3.2 분석기법 적용

이즈 이론

통계학의 확률은 크게 빈도 확률과 베이지안 확률로 구분할 수 있다. 빈도확률은 객관적으로 확률을 해석하고, 베이지안 확률은 주관적으로 확률을 해석하는 것으로 볼 수 있다.

개념 1

● 베이즈 정리

베이즈 정리란 사전확률과 우도확률을 통해 사후확률을 추정하는 정리로 데이터를 통해 확률을 추정할 때 현재 관측된 데이터의 빈도만으로 분석하는 것이 아니라, 분석자의 사전지식까지 포함해 분석하는 방법이다.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

$P(H)$: H가 발생할 사전 확률

$P(E)$: E가 발생할 확률, 상수 취급

$P(H|E)$: 사건 E가 발생한 후 갱신된 H의 사후 확률

개념 2

● 베이즈 예제

특정 회귀질환에 대한 검사를 진행한다고 하자.
진단키트는 해당 희귀질환이 있는 사람을 검사할 경우 99 %의 확률로 양성반응을 보이고, 해당 질환이 없는 사람을 검사할 경우 1 %의 확률로 양성반응을 보인다. 이 희귀질환은 전체 인구의 0,1 %만 걸린 질환이다. 만약 질환의 발생 여부를 모르는 환자가 이 진단키트를 통해 검사한 결과 양성반응을 보였다면 그 환자가 실제 희귀질환을 가지고 있을 확률은 얼마인가?

풀이)

P(희귀질환에 걸렸을 확률) = 0.001

P(양성|희귀질환에 걸렸을 확률) = 0.99

P(양성) = 0.999 X 0.01 + 0.0001 X 0.99 = 0.01098

개념 3

● 나이브 베이즈 분류

나이브베이즈 분류 모형은 베이즈 정리를 기반으로 만들어진 지도학습 모형으로, 스팸 메일 필터링, 텍스트 분류 등에 사용된다.

나이브 베이즈는 데이터의 모든 특징변수가 서로 독립적이라는 가정하에 분류를 실행한다.

개념 4

● 딥러닝 분석

딥러닝은 인공신경망에 기반하여 설계된 개념으로 연속된 여러 개의 층을 가진 인공신경망을 통해 계층적으로 데이터를 학습시키는 방법이다.

딥러닝은 일반적으로 2개 이상의 은닉층을 가진 신경망을 통해 데이터를 학습하는 것을 말한다.

기존의 다층 퍼셉트론은 다음과 같은 문제점을 가지고 있다.

학습 데이터에 과적합되고, 은닉층의 수가 증가할수록 연산량이 폭증한다.

또한, 역전파 학습 과정에서 기울기 소실이 발생할 수도 있다.

개념 5

● 기존 신경망의 문제를 극복한 딥러닝

1. 사전학습

과적합이 발생하지 않게 신경망의 가중치와 편향을 초기화하는 방법이다.

2. 정규화

Regularization 혹은 정규화(일반화) 라고 하며, 패널티를 줌으로써 모델의 복잡성을 줄이는 방법을 말한다.

3. 드롭아웃

신경망이 복잡해지면서 가중치 감소만으로는 과적합을 방지하기가 힘들어지는데, 이때 사용하는 방법이 드롭아웃이다.

드롭아웃은 일정 비율의 뉴런을 임의로 정해 삭제하며 학습에서 배제하는 방법이다.

4. 배치 정규화

배치 정규화는 각 층의 출력값의 분포가 일정해지게 정규화 하는 방법으로 배치 정규화를 적용하면 각 층의 출력값은 정규분포를 따르게 된다.

5. 활성함수 변경

Sigmoid 함수와 같이 출력값의 범위를 한정시키는 활성함수를 사용함으로써 발생하는 기울기 소실 문제는 ReLU 등의 비선형 함수가 개발되면서 해결되었다.

개념 6

● 합성곱신경망(CNN)

CNN은 데이터의 특징을 추출해 이 특징들의 패턴을 파악함으로써 이미지 처리, 자연어 처리 등에 활용한다.

합성곱 층과 풀링 층으로 구성되어 있으며 합성곱 과정과 풀링 과정을 통해 분석이 진행된다.

1. 합성곱 과정

합성곱 층에서는 필터(커널)를 통해 전체 데이터를 스캔하며 대응하는 원소들의 곱을 모두 더한 값으로 출력한다. 이때, 필터의 이동량은 스트라이드라고 한다.

2. 풀링 과정

풀링은 합성곱 과정을 거친 데이터의 사이즈를 줄여주는 과정이다. 예시로는 맥스풀링 등이 있다.

CNN Feature Map 계산

원본 이미지 크기 : $n \times n$, 스트라이드 : s, 패딩 : p, 필터 : $f \times f$

이때, $Feature\ Map = (\frac{n+2p-f}{s} + 1, \frac{n+2p-f}{s} + 1)$

개념 7

● 순환신경망(RNN)

RNN은 문장이나 시계열 데이터와 같이 순차적인 형태의 데이터 학습에 최적화된 알고리즘이다.

기존의 신경망 모델은 각 변수가 독립적이라는 가정을 기반으로 하지만, RNN은 현재 결과과 이전 결과 사이에 연관성이 있다는 가정을 기반으로 하므로 음성, 문장, 시계열 데이터 등 순차적인 데이터를 다룰 수 있다.

개념 8

● 텍스트 마이닝

텍스트 마이닝이란 비정형 텍스트에서 특정 단어의 출현 빈도, 단어 간의 연관성, 단어의 긍정 부정의 방향성 등을 파악하고, 이를 통해 의미 있는 정보를 추출하는 방법이다.

개념 9

● 텍스트 마이닝 수행 단계

1. 데이터 수집

텍스트 데이터를 기사, 논문, SNS 등을 통해 수집하는 단계

2. 텍스트 전처리

수집한 코퍼스 데이터를 사용하려는 용도에 맞게 처리하는 단계다.

2-1 클렌징
코퍼스 데이터 내에 존재하는 노이즈를 제거하는 단계

2-2 토큰화
코퍼스를 토큰이라는 단위로 나누는 작업을 말한다.

2-3 불용어 제거
문맥적으로 큰 의미 없는 단어를 제거하는 단계

2-4 어간 추출
단어로부터 접사를 제거해 어간을 추출하는 단계

2-5 표제어 추출
표제어 추출이란 문장 속에서 다양한 형태로 활용된 단어의 표제어를 추출하는 것을 말한다.

3. 피처 벡터화

전처리된 데이터에서 문서별 단어의 사용빈도를 이용해 단어 문서 행렬을 생성하는 단계이다.

4. 텍스트 분석 및 시각화

피처 벡터화를 통해 숫자형으로 변환할 데이터를 이용해 분석 및 시각화를 하는 단계

개념 10

● 사회연결망 분석

사회연결망 분석은 사회를 관계성을 중심으로 설명하는 것으로, 개인, 집단, 사회의 관계를 네트워크로 파악하는 개념이다.

1. 중심성

중심성은 하나의 노드가 전체 연결망에서 중심에 위치하는 정도를 표현하는 지표

1-1 연결 정도 중심성

연결망 내에서 하나의 노드에 연결된 노드들의 합을 기반으로 중심성을 측정하는 방법

1-2 근접 중심성

각 노드간의 거리를 기반으로 중심성을 측정하는 방법

2. 밀도

연결망에서 노드 간의 연결 정도를 나타내는 지표

3. 중심화

하나의 연결망이 특정 노드에 집중되어 있는 정도를 보여주는 지표

개념 11

● 앙상블 분석

앙상블 분석이란 주어진 데이터를 여러 개의 학습용 데이터셋으로 분할하고 각각의 학습용 데이터셋을 통해 여러 개의 예측모형을 만든 후 여러 예측모형의 결과를 종합해 하나의 최종 결과를 도출하는 방법이다.

개념 12

● 배깅(bagging)

배깅은 Bootstrap aggregating의 줄임말로, 데이터셋에서 중복을 허용하여 랜덤하게 데이터를 추출하는 부트스트랩 방식을 통해 여러 개의 크기가 같은 표본을 추출하고 각 표본에 대해 예측 모델을 적용한 후 그 결과를 집계하는 방법이다.

개념 13

● 부스팅(Boosting)

부스팅은 예측력이 약한 모형을 연결하여 순차적으로 학습함으로써 예측력 강한 모형을 만드는 기법이다.

개념 14

● 랜덤포레스트

랜덤 포레스트는 배깅의 일종으로 배깅에 변수 랜덤 선택과정을 추가한 방법이다.

개념 15

● 비모수 통계

비모수적인 감정은 모집단의 분포를 가정하지 않고 분석을 실시하는 방법을 말한다.

개념 16

● 부호검정

부호검정은 중앙값을 통해 가설을 검정하는 방법이다.

개념 17

● 만-위트니 검정(윌콕슨의 순위함 검정)

만-위트니 검정은 독립된 두 집단의 중심 위치를 비교하기 위해 사용한다. 검정 방법은 아래와 같다.

step 1

두 모집단을 통합한 후 오름차순으로 정렬하여 가장 작은 값부터 순서를 매긴다. 이때 같은 값이 있는 경우에는 순위의 평균을 할당한다. 예를 들어, (1,2,3,4,4,5)의 순위는 (1,2,3,4.5,4.5,6)이다.

step 2

표본별 순위의 합을 계산한다.

step 3

표본별 U값을 계산한다. 표본 1의 U_1값은 다음과 같이 계산하고 표본2의 U_2값 또한 동일하게 계산한다. 최종 U값은 $min(U_1, U_2)$로 계산한다.

$$U_1 = R_1 - \frac{n(n+1)}{2}$$

step 4

U값을 이용해 두 집단의 차이를 검정한다. 만-위트니 U 테이블을 통해 임계값을 확인하고 U값이 임계값보다 작으면 귀무가설을 기각한다. 표본의 크기가 큰 경우 U는 근사적으로 정규분포를 따르는데, 이를 표준화한 z값을 통해서도 차이를 검정할 수 있다. z값을 구하는 식은 다음과 같다.

$$Z = \frac{U - nm/2}{\sqrt{nm(n+m+1)/12}}$$