『 교통사고 통계자료 분석 기법연구 』

데이터 단순화 기법

|  |  |  |  |
| --- | --- | --- | --- |
| 주성분분석 | 원래 변수의 선형 결합으로 자료의 공분산 구조를 설명 |  |  |
| 요인분석 | p개의 관측 가능한 양적 변수들 사이에서 공분산 관계 설명할 수 있는 m개의 인자(가설적인 변수)를 찾는 다변량 분석법(m<p) | 즉 변수가 여러 개 있는 경우에 비슷한 특성을 지닌 변수들끼리 모아 몇 개의 집단으로 나눈 후 각 집단으로 대표할 수 있는 새로운 대표군인 인자를 찾는 것. | 원래의 변수에 의한 변동의 대부분을 훨씬 적은 수의 인자들을 이용해서 설명할 수 있게 됨 |

분석기법

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 회귀분석 | 거시적 교통사고량 추이 예측에 가장 많이 쓰인 기법 | 설명변수들간 관계에 따라 선형 / 비선형 | 교통사고량:  지역별(cross sectional data)/시간별(time-series)로 분석 가능, 시간과 지역 변화시킨 복합자료를 사용하기도 | 한계점:  현 상태의 지속성이라는 가정에 근거. 극복하기 위해서는 사고 모형에 정책변수를 포함시키는 노력 필요+교통사고 모형에 세분화된 설명변수들과 종속변수들을 포함한 개별모형(disaggregate models)의 개발도 중요 |
| 시계열분석 | 시계열 자료: 시간의 흐름에 따라 주기적으로 변화. 동일한 계절/월/분기/년 등의 시계열 관측값사이에 존재하는 높은 상관관계-계절성을 가짐-non stationary | Non-stationary 형태를 보정하기 위해서는 계절성을 제거하기 위한 차분방법 사용해 모형 만듦.  차분과정을 모형화 시켜 만든 시계열 모형: 계절성을 가진 ARIMA 모형 | 원래의 변수에 의한 변동의 대부분을 훨씬 적은 수의 인자들을 이용해서 설명할 수 있게 됨 |  |
| 판별분석 | 여러 개의 모집단에서 추출한 표본들이 지니고 있는 관측값 벡터들을 이용해 이 표본들이 어느 모집단에서 추출된 것인지를 판단하기 위해 가장 잘 분류해 줄 수 있는 기준, 잘못 분류될 확률을 최소로 하는 판별 함수를 찾는 방법 | 관측값 벡터들이 다변량 정규분포를 따른다는 가정을 요함. |  |  |
| 로지스틱 회귀분석 | “의 목적과 같음 | “의 가정 필요X |  |  |
| 대수선형분석 | 종속변수 또는 반응변수가 명의척도로 포아송분포를 따르는 범주형 자료로 표시된 변수들간의 관계를 특정 교호작용 항에 대한 모형의 검증을 실시해 변수들 사이의 관계를 분석하는 방법 | 일시,연령 등 제외한 대부분이 범주형 자료이므로 응용분야가 넓다 |  |  |
| Generalized Linear Model | 고전적인 선형모형은 포아송분포나 이항분포를 따르는 종속변수에 적합하지 않지만 GLM은 명의 척도로 측정된 종속변수까지 포용해 응용범위가 넓음 | 교통사고 수를 지역적으로 분석할 때 한 지역에서 발생한 사고수가 포아송분포 따른다는 가정 하에 평균사고량의 변화를 교통사고가 발생한 지점의 여러가지 환경요인으로 회귀분석할 수 있다 | 교통사고율 모형에 적용시 가능한 문제점: 과잉산포, 불확실성에 대한 추정, 서로 다른 교통사고 형태의 상관관계, 개별적인 사고발생율의 집합 예측 |  |
| 생존분석 | 관심있는 분석단위의 수명분포를 규명하는게 목적 | 비교분석단위들의 수명분포간에 차이가 있을 때 그들의 묘수를 분석단위와 관련된 주변 환경변수의 함수로 가정하고 모형추정 하면 관련 변수들의 수명 분포에 대한 영향력 및 수명 예측 가능해 교통사고 자료에 응용분야 넓음 |  |  |
| 베이지안분석 | 평균사고수에 대한 확률분포를 사고수가 실제로 주어진 상태에서 update한느 분석 |  |  |  |
| 수량화이론 | 명의척도나 순위척도로 조사된 자료에 더미변수와 같은 기계적인 수치를 부여하기보다는 최적의 데이터 해석이 가능하도록 수치를 부여하여 분석하는 것. Guttman의 예측이론으로부터 출발함 | 부여하는 수치는 질적 자료를 회귀분석에 적용해서 추정된 각 더미변수의 회귀계수와 유사 |  |  |

[주성분분석사례] Principal Component Analysis

1. 충화추출법으로 선정-네덜란드 32개 교차로, 312건 사고

설명변수: 교차로의 기하학적인 구조(크기,각도,안과밖의 폭,안쪽지름), 평균24시간동안 차량통행량(이륜차 유입량과 이출량,차량 유입량과 이출량,총 유입량과 이출량 등)…20개

종속변수: 사고의 종류(진입/회전사고,접근사고,단독사고,기타사고,보행자사고)

변수 가장 잘 대표하는 선형결합 제1성분,제2성분 구해 두 주성분을 축으로 원래 변수들의 관계구조를 밝힘

사고 종류 등 명의 척도를 scale해서 쓰는 정성적 주성분분석함

2. 부산시내 사고 다발 지점 45개

도로환경별 특성->7개 주성분, 각 사고 다발지점별로 7개 주성분별 값들을 구해 사고 지점을 유형화

[요인분석 사례] Factor Analysis

1. 사망사고 원인의 우선 순위 알아보기 목적. 서울 세 지역 136건 사망사고.

사고 가해 운전자의 문제 행동 요인을 39개의 질적 분석 단위(야간 교통 특성 무시, 과속, 심야, 좌우 탐색 불량, 전방 주시 태만, 좌우 후방만 주의..)로 측정

39개 변수들의 상호 상관성을 고려해 몇 개의 인자로 사망사고의 특성을 규명하기 위해 요인분석.

그 결과 사망사고의 위험 요인 순서: 심야의 과속, 도로 등의 불량, 관찰 요인 등

2. 14개의 도로의 환경에 대한 설문조사->커브운전을 할 때 속도에 영향을 주는 변수

요인분석 결과 커브에서의 속도 선택은 4가지 도로 환경요인(opposing traffic의 선정, cross-section 특성, alignment, signing)에 의해 설명.

속도 규제를 위반한 운전자들과 위반하지 않는 운전자들의 반응 비교는, 후자는 주로 커브에서 신호등 요인에 의해 그들의 속도를 정함을 보여주고, 반면에 전자는 주로 도로 배치 요인에 의해 결정됨

①주성분분석이나 요인분석을 통해 종속변수나 설명변수의 구분 없이 다량의 교통관계 변수들의 관계에 대한 파악을 한 후 ②교통사고량 또는 상해정도를 관련 변수로 모형화하기 위해 종종 쓰이는 회귀분석, 판별분석

[회귀분석 사례]

[시계열분석 사례]

교통사고 월별 자료를 Box-Jenkins 시계열 모형에 적용해 모형의 예측치를 분석하고 사고 원인별로 파악해 교통사고 발생의 모형 제시함

[판별분석 사례] Discriminant Analysis, Logistic Regression

1. 대형 교통사고의 요인들 중 인적 요인인 운전자의 법규위반 유형, 가해 운전자의 사망여부를 인적, 차량적, 도로 환경적 요인들에 대한 판별 함수를 이용해 분석

2. 3609가지 도로사고 진단코드로 각 신체부위의 상해 정도를 AIS점수로 바꾸고 다시 minor/major로 분류 후 상해자의 나이, 성별, road user type 의 판별함수로 로지스틱 회귀분석 이용해 상해정도를 판별함

3. 단독 사고에서의 모터사이클 운전자 사고정도 심각성(damage,injury..)을 5가지로 구분해 이들의 변화를 환경요소들, 도로컨디션,차종의 특성, 운전자 태도와 연관시키기 위해 multinomial-logit 모형 이용.

[대수선형분석 사례] Log-Linear Analysis

인적요인인 음주,과속운전 여부,보행자의 무단횡단에 따른 사망정도가 통계적으로 유의미한 차이가 있는지를 대수선형분석으로 비교함.

[GLM 사례] Generalized Linear Model

최소자승법 기초 전통 회귀분석보다 포아송분포 가정 GLM이 더 효과적임을 증명. GLM적용 교통사고율 모형의 문제점 해결방안도 제시

[생존분석 사례] Survival Analysis

차량을 단위로 사고 나기까지 걸린 시간을 수명이라 보고 생존분석 이용해 수명분포의 묘수를 사고와 관련한 요인들(사고가 일어난 계절, 시간, 트럭운전자의 나이, 화물의 무게, 사고가 나기 전

휴식을 취한 시간)의 함수로 모형화.

[베이지안분석 사례]

3800km 영국 고속도로의 교통사고자료

교통사고수는 Poiss 분포를 따르며 평균사고수를 사고환경 변수의 함수로 가정

사고환경 변수의 값이 동일한 조건으로 주어진다고

해도 평균 사고수가 반드시 일치하는 것은 아니므로, 평균사고수의 확률오차

에 대한 분포 가정을 필요로 했다. 이때 확률오차 분포의 모수는 알려져 있다

고 가정하지 않음으로, 모수 추정을 위해 Empirical Bayes절차를 밟았다. 이

러한 Poisson 모형은 6개의 고속도로 범주에 대한 평균사고의 회귀추정을 위

해 사용되었으며, Empirical Bayes절차는 교통사고건수의 추정을 향상하기 위

해서 사용되었다. 고속도로상하에서의 사고는 노출(교통흐름)과 Minor

Junction의 빈도에 비선형함수로 나타났다

[수량화이론 사례] Qualification Theory

7개 사고 유형별 사고율에 영향을 미치는 요인 분석

종속변수: 1백만 진입차량당 사고건수

설명변수: 25개의 교통관련 변수들

+사고 통계는 대부분 범주형 자료. 범주형 자료를 위한 다변량분석의 응용 기대.

질적자료에 대한 주성분분석, 정준상관분석, Correspondence Analysis

메타분석. 다양한 교통사고에 대한 연구결과들을 또 하나의 자료로 동일한 모수임에도 결과가 상이하다면 다른 이유가 될 만한 인자들의 함수로 모수를 재추정함으로써 그 원인을 설명하는 분석

교통관련 인자들의 함수로 사고 수를 포아손 회귀분석하여 통계적으로 유의한 관련인자를 찾아냄

『 DSRC 데이터를 활용한 k-NN 알고리즘 기반 교통사고 위험예측 기법 연구 』

연구의 공간적 범위: 대구광역시 도시부 간선도로

모형 검증: 대구 주요 간선도로 중 하나인 달구벌대로의 최상위 사고다발링크 대상

사고위험예측: 머신 러닝. 기존 문헌 고찰 통해 사용할 알고리즘 선정함

머신러닝- 경험적 데이터를 기반으로 학습하고 학습된 내용을 통해 예측 수행하는 알고리즘 -> 예측 기반이 되는 학습 데이터 구축: 2년간 대구광역시 사고데이터, 교통소통정보, 기상데이터

사고위험예측 알고리즘 개발-개발알고리즘의 검증 및 절차 설계-검증과정에서 파라미터 최적화하여 보정해 최적의 모형 도출

1. 연구의 배경 및 목적 선정

연구 범위 선정: 서울시 3개구

연구 방법 및 계획 수립

기존 문헌 고찰

2. 자료수집

3. 데이터 가공 및 통합 테이블 구축

4. 사고 위험 예측 모형 개발

변수 선정

변수의 중요도 평가

알고리즘에 필요한 적정 하이퍼파라미터 선정

5. 사고 위험 예측 모형 검증

6. 결론 및 향후과제

[회귀분석 단점]

한상진, 김근정, 오순미(2008)는 회귀분석을 이용한 사고예측모형은 다중

공선성으로 인해 중요한 변수가 누락될 가능성이 있고, 각 설명변수들과 사

고율 사이의 인과관계를 구하기가 어려우며 변수별로 함수식을 다르게 하

는 경우가 거의 없어 변수의 특성이 모형식에 반영하기 어렵다는 회귀분석

의 한계점에 대해 정리하였다.

[기존 사고예측 연구의 한계- 도로와 교통 조건 양호구간 사고에 대한 설명 미흡]

백승걸, 장현호, 강정규(2005)는 기존 사고 예측 관련 연구들은 주로 특정

지점에서의 도로기하 구조조건, 교통 및 환경조건들과 교통사고와의 관계를

설명하기 위한 모형의 개발에 초점을 두었으며, 그렇기 때문에 도로와 교통

조건이 양호한 구간에서의 높은 교통사고율을 설명하기 어렵다는 한계를

가지고 있다고 언급하였다.

[국내 사고예측 연구 모형 정확성↓ / 국외연구: 비선형 회귀분석 주, 다양한 변수]

박준태, 이수범, 이동민(2011)은 국내의 사고예측과 관련된 연구는 대부분

상관분석 등을 통해 교통사고를 설명할 수 있는 변수들을 선정하여 선형회

귀, 비선형회귀, 로지스틱 곡선, 음지수 분포 등의 회귀식과 회귀분석, 시계

열분석, 수량화 분석 등을 이용하여 변수들과 교통사고와의 관계를 규명하

였으나, 연구의 대부분이 사고 관련 자료의 불충분으로 인한 변수의 부족과

표본수의 부족, 사고건수 대신 상충수와의 비교 등으로 인하여 모형에 대한

정확성이 낮게 나타나고 있다고 언급하였다. 또한, 국외의 사고예측과 관련

된 연구는 비선형 회귀분석을 이용한 연구가 주를 이루었으며, 사고에 영향

을 미치는 변수를 고려함에 있어 보다 다양한 변수를 고려하고 있다고 하

였다.

[최근연구-미시적 분석(도로구간의 개별특성 분석), 인공신경망 등 / 한계점: 상관관계 높은 변수로만 모형 개발해 다양한 요인 반영 하지 못함+자료의 분산↑이라 모형 적용 어려움]

이재명, 김태호, 이용택, 원제무(2008)는 최근 사고예측 관련 연구에서는

도로구간의 개별특성 자료를 분석하는 미시적 분석을 도입하고 있으며 분

석기법도 확률회귀분석모형(포아송, 음이항 회귀분석) 또는 변수의 비선형

특성을 설명하는 데에 우수한 인공신경망모형 등을 다양하게 이용하고 있

다고 언급하였다. 그러나 기존 교통사고예측모형은 설명력을 높이기 위하여

교통사고와의 상관관계가 높은 변수로만 선형회귀모형을 개발하여 다양한

요인을 모형 내에 반영하지 못하거나, 교통사고의 비선형적 특성을 반영하

기 위해 비선형 모형을 이용하더라도 교통사고관련 자료의 분산이 매우 커

서 모형의 적용성에는 한계를 안고 있다는 문제점을 제시하였다.

[이항 로지스틱 회귀분석]

Pirdavani et al.(2014)는 루프검지기의 데이터와 실시간 교통 데이터, 사

고데이터를 통합하여 실시간 충돌위험예측 모델을 개발하였다. 예측모델은

이항 로지스틱 회귀분석을 사용하였으며, Confusion Matrix와 ROC Curve

를 이용하여 모형을 검증하였다. 모델의 검증 결과, 사고가 발생하는 경우

의 60% 이상을 예측할 수 있는 반면, 사고가 발생하지 않는 경우는 90%

이상을 예측하는 성능을 나타냈다.

[빈발패턴에 기반한 새로운 변수선택 방법(빈발패턴나무+랜덤포레스트)]

Lin, Wang과 Sadek(2015)는 실시간 교통사고 위험 예측 모형을 위해 빈

발패턴(Frequent Pattern)에 기반한 새로운 변수 선택 방법에 대해 연구하

였다. 빈발패턴나무와 랜덤 포레스트(Random Forest)를 이용하여 8가지 사

고발생에 위험을 주는 변수를 선정하였으며, k-NN 알고리즘과 베이지안

네트워크를 이용하여 사고예측을 통해 어떤 방법의 변수 선정이 예측률이

더 높았는지 확인하였다.

[k-NN알고리즘 > C-means알고리즘]

Lv, Tan과 Zhao(2009)는 k-NN 알고리즘을 활용하여 고속도로 교통사고

예측모형을 개발하였다. k-NN 알고리즘을 적용함으로써 교통사고의 유발

가능성이 더 높은 교통 조건을 식별하고, 사고 유발 요인이 교통사고 발생

에 미치는 영향을 고려하였다. ATMS 실시간 교통데이터와 사고이력자료

를 통합하여 분석데이터로 활용하였으며, C-means 알고리즘과의 예측모형

의 성능 비교 결과 k-NN 알고리즘이 더 우수함을 증명하였다. 또한, 이 연

구는 k-NN 알고리즘을 활용하여 교통사고 예측모형을 적용한 최초의 연구

라 할 수 있다.

[변수설정:랜덤포레스트 모델, 모형개발:SVM + k-means clustering

=>랜덤포레스트 변수설정+하이브리드모델 예측력↑↑]

Sun과 Sun(2016)은 랜덤 포레스트 모델을 사용하여 중요한 변수를 선정하

는데 사용하고, 서포트 벡터 머신(Support Vector Machine)과 k-means

clustering을 결합한 하이브리드 모델로 사고예측모형을 개발하였다. 변수

선택 없이 서포트 벡터 머신을 활용한 경우와, 변수를 선택한 서포트 벡터

머신을 활용한 경우, 서포트 벡터 머신과 k-means clustering을 결합한 하

이브리드 모델 총 3가지 경우를 비교하였다. 비교 결과, 하이브리드 모델과

변수를 선택하지 않았을 때 보다 랜덤 포레스트를 이용하여 변수를 선택한

경우의 예측 성능이 더 뛰어난 것을 확인하였다.

[국내:주로 모수적 기법 이용, 모수적 기법의 문제점 / 국외: 비모수적 기법 다수]

국내의 사고예측모형 관련 연구는 주로 모수적 기법을 이용한 연구가 다

수였으며, 모수적 기법의 문제점을 설명한 연구들이 있었으나, 비모수적 기

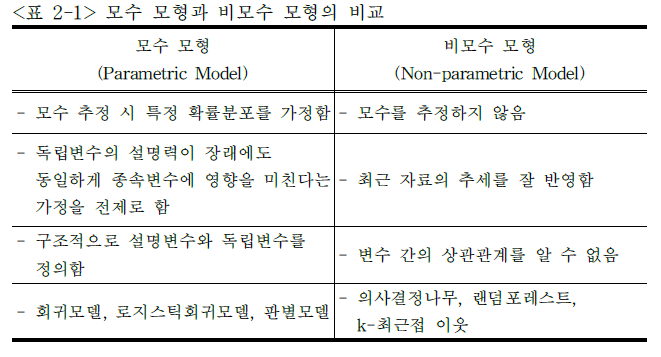
법을 활용한 연구는 많이 이루어지지 않았다. 국외의 사고예측모형 관련 연

구에서는 실시간으로 얻을 수 있는 교통데이터를 활용한 연구가 많았다. 교

통데이터, 사고 이력데이터 등 여러 가지 수집 데이터를 결합하여 사고위험

을 예측하는 연구를 하였으며, 다수의 연구가 비모수 기법을 이용하였다.

[k-NN 알고리즘] p.7~



[3장 자료구축] p.22~ 자료수집 과정에 나온 지도 참고하자~

[기상데이터 활용방안]

1. 날씨 좋을 때/비 올 때 사고의 심각도. 안개- 지리에 따라 사고의 심각도가 다름

2. 기상/교통사고 관계분석

로지스틱 회귀모형, 의사결정나무 모형

->교통사고 발생 경우 기온↑, 비ㅇ -> 강수 유무 + 기온

3. 기온↑, 습도↓, 강수량↑수록 교통사고 多

[데이터 전처리 과정] 예시

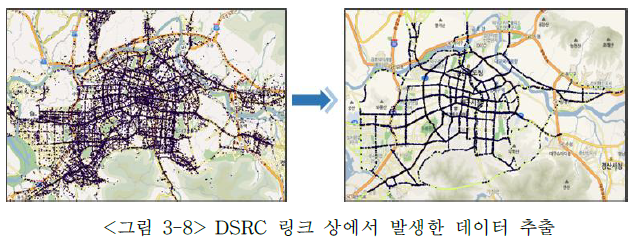
여러 데이터 하나의 통합 테이블로 구축하기 위해 공통 필드 맞추기 필요

동일한 일시(연, 월, 일, 시) 1시간 단위 기준으로 통합함(5분단위 데이터 평균함)

결측 데이터는 평균으로 산출되도록.

표준편차가 사고 유무에 큰 변수라는 기존 연구에 따라 표준편차 산출함.

각 사고발생이 어느 링크에서 발생했는지 파악위한 전처리 작업+링크 상에서 발생한 사고자료 추출하는 전처리 작업 필요 -> 우리도 필요(운행제한시설물 시작점~끝점.., 도로 위 사고/이상한 곳 사고.., 자동차 전용도로인지 구분할 때!)

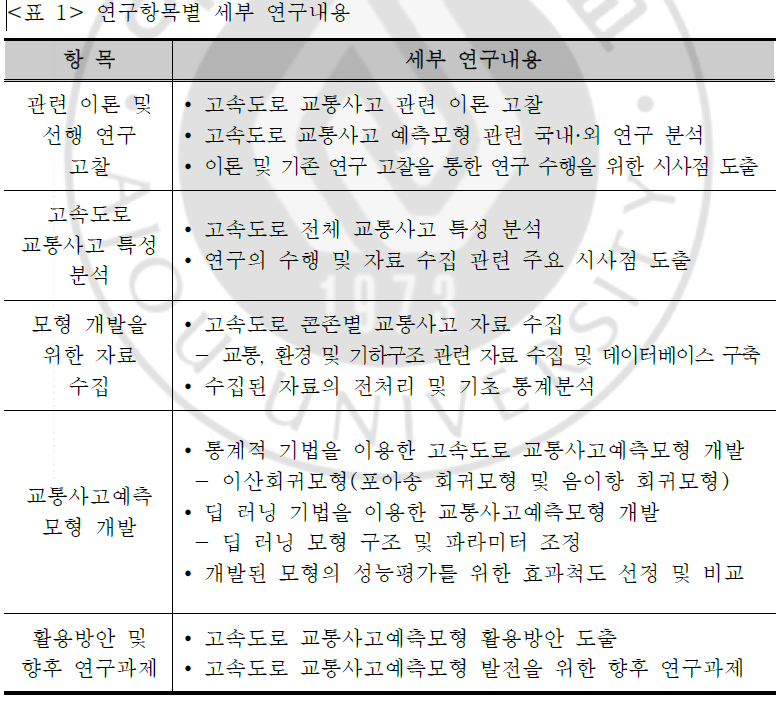


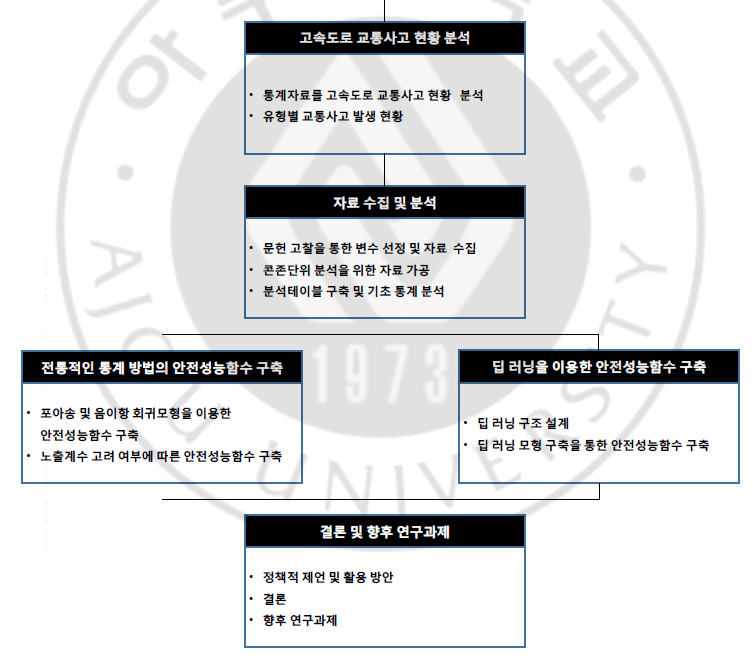
우리는 R로 표현한 정제 안된 데이터 찍은 지도 제안서에 보여주고 위 지도에서 도로가 아닌 부분?은 빼는 전처리 과정을 거치겠다 이런식으로?

k-NN은 k개의 데이터를 기반으로 예측 결과->중복 데이터는 가중 역할

[4장 사고위험예측 모형 설계]

『 딥 러닝을 이용한 고속도로 교통사고 예측모델 개발 』





다변량 분석의 경우 다중회귀분석,로지스틱,프로핏,신경망 ,SVM 같이

다수의 독립변수와 하나 또는 다수의 종속변수 사이의 관계를 파악하는 기법과

종속변수와 독립변수의 구분없이 데이터 특성을 논하는

주성분,군집,컨조인트 등등 의 방법이 있습니다.

보통의 데이터 형태는 아래와 같은 형태를 가지고 있습니다,.

이건 실제 분석을 진행했던 데이터 일부인데..

실무 데이터라 데이터 일부와 데이터 설명은 생략하겠습니다.^^

140 0 839 452 0

119 6702 878 452 16

115 7585 934 452 32.7

95 7611 956 452 57.9

117.25 5474.5 117 0 57.9

84 7012 986 452 87.4

113 7573 1011 452 113.1

107 7426 1033 452 141.2

표처럼 맨 오른쪽 값을 종속변수라고 하면 나머지 4개 변수와 독립변수간의 인과관계를 찾거나

위 5개 변수 모두를 포함하는 특성을 찾는 과정입니다

다중선형회귀: 수치형 설명변수, 연속형 종속변수, 선형가정, 회귀계수 추정

로지스틱 회귀: 범주형 종속변수

[머신러닝 기법 비교]

선형회귀 vs K-NN

선형회귀: 모수적 기법

K-NN: 비모수적 기법, 변수 많아질수록 오차↑

데이터 분포 양상과 모수적 모델 가정 모양 일치 경우 모수적 기법 성능↑

“ 불일치 경우 비모수적 기법 성능↑

LDA vs 로지스틱회귀

선형판별분석LDA: 범주 2가지, 선형식형태, 다변량 정규분포 확률함수

로지스틱 회귀: 선형식형태, 최대우도추정

데이터 가정 들어맞는 경우 LDA 성능↑

데이터 정규분포 따르지 않을 때 로지스틱 회귀 성능↑

서포트벡터머신SVM vs 로지스틱 회귀

손실함수가 비슷해 결과도 유사

QDA: 비선형식 형태

https://ratsgo.github.io/machine%20learning/2017/10/04/comparison/

<http://www.dodomira.com/2016/08/19/frequently_used_analyitic_method/> 통계방법론30가지