머신러닝(Machine Learning)

머신러닝 소개 및 시작하기

목 차

01 머신러닝으로 풀 수 있는 문제

(1) 편지 봉투에 손으로 쓴 우편번호 숫자 판별

(2) 의료 영상 이미지에 기반한 종양 판단

=> 입력이 이미지, 출력은 종양이 양성인지의 여부.

(3) 의심되는 신용카드 거래 감지

=> 신용카드 거래 내역이 **입력**이 되고 부정 거래인지가 **출력**이 된다.

01 머신러닝으로 풀 수 있는 문제

(4) 블로그 글의 주제 구분

=> 많은 양의 **텍스트 데이터를 요약**하고 그 안에 담긴 **핵심 주제**를 찾기.

(5) 고객들을 취향이 비슷한 그룹으로 묶기

=> 어떤 고객들의 취향이 비슷한 지 비슷한 취향의 고객을 그룹으로 묶고 싶을 때,

(6) 비정상적인 웹 사이트 접근 탐지

=> 정상 패턴과 비정상 패턴을 찾아본다.

(7) 영화 추천에서 음식 주문, 쇼핑, 맞춤형 온라인 라디오 방송 등.

01 머신러닝으로 풀 수 있는 문제

(8) 얼굴 인식

=> 스마트폰 얼굴 인식. 픽셀 데이터를 이용한 학습을 통해 사람의 얼굴 확인 가능.

01 머신러닝이란?

(가) 머신러닝은 데이터에서 지식을 추출하는 작업이다.

(나) 머신러닝은 통계학, 인공지능, 컴퓨터 과학이 얽혀 있는 연구 분야

(다) 비정상적인 웹 사이트 접근 탐지

=> 정상 패턴과 비정상 패턴을 찾아본다.

02 문제와 데이터 이해하기

▶ 머신러닝의 가장 중요한 과정은 사용할 데이터를 이해하고 데이터가 해결해야 할 과제를 이해하는 일이다.

A. 어떤 질문에 대한 답을 원하는가? 가지고 있는 데이터가 원하는 답을 줄 수 있는가?

B. 내 질문을 머신 러닝 문제로 가장 잘 기술하는 방법은 무엇인가?

C. 문제를 풀기에 충분한 데이터를 모았는가?

D. 내가 추출한 데이터의 특성은 무엇이며 좋은 예측을 만들어 낼 수 있을 것인가?

E. 머신 러닝 애플리케이션의 성과를 어떻게 측정할 수 있는가?

F. 머신 러닝 솔루션이 다른 연구나 제품과 어떻게 협력할 수 있는가?

03 용어 이해하기

▶ 샘플(sample), 데이터 포인트(data point), 특성(feature)

(가) 샘플(sample) 또는 데이터 포인트

하나의 개체 또는 행을 샘플이라고 말한다.

(나) 특성(feature)

샘플의 속성, 즉 열을 말한다.

(다) 특성 추출(feature extraction) or 특성 공학(feature engineering)

좋은 입력 데이터를 만들어 내는 것.

04 Why Python?(왜 파이썬인가?)

- 범용 프로그래밍 언어의 장점.
- 다양한 라이브러리

  파이썬은 데이터 처리 및 적재, 통계, 자연어 처리, 시각화, 이미지 처리 등의 필요한 라이브러리들을 가지고 있다.

- 터미널, 주피터 노트북처럼 대화형 프로그래밍이 가능하다.
- 그래픽 사용자 인터페이스(GUI)나 웹 서비스도 만들 수 있음.

05 라이브러리 – 필수 프로그램

- Numpy, SciPy - 파이썬 과학 라이브러리
- pandas : 데이터 처리 및 적재 등
- matplotlib, seaborn, plotly : 시각화
- scikit-learn : 머신러닝 라이브러리
- tensorflow, keras, pytorch : 딥러닝 라이브러리

05 라이브러리 – Numpy

- 다차원 배열을 위한 기능
- 선형대수 연산과 푸리에 변환과 같은 고수준 수학 함수
- 유사 난수 생성기
- 배열이 기본 데이터 구조
- 데이터의 모든 원소는 동일한 데이터 타입이어야 함

05 라이브러리 – SciPy

- 과학 계산용 함수를 모아놓은 파이썬 패키지
- 고성능 선형대수, 함수 최적화, 신호 처리, 특수한 수학 함수와 통계분포
- 가장 중요한 기능 scipy.sparse - 희소 행렬 기능

05 라이브러리 – Matplotlib, Seaborn

- 파이썬의 대표적인 과학 계산용 그래프 라이브러리
- 데이터 분석 결과를 다양한 시각화를 통해 매우 중요한 통찰을 얻음.
- 적은 코드로 좀 더 세련된 시각화를 해 내기 (Seaborn)

05 라이브러리 – Pandas

- 파이썬의 대표적인 데이터 처리와 분석을 위한 파이썬 라이브러리
- 기본적인 데이터 구조 – DataFrame
- Numpy와 달리 각 열의 자료형 타입이 달라도 됨.
- SQL, 엑셀 파일, CSV 파일 같은 다양한 파일을 읽고 쓰기가 가능

05 라이브러리 – mglearn

- 이 책에서 다양한 설명을 위해 만들어진 라이브러리

06 파이썬 2 vs 파이썬 3

- 파이썬 2는 지원을 하지 않음. 현재 Python3 버전을 사용함.
- 2020년 9월 기준 Python 3.8.x 임.

07 머신러닝 기본 용어 이해

- 출력될 수 있는 값, 레이블의 범주를 클래스(class)라고 한다.
- 특정 포인트에 대한 출력을 레이블(label)이라고 한다.

08 iris 데이터 셋 이해

▶ Iris 데이터 셋

SETOSA

virginica

Verisicolor

Wiki 참조

150개의 샘플(sample), 5개의 특성(feature)

09 훈련 데이터(train)와 테스트 데이터(test)

▶ 모델의 평가

머신러닝 모델을 만든 후, 모델이 잘 작동하는지 판단해야 한다.

▶ 정확한 평가를 위해 데이터 나누기

머신러닝 모델을 학습할 때 사용하는 **훈련 데이터 셋(train set)**

머신러닝 모델을 평가할 때 사용하는 **테스트 데이터 or 테스트 세트(test set) or 홀드아웃 세트**

10 데이터에 대해 가장 먼저 할 일

▶ 머신러닝을 사용해서 풀 문제인지?

▶ 필요한 정보가 누락되지 않았을까? 또는 이상치는 없을까?

▶ 데이터를 탐색을 통해 알아보자(시각화)

2개의 특성 –산점도

3개의 특성 – 산점도 행렬 – (단점 : 데이터가 많을 경우, 오랜 시간이 필요)

History

날짜	내용	비고
2020/09/10	내용 업데이트	

