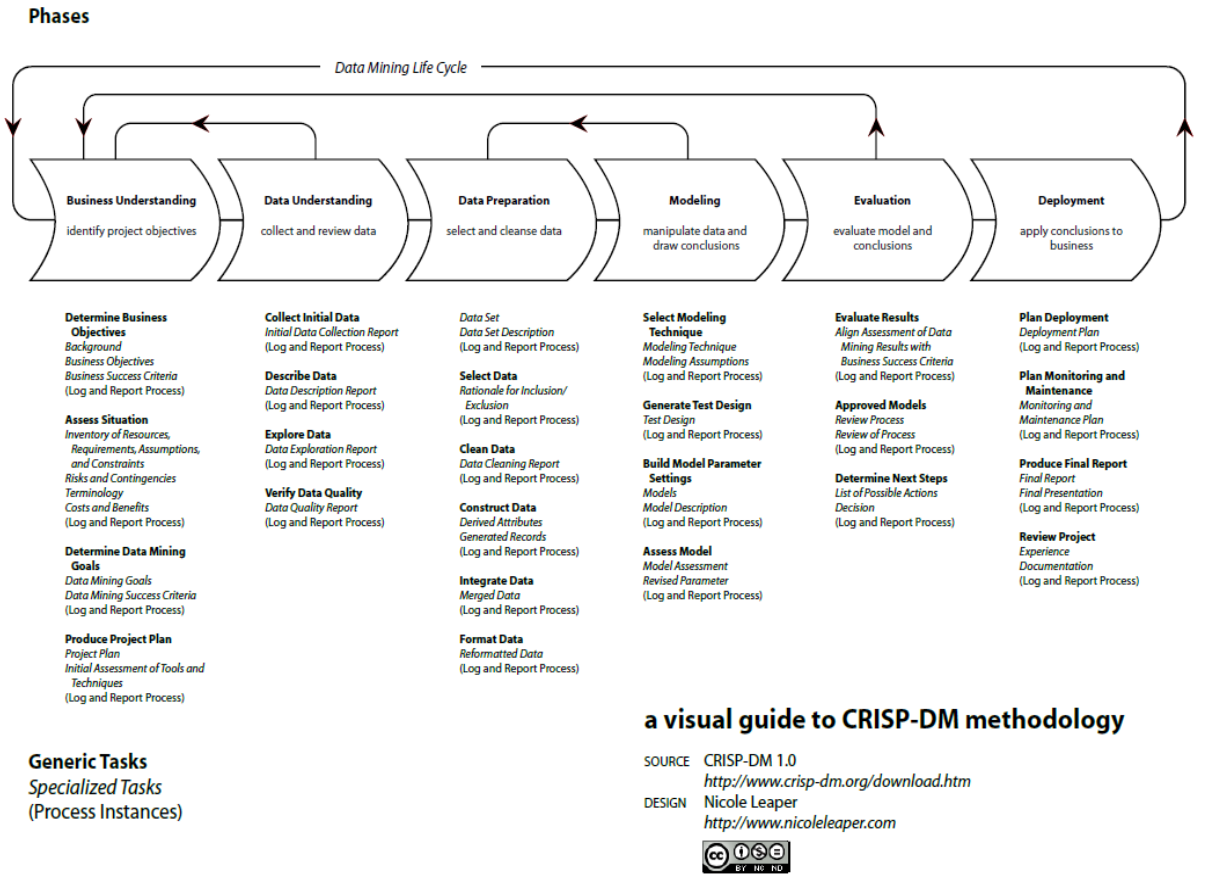
Python 기반 데이터 전처리 실습

2021. 8. 13

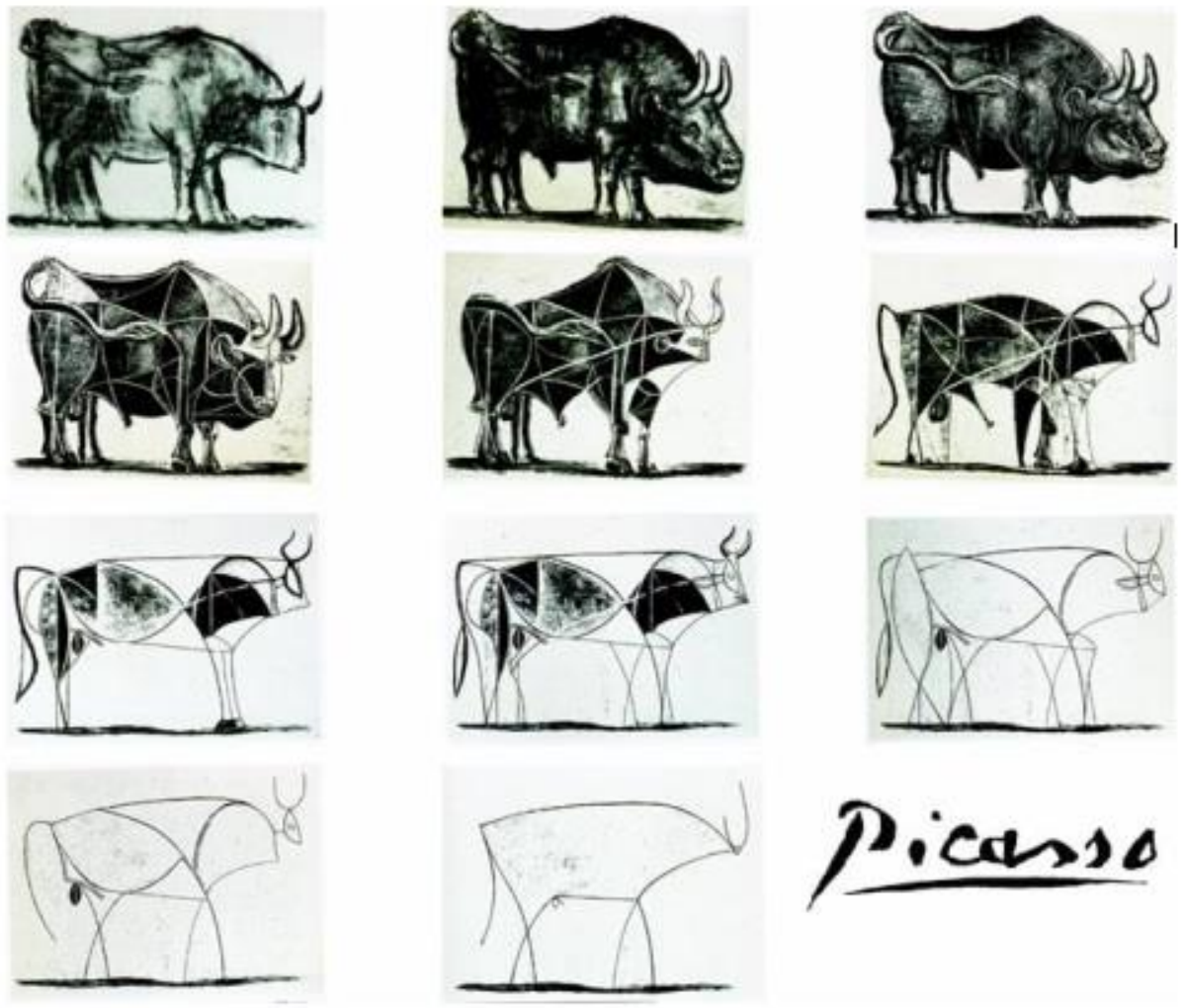
정 준 수 Ph.D

CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM(Cross Industry Standard Process for Data Mining)은 데이터 마이닝 전문가가 사용하는 일반적인 접근 방식을 설명한 가장 널리 사용되는 공개 표준 분석 모델입니다.

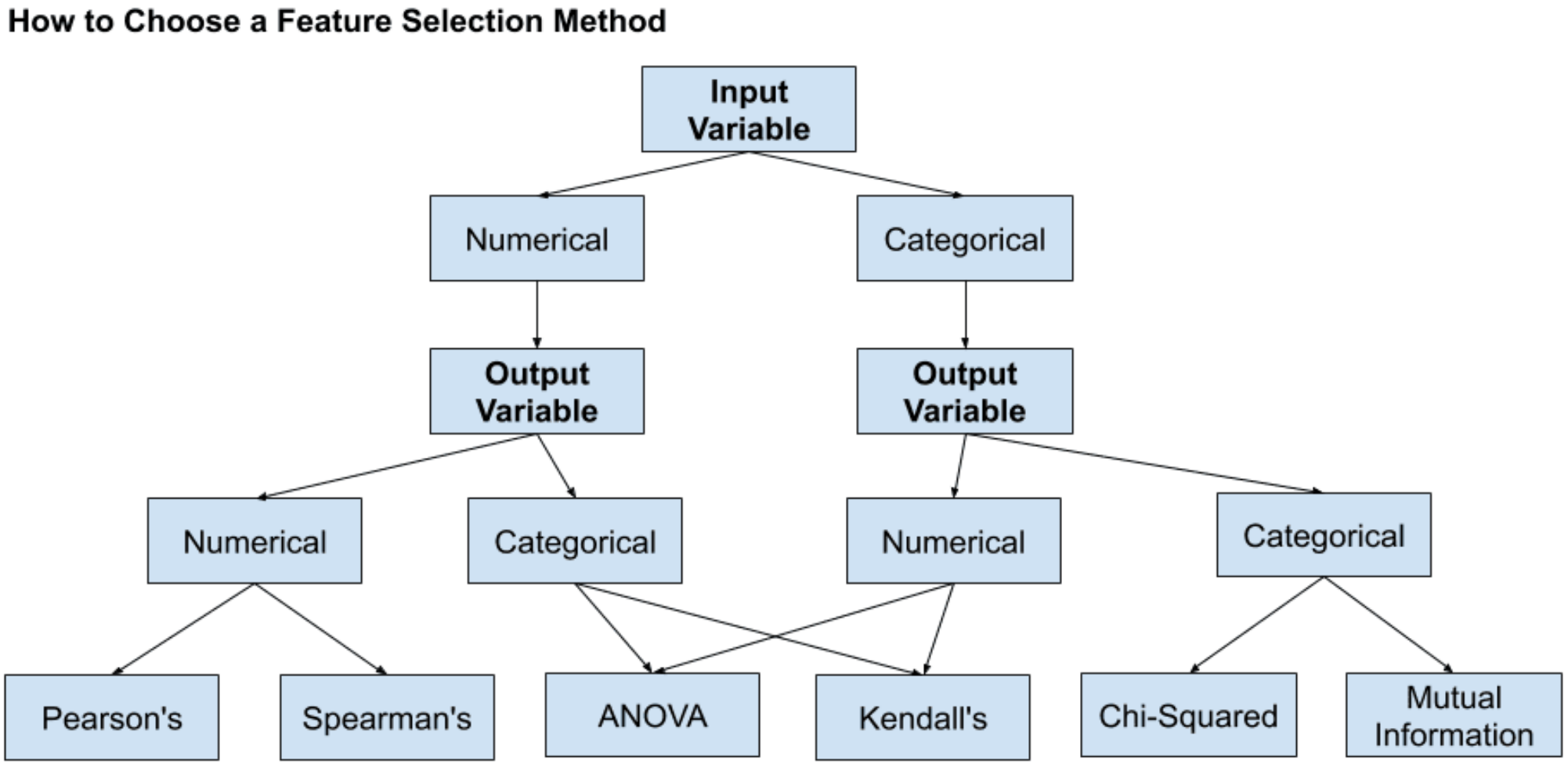


◈ 추상화 (Abstract)



Pablo Picasso, Bull (plates I - XI) 1945

Feature Selection 분류 방법



머신러닝에서 일반적 Data Preparation 과정 정리

1. 데이터 준비 과정의 중요성
2. 결측치의 처리방법
3. 특징 추출 (Recursive Feature Elimination)
4. 데이터 정규화
5. 원 핫 인코딩으로 범주 변환 (One Hot Encoding)
6. 숫자 변수의 범주형 변수로 변환
7. PCA를 통한 차원 축소

1. 데이터 준비 과정의 중요성

데이터 준비 과정은 원시 데이터를 모델링에 적합한 형식으로 변환하는 작업이 중심 내용이며, 데이터를 준비하는 것은 예측 모델 생성 프로젝트에서 가장 중요한 부분이며 가장 많은 시간이 소요된다. 모든 데이터 준비 과정은 기계 학습 알고리즘에 맞추어져 진행 되며, 실제 데이터와 매개 변수 등을 활용한다. 실질적인 데이터 준비 과정에는 데이터 정리, 특징 추출, 데이터 변환, 차원 축소 등에 대한 지식이 필요하다.

예측 모델링 프로젝트에는 데이터로부터 학습이 진행됩니다. 데이터는 해결하려는 문제를 특징 짓는 도메인으로 부터 경험 할 수 있는 사례를 제공하지만, 분류 또는 회귀와 같은 예측 모델링 프로젝트에서 원시 데이터는 일반적으로 직접 사용할 수 없습니다. 이것이 사실인 네 가지 주요 이유가 있습니다.

원시 데이터는 기계 학습 모델을 맞추고 평가하는 데 사용되기 전에 사전 처리 되어야 합니다. 기계 학습 프로젝트의 데이터 준비 단계에서 사용하거나 탐색 할 수 있는 일반 또는 표준 작업이 있습니다.

- 데이터 정리 : 데이터의 오류 또는 오류를 식별하고 수정
- 특징 선택 : 작업과 가장 관련된 입력 변수 식별
- 데이터 변환 : 변수의 척도 또는 분포 파악
- 특징 엔지니어링 : 사용 가능한 데이터에서 새로운 변수 도출
- 차원 감소 : 데이터의 간결한 예측 생성

2. 결측치의 처리방법

실제 데이터에는 종종 결 측값이 있습니다. 데이터에는 기록되지 않은 관찰 및 데이터 손상과 같은 여러 가지 이유로 인해 누락 된 값이 있을 수 있습니다. 누락 된 데이터 처리가 중요합니다.
많은 기계 학습 알고리즘이 결 측값이 있는 데이터를 지원하지 않기 때문입니다. 누락 된 값을 데이터로 채우는 것을 데이터 대치라고 하며 데이터 대치에 대한 일반적인 접근 방식은 각 열 (예 : 평균)에 대한 통계 값을 계산하고 해당 열의 모든 누락 된 값을 통계로 바꾸는 것입니다.

(결측치 처리 예제 – 단순 평균값 입력)

SimpleImputer() 클래스를 사용하여 NaN 값으로 표시된 모든 누락 된 값을 열의 평균으로 변환 할 수 있습니다.

<실습 프로그램>

https://gist.github.com/JSJeong-me/fdbba476a9cff9400ba32064a92f54e8

3. 특징 추출 (Recursive Feature Elimination)

특징 선택은 예측 모델을 개발할 때 입력 변수의 수를 줄이는 프로세스입니다. 모델링의 계산 비용을 줄이고 경우에 따라 모델의 성능을 향상시키기 위해 입력 변수의 수를 줄이는 것이 바람직합니다.
간단히 RFE는 인기있는 기능 선택 알고리즘입니다.
RFE는 구성 및 사용이 쉽고 대상 변수를 예측하는 데 더 많거나 가장 관련성이 높은 학습 데이터 세트에서 이러한 기능 (열)을 선택하는 데 효과적이기 때문에 널리 사용됩니다.

(특징 추출 예제 – scikit-learn 사용)

scikit-learn Python 기계 학습 라이브러리는 기계 학습을 위한 RFE 구현을 제공합니다. RFE 변환을 사용하려면 먼저 추정치 인수를 통해 지정된 선택한 알고리즘과 인수를 선택하기 위해 n 개의 기능을 통해 선택할 기능수로 클래스를 구성합니다. 다음의 예는 5 개의 중복 입력 기능이 있는 합성 분류 데이터 세트를 정의합니다. 그런 다음 RFE를 사용하여 의사 결정 트리 알고리즘을 사용하여 5 개의 기능을 선택합니다.

<실습 프로그램>

https://gist.github.com/JSJeong-me/7d6a3f852eb0e9eb451e4c153af6cc6f

4. 데이터 정규화

기계 학습을 위해 숫자 데이터를 확장하는 방법을 알아 봅니다. 많은 기계 학습 알고리즘은 숫자 입력 변수가 표준 범위로 조정될 때 더 잘 수행됩니다. 여기에는 선형 회귀와 같은 입력의 가중 합계를 사용하는 알고리즘과 k- 최근 접 이웃과 같은 거리 측정을 사용하는 알고리즘이 포함됩니다.

모델링 전에 수치 데이터를 스케일링하는 가장 널리 사용되는 기술 중 하나는 정규화입니다. 정규화는 각 입력 변수를 0-1 범위로 개별적으로 조정합니다. 이는 가장 정밀도가 높은 부동 소수점 값의 범위입니다. 각 변수에 대한 최소 및 최대 관찰 가능 값을 알고 있거나 정확하게 추정 할 수 있어야합니다.

(데이터 정규화 예제 – MinMaxScaler 사용)

scikit-learn 객체 MinMaxScaler를 사용하여 데이터 세트를 정규화 할 수 있습니다. 아래 예제는 합성 분류 데이터 세트를 정의한 다음 MinMaxScaler를 사용하여 입력 변수를 정규화합니다.

<실습 프로그램>

https://gist.github.com/JSJeong-me/fa8e38c0d0960f520731de45f7e1b6eb

5. 원 핫 인코딩으로 범주 변환 (One Hot Encoding)

범주 형 입력 변수를 숫자로 인코딩하는 방법을 알아 봅니다. 기계 학습 모델에서는 모든 입력 및 출력 변수가 숫자여야 합니다. 즉, 데이터에 범주형 데이터가 포함된 경우 모델을 적합하고 평가하기 전에 이를 숫자로 인코딩 해야합니다. 범주 형 변수를 숫자로 변환하는 가장 널리 사용되는 기술 중 하나는 원 핫 인코딩입니다. 범주 형 데이터는 숫자 값이 아닌 레이블 값을 포함하는 변수입니다.

범주 형 변수에 대한 각 레이블은 서수 인코딩이라고하는 고유 한 정수에 매핑 될 수 있습니다. 그런 다음 서수 표현에 원 핫 인코딩을 적용 할 수 있습니다. 여기서 변수의 고유 한 정수 값 각각에 대해 하나의 새 이진 변수가 데이터 세트에 추가되고 원래 범주 형 변수가 데이터 세트에서 제거됩니다.

(데이터 Encoding 예제 – One Hot Encoding 사용)

이 핫 인코딩 변환은 OneHotEncoder 클래스를 통해 scikit-learn Python 기계 학습 라이브러리에서 사용할 수 있습니다. 유방암 데이터 세트에는 범주 형 입력 변수 만 포함됩니다. 아래 예제는 데이터 세트를 로드하고 하나의 핫 인코딩은 각 범주 형 입력 변수를 인코딩합니다.

<실습 프로그램>

https://gist.github.com/JSJeong-me/664af116682e32f0e1fb0141c280d879

6. 숫자 변수의 범주형 변수로 변환

숫자 변수를 범주형 변수로 변환하는 방법을 알아 봅니다. 일부 기계 학습 알고리즘은 일부 의사 결정 트리 및 규칙 기반 알고리즘과 같은 범주 형 또는 순서 형 입력 변수를 선호하거나 요구할 수 있습니다. 이것은 데이터, 다중 입력 데이터 분포, 고도의 지수 분포 등. 많은 기계 학습 알고리즘은 비표준 분포를 가진 숫자 입력 변수가 새로운 분포 또는 완전히 새로운 데이터 유형을 갖도록 변환 될 때 더 나은 성능을 제공합니다.

한 가지 접근 방식은 숫자 변수의 변환을 사용하여 각 숫자 값에 레이블이 할당되고 레이블에 순서가 지정된 (순서적) 관계가 있는 이산 확률 분포를 갖는 것입니다. 이를 이산화 변환이라고 하며 숫자 입력 변수의 확률 분포를 이산 적으로 만들어 데이터 세트에 대한 일부 기계 학습 모델의 성능을 향상시킬 수 있습니다.

(숫자 변수의 범주형 변수 변환 예제 – KBinsDiscretizer 사용)

이산화 변환은 KBinsDiscretizer 클래스를 통해 scikit-learn Python 기계 학습 라이브러리에서 사용할 수 있습니다. 생성 할 이산 구간의 수 (n 구간), 변환 결과가 서수인지 아니면 하나의 핫 인코딩 (인코딩)인지, 변수 값을 나누는 데 사용되는 분포 (전략)를 지정할 수 있습니다. 아래 예제에서는 10 개의 숫자 입력 변수 가있는 합성 입력 변수를 생성 한 다음 각각을 서수 인코딩을 사용하여 10 개의 개별 Bin(바구니)으로 인코딩합니다.

<실습 프로그램>

https://gist.github.com/JSJeong-me/34721c717cfc4a4f9e5eaee4f33124c2

7. PCA를 통한 차원 축소

차원 감소를 사용하여 데이터 세트의 입력 변수를 줄이는 방법을 알아 봅니다. 데이터 세트에 대한 입력 변수 또는 기능의 수를 차원이라고 합니다. 차원 감소는 Features 수를 줄이는 기술을 의미합니다. 데이터 세트의 입력 변수. 더 많은 입력 기능은 종종 예측 모델링 작업을 모델링하기 더 어렵게 만들고, 일반적으로 차원의 저주라고 합니다. 고차원 통계에서는 차원 감소 기술이 데이터 시각화에 자주 사용되지만, 이러한 기술은 예측 모델에 더 적합하도록 분류 또는 회귀 데이터 세트를 단순화하기 위해 적용된 기계 학습에서 사용할 수 있습니다. 기계 학습에서 차원 감소를 위한 가장 인기있는 기술은 주성분 분석 (줄여서 PCA) 일 것입니다.

(PCA를 통한 차원 축소 예제 – PCA)

scikit-learn 라이브러리는 데이터 세트에 적합하고 향후 학습 데이터 세트 및 추가 데이터 세트를 변환하는 데 사용할 수 있는 PCA 클래스를 제공합니다. 아래 예제는 10 개의 입력 변수가 있는 합성 이진 분류 데이터 세트를 만든 다음 PCA를 사용하여 데이터 세트의 차원을 가장 중요한 세 가지 구성 요소로 줄입니다.

<실습 프로그램>

https://gist.github.com/JSJeong-me/ec0826fc9da3f7e1bac19bcf95aef47f

<실습 예제 프로그램>

https://github.com/JSJeong-me/JBNU-2021

강사 소개

정 준 수 / Ph.D (heinem@naver.com)

- 前) 삼성전자 연구원
- 前) 삼성의료원 (삼성생명과학연구소)
- 前) 삼성SDS (정보기술연구소)
- 現) (사)한국인공지능협회, AI, 머신러닝 강의
- 現) 한국소프트웨어산업협회, AI, 머신러닝 강의
- 現) 서울디지털재단, AI 자문위원
- 現) 한성대학교 교수(겸)
- 전문분야: 시각 모델링, 머신러닝(ML), RPA
- https://github.com/JSJeong-me/