SUMMARY

〈공공데이터가 민간데이터와 다른 특징〉

1) 공공데이터는 특정영역에서 독과점적인 위치 ⇒ 표본데이터가 비교적 모집단과 유사

2) 독과점적인 공공데이터의 비편향성 ⇒ 전국민의 데이터 확보 가능

〈공공 분야 빅데이터 활용 영역 및 행태〉

1) 공공 부문의 데이터가 활용하기에는 더 적절 but 의미 있는 결과를 산출하는 역량은 민간보다 떨어짐

2) 성공적으로 빅데이터가 활용됐다고 평가받는 분야가 의료, 범죄 등의 일부 영역으로 한정

〈공공 데이터 활용의 예 : 범죄 발생 방지〉

① 범죄에 대한 예측

- 범죄에 대한 예측의 세가지 기준 : 어디에서, 어떤 범죄가, 왜
- 범죄 예측의 핵심 : 범죄기록 데이터를 평가해 일관된 패턴 찾기
- 범죄 발생 가능성이 높은 고위험 지역 및 시간에 대한 예측 + 회귀분석 및 신경망분석 등의 분석기법과 조합 ⇒ 분석 대상의 상대적 위험 수준까지도 예측 가능

② 범죄자에 대한 예측

- 범죄자에 대한 예측의 세가지 기준 : 누가, 언제, 어떤 범죄를 저지를 것인지
- 재범 가능성이 높은 범죄자들의 미래 범죄 위험 가능성을 평가 및 예방
- 랜덤 포레스트라는 기법을 활용해 재범 가능성의 위험수준(없음, 경범죄, 중범죄)을 구분 ⇒ 획일적 감시 전략을 위험군별 감시전략으로 전환

③ 범죄자 신원에 대한 예측

- 이미 발생한 범죄가 누구의 소행인지 분석하는 활동
- 범죄 장소 및 범죄자를 프로파일링하고 향후 발생할 범행의 시간 및 장소 등을 예측

④ 피해자에 대한 예측

- 범죄 발생 가능성이 높은 지역에서 피해를 당할 가능성이 높은 피해자를 예측

〈공공 데이터 구축을 통한 범죄 예방 시스템의 시사점〉

1) 데이터 간 통합의 중요성

- 데이터들을 묶어서 예측의 정확도를 향상

2) 계량화된 빅데이터에 기반한 예측만으로 판단하는 것은 지양

- 비정형데이터 또한 연계하여 범죄 예방 시스템을 구축
- 컴퓨터가 하는 예측을 기계적으로 믿지 않고, 최종적인 판단은 모든 변수를 고려해 사람이 판단

3) 상호작용을 통한 예측모형의 정교화를 추구

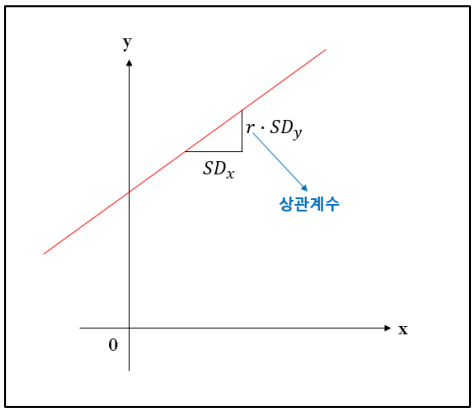
- 장기간의 데이터를 수집해 입력해 만든 모형을, 지속적으로 검토하는 방식을 도입 (모형-활동-평가)

4) 혁신을 수행하는 기관 자체의 '효율성 개선'외에 '추가적인 가치(고객 편익 증가 혹은 비용 절감)'도 제공

- 예측 치안을 통해 범죄 예방 외에도 시민들의 두려움 감소에도 효과적이어야 함

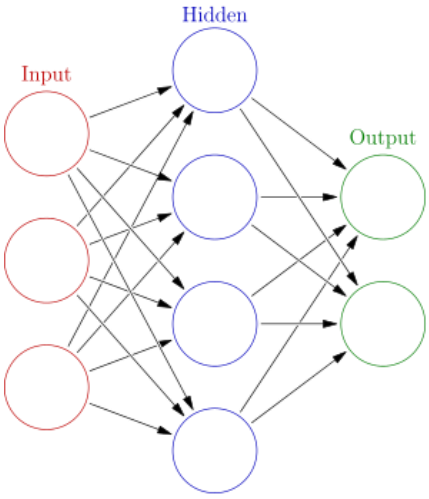
WORD

회귀분석(regression analysis)



1. 회귀분석(regression analysis) : 여러 변수간의 관계를 밝히기 위한 통계적 기법이다.
2. 단순회귀분석(simple regression analysis) : 하나의 변수와 다른 또 하나 변수간의 관계를 분석하는 방법
3. 중회귀분석(multiple regression analysis) : 하나의 변수와 둘 또는 그 이상 변수간의 관계를 분석하는 방법
4. 표준편차선(SD line) : 두 변수의 표준편차 값이 같은 점들을 이은 직선
5. 회귀직선 (regression line)
6. 두 변수를 x,y 라고 하고 상관계수를 r이라고 하자. x가 1SDx만큼 변화할 때, t 가 r*SDy 만큼 변함
7. 평균의 그래프 : x축이 키, y축이 몸무게라면. 키에 따른 몸무게의 평균을 연결한 그래프
8. 회귀직선은 평균의 그래프를 하나의 직선으로 근사시킨 것이라고 볼 수 있다. 만일 평균의 그래프 자체가 직선이라면 그 직선이 바로 회귀직선이다.
9. 평균의 그래프가 비선형이면 회귀직선으로의 근사는 부적절하다.

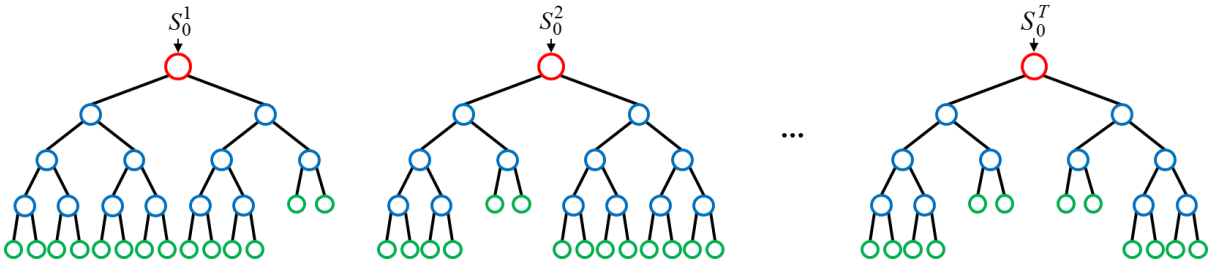
신경망 분석(Neural networks analysis)



: 인공지능 분야에서 신경망은 보통 인공신경망을 지칭한다. 인공신경망은 본질적으로 함수 f : X → Y,X에 대한 분포, 또는 X와 Y에 대한 분포를 정의하는 간단한 수학적 모델이고, 가끔씩 특정한 학습 알고리즘이나 학습 규칙과 긴밀하게 연계되어 있기도 한다. 인공신경망이라는 단어는 보통 이러한 함수들의 모임에 대한 정

의를 뜻하고, 이 모임의 구성원들은 식의 인자를 바꾸거나, 연결 가중치를 바꾸거나, 뉴런의 수나 연결 정도와 같은 구조에 대한 상세적인 것을 바꿈으로써 얻어진다.

랜덤 포레스트(random forest)



: 기계 학습에서의 랜덤 포레스트(영어: random forest)는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 부류(분류) 또는 평균 예측치(회귀 분석)를 출력함으로써 동작한다.

무죄추정의 원칙

: 피고인 또는 피의자는 유죄판결이 확정될 때까지는 무죄로 추정한다는 원칙으로 프랑스의 권리선언에서 비롯된 것이다.

프로파일링

: 자료수집'이 원 뜻이나 수사용어로는 범죄유형분석법을 말한다. 범죄 현장을 분석해 범인의 습관, 나이, 성격, 직업, 범행 수법을 추론한 뒤 이를 바탕으로 범인을 찾아내는 수사 기법이다.

비정형 데이터

: 형식이 정해지지 않은 데이터. 데이터는 형식이 정해진 정형 데이터(formal data)와 형식이 정해지지 않은 비정형 데이터가 있다. 페이스북, 트위터 등 소셜 네트워킹 서비스(SNS, 누리소통망 서비스)의 확산으로 데이터베이스에 잘 정리된 데이터가 아닌, 웹 문서, 이메일, 소셜 데이터 등 비정형 데이터가 주를 이루고 있다.

2종 오류(type Ⅱ error)

	진실	
의사결정 (Action)	올바른 의사결정	Type II 오류 (β) 잘못된 귀무가설을 기각하지 못하는 오류
	Type I 오류 (α) 올바른 귀무가설을 기각하는 오류	올바른 의사결정 (검정력: Power) 올바른 대안을 찾아내는 능력

: 통계적 가설검정시 발생하는 두 가지(1종, 2종) 가능한 오류 중의 하나로, 연구자가 상정한 영가설이 실제로는 참이 아님에도 이를 참이라고 수용할 확률이다.

ISSUE

[데이터 활용에 관해 : 심층 토론]

1-1. 본문에는 공공부분의 데이터가 민간부분보다 독과점 형태로 되어있어서 비편향적이지만 데이터를 수집해 의미 있는 결과를 산출하는 역량은 공공부문이 민간보다 떨어진다고 말하고 있다. 공공데이터가 민간데이터보다 편향되어 있지 않아 품질이 훨씬 좋은 데이터임에도 불구하고, 민간보다 활용하는 역량이 떨어지는 이유가 무엇일까? 여러 이유를 생각해보자.

1-2. 본문에는 데이터 활용에 있어서 4가지를 강조하고 있다.
① 데이터 간 통합의 중요성
② 계량화된 데이터에 기반한 예측만으로 판단하는 것을 지양
③ 상호작용을 통한 예측모형의 정교화
④ 기관 자체의 '효율성 개선' 외에 '추가적인 가치'도 제공

이러한 것들이 데이터 분석에 중요한 이유는 무엇일까? 또한 이 4가지가 적용되기 힘든 현실적인 문제는 무엇인가? (밑에 박스에 있는 본문 내용을 다시 한번 숙지 해보시기 바랍니다.)

첫째, 데이터 간 통합의 중요성이다. 여기서의 통합은 범죄와 관련해 경찰이 보유하고 있는 데이터들을 단지 묶어놓는 것만을 의미하지는 않는다. 경찰 보유 데이터와 타 부문이 가지고 있는 각종 데이터를 통합함으로써 기존 데이터만 활용한 예측보다 정확도를 향상시키고 지리학 데이터와의 결합을 통해 빅데이터 분석에서 가장 난해하다는 시각화까지도 해결하고 있다.이러한 데이터 통합은 경찰 입장에서는 통합이지만 경찰에 데이터를 제공하는 기관 입장에서는 일종의 데이터 개방이 일어나고 있음을 의미한다. 공공 부문의 부문 간 데이터 개방과 통합으로 새로운 가치가 발굴되고 있는 것이다. 흥미로운 점은 아직까지 어떤 국가도 통합을 위한 통일된 규칙은 제시하지 못한 반면 통합의 순서와 방향은 대체로 일치한다는 것이다. 데이터 통합은 대체로 과거의 범죄 데이터를 데이터베이스화하는 것을 1단계로, 여기에 신고·제보 데이터를 결합하는 2단계, 타 공공 부문이 확보하고 있는 인적사항 관련 데이터를 결합하는 3단계, 센서·CCTV 등 지역에 설치된(혹은 민간이 보유하고 있는) 인프라에서 확보되는 실시간 정보까지도 통합하는 4단계의 구조를 보이고 있다. 1~2단계가 경찰 내부의 과거 및 실시간 데이터를 통합하는 것이라면 3~4단계는 외부의 과거 및 실시간 데이터까지도 통합하는 단계라 볼 수 있다. 실시간 데이터의 빠른 입력과 이에 따른 신속한 처리 및 전송의 적시성 때문에 과거 데이터보다 통합되는 시점이 후순위로 나타나고, 개인정보 공유 등의 문제 해소가 필요하기 때문에 내부 데이터보다 외부 데이터의 통합이 후순위로 나타나는 경향을 보인다.

둘째, 빅데이터에만 의존한 판단은 하지 않는다는 것이다. 즉 계량화된 데이터에 기반한 예측만으로 판단하는 것은 지양한다는 얘기다. 여기에는 빅데이터를 활용한 범죄 예방 시스템 구축이 논의되고 시작됐던 시점에 비정형 데이터 활용 수준이 높지 못해 경찰들이 가지고 있던 '노하우'라는 비정형 데이터를 연계할 수 없었던 것이 하나의 원인으로 작용한다. 그뿐만 아니라 어떤 예측모형을 만들 때에는 그러한 예측을 하게 된 이유, 즉 이론적 근거를 기반으로 만들어야 하는데 범죄는 유발요인에 대한 명확한 규명이 쉽지 않아 모형화에 한계가 있다는 것도 중요한 이유였다. 결국 컴퓨터가 하는 예측은 말 그대로 분석일 뿐이지 평가가 아니므로 최종적인 판단은 사람이 수행해야 한다는 것이다.

셋째, 상호작용을 통한 예측모형의 정교화를 추구한다는 것이다. 범죄 유발요인에 대한 명확한 규명 없이 과거 데이터만으로 미래를 예측한다는 것은 상당한 위험을 내포하고 있다. 편향적 예측, 재범우려가 높은

위험도가 낮은 인물로 분류하는 예측모형의 오류(2종 오류) 등이 있을 수 있다. 또 범죄가 발생할 것으로 예측된 장소에 경찰력이 사전 대비를 할 경우, 해당 장소에서 범죄를 저지르려 했던 잠재적 죄인은 범행을 멈추기도 하지만 시간과 장소를 바꿔 범행을 재차 시도할 수도 있다. 이 경우 재차 시도하는 범죄에 대한 추가적인 예측은 기존 데이터만으로는 한계가 있을 수 있다. 따라서 빅데이터를 이용한 범죄 예측 모형은 장기간의 데이터를 일관적으로 수집해 입력하는 방식보다는 지속적으로 검토하는 방식을 활용하는 경향을 보인다. 그리고 모형의 개선뿐 아니라 실제 치안활동 운영 방식에 대한 개선도 동시에 수행하고 그 결과를 실시간으로 입력해 예측 치안의 방향성을 다시 검토하는 등 '모형-활동-평가 간 상호작용'을 통해 모형의 정확도를 개선하고 있다

넷째, 혁신을 수행하는 기관 자체의 '효율성 개선' 외에 '추가적인 가치(고객 편익 증가 혹은 비용 절감)'도 제공해야 한다. 범죄 예측은 기존보다 절감된 예산하에서 범죄 발생률을 억제하는 비용 효율적 특성 외에 사회 전반에서 '삶의 질 제고'라는 고객 편익 증가를 수반한다. 특히 일반 시민들의 삶의 질에는 범죄율이라는 통계값보다 '자신이 범죄의 피해자가 될 수 있다는 두려움'이 더 영향을 미친다. 예측 치안을 통한 취약 가능 지역에의 역량 집중과 이를 기반으로 수행하는 다양한 치안 활동 운영 전략들은 범죄 외에도 무질서나 두려움 감소에도 효과적인 것으로 알려져 있다. 실제로 범죄 예측을 치안 운영 전략에 활용하는 많은 기관들은 '비용 절감이나 범죄 해결뿐만 아니라 지역사회와의 전반적인 협력을 통한 삶의 질 향상 기여'를 운영 목표로 하고 있다.

[데이터 분석에 관해 : 활용 실습]

2. 본문에는 공공 데이터를 이용해 범죄 발생을 방지하는 프로세스를 자세하게 설명하고 있다. 본문에 나온 데이터분석 프로세스 외에 다른방법으로 범죄예방 기법을 만들 수 있을까? 아래에 있는 데이터를 활용해 머신러닝 기법으로 범죄 예방을 할 수 있는 예측 기법을 생각해보자. 어떠한 변수를 기반으로 어떠한 머신러닝 기법을 통해 어떠한 예측을 할 것인지 구체적으로 제시해야 한다.

활용 데이터
범죄가 발생하는 지역 및 빈도와 시기, 범죄자 범죄 경력, 과거수감기록, 마지막 중 범죄 후 경과 시간, 지금까지 일어난 범죄자의 직업, 나이 등과 같은 범죄자 신원, 범죄자가 범죄를 저지르기 직전에 썼던 기록, 지금까지 범죄자들이 남겨놓은 증거물들
(여기에 있는 데이터 외에 현대의 기술만으로 수집 가능한 모든 데이터들 변수로 사용 가능)

활용 예시 예 : 나이브베이즈 기법을 이용해 인터넷에 올라오는 글들을 분석한 후 범죄를 저지를 가능성이 있는 잠재적 범죄자를 예측

[머신러닝 기법들]

k-최근접 이웃 알고리즘
: k-NN은 함수가 오직 지역적으로 근사하고 모든 계산이 분류될 때까지 연기되는 인스턴스 기반 학습 또는 게으른 학습의 일종이다. k-NN 알고리즘은 가장 간단한 기계 학습 알고리즘에 속한다. 분류와 회귀 모두 더 가까운 이웃일수록 더 먼 이웃보다 평균에 더 많이 기여하도록 이웃의 기여에 가중치를 주는 것이 유용할 수 있다.

나이브 베이즈 분류
: 나이브 베이즈는 분류기를 만들 수 있는 간단한 기술로써 단일 알고리즘을 통한 훈련이 아닌 일반적인 원칙에 근거한 여러 알고리즘들을 이용하여 훈련된다. 모든 나이브 베이즈 분류기는 공통적으로 모든 특

성 값은 서로 독립임을 가정한다. 예를 들어, 특정 과일을 사과로 분류 가능하게 하는 특성들 (둥글다, 빨갛다, 지름 10cm)은 나이브 베이즈 분류기에서 특성들 사이에서 발생할 수 있는 연관성이 없음을 가정하고 각각의 특성들이 특정 과일이 사과일 확률에 독립적으로 기여 하는 것으로 간주한다.

의사결정 트리

: 결정 트리(decision tree)는 말그대로 결정을 내리기 위해 사용하는 트리로, 복잡한 문제를 간단한 문제들로 이루어진 계층 구조형태로 나누기 위한 기술이다. 간단한 문제에 대해서는 매개변수(예: 모든 노드의 테스트 매개변수, 종단 노드에서 매개변수 등)를 사용자가 직접 설정할 수 있지만, 보다 복잡한 문제의 경우 학습 데이터로부터 트리 구조와 매개변수를 자동으로 학습한다.

회귀 방법

: 통계학에서, 회귀분석(回歸分析, 영어: regression analysis)은 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한뒤 적합도를 측정해 내는 분석 방법이다.

신경망분석

: 인공지능 분야에서 신경망은 보통 인공신경망을 지칭한다. 인공신경망은 본질적으로 함수 $f : X \rightarrow Y$,X에 대한 분포, 또는 X와 Y에 대한 분포를 정의하는 간단한 수학적 모델이고, 가끔씩 특정한 학습 알고리즘이나 학습 규칙과 긴밀하게 연계되어 있기도 한다. 인공신경망이라는 단어는 보통 이러한 함수들의 모임에 대한 정의를 뜻하고, 이 모임의 구성원들은 식의 인자를 바꾸거나, 연결 가중치를 바꾸거나, 뉴런의 수나 연결 정도와 같은 구조에 대한 상세적인 것을 바꿈으로써 얻어진다.

서포트 백터 머신

: 서포트 벡터 머신(support vector machine, SVM은 기계 학습의 분야 중 하나로 패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만든다. 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다. SVM은 선형 분류와 더불어 비선형 분류에서도 사용될 수 있다. 비선형 분류를 하기 위해서 주어진 데이터를 고차원 특징 공간으로 사상하는 작업이 필요한데, 이를 효율적으로 하기 위해 커널 트릭을 사용하기도 한다.

연관규칙

: 분류 규칙 학습은 대규모 데이터베이스의 변수 간에 흥미로운 관계를 발견하기 위한 규칙 기반의 기계 학습 방법이다. 그것은 규칙적인 규칙의 개념을 사용하여 데이터베이스에서 발견된 강력한 규칙을 찾아내기 위한 것이다.

k-평균 군집화

: K-평균 알고리즘(K-means algorithm)은 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다. 이 알고리즘은 EM 알고리즘을 이용한 클러스터링과 비슷한 구조를 가지고 있다.