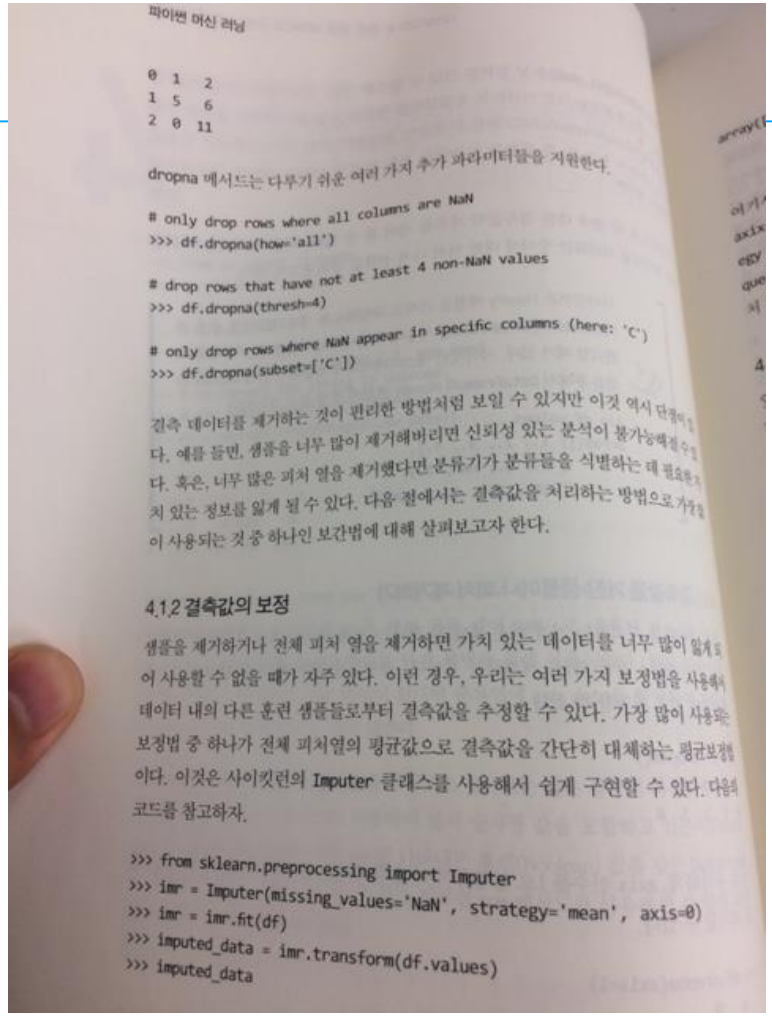
Materials and IPython notebooks for "Python for Data Analysis" by Wes McKinney, published by

O'Reilly Media

Python code execution







파이썬 머신 러닝

```
0  1   2
1  5   6
2  0  11
```

dropna 메서드는 다루기 쉬운 여러 가지 추가 파라미터들을 지원한다.

```
# only drop rows where all columns are NaN
>>> df.dropna(how='all')

# drop rows that have not at least 4 non-NaN values
>>> df.dropna(thresh=4)

# only drop rows where NaN appear in specific columns (here: 'C')
>>> df.dropna(subset=['C'])
```

결측 데이터를 제거하는 것이 편리한 방법처럼 보일 수 있지만 이것 역시 단점이 있
다. 예를 들면, 샘플을 너무 많이 제거해버리면 신뢰성 있는 분석이 불가능해질 수
다. 혹은, 너무 많은 피처 열을 제거했다면 분류기가 분류들을 식별하는 데 필요한 가
치 있는 정보를 잃게 될 수 있다. 다음 절에서는 결측값을 처리하는 방법으로 가장 많
이 사용되는 것 중 하나인 보간법에 대해 살펴보고자 한다.

4.1.2 결측값의 보정

샘플을 제거하거나 전체 피처 열을 제거하면 가치 있는 데이터를 너무 많이 잃게 되
어 사용할 수 없을 때가 자주 있다. 이런 경우, 우리는 여러 가지 보정법을 사용해서
데이터 내의 다른 훈련 샘플들로부터 결측값을 추정할 수 있다. 가장 많이 사용되는
보정법 중 하나가 전체 피처열의 평균값으로 결측값을 간단히 대체하는 평균보정법
이다. 이것은 사이킷런의 Imputer 클래스를 사용해서 쉽게 구현할 수 있다. 다음의
코드를 참고하자.

```
>>> from sklearn.preprocessing import Imputer
>>> imr = Imputer(missing_values='NaN', strategy='mean', axis=0)
>>> imr = imr.fit(df)
>>> imputed_data = imr.transform(df.values)
>>> imputed_data
```

!!!

API로 노출

Python + node.js + C# + …

Python 등에서 API 사용 DEMO

Restful Front-End를 이용한 Machine Learning API 호출 + 대량 Batch 분석 DEMO

Q & A

Python & Azure Machine Learning

Python을 이용한 머신러닝 시작부터 예측모델 배포까지
https://github.com/CloudBreadPaPa/azure-ml-python-seminar

한국마이크로소프트 | 김대우

2017-01-21