1. **머신러닝 및 인공지능 개요**

기계학습: 인간이 개발한 알고리즘을 컴퓨터 언어를 통해 기계에게 학습시키는 행위

알고리즘: Least square estimation , Gradient

컴퓨터언어: 파이썬 , R

다변량 데이터: 변수가 여러 개

변수는 독립변수/종속변수로 나누어짐

머신러닝 모델링: x와y의 관계를 설명하는 함수식을 찾는 것 -> 이 함수식 가지고 예측 가능

Training Data : 모델 구축 시 사용되는 데이터 + Testing Data : 구축된 모델을 검증하는데 사용되는 데이터

왜 딥러닝 ? : 딥러닝은 데이터의 수가 늘어날수록 성능이 선형적으로 끊임없이 증가함

1. **수치예측, 범주예측 (범류)**

예측 모델링은 종류가 수치예측 , 범주예측 2가지임. 예측 결과가 연속형 이면 수치예측 범주형이면 범주예측이다.

수치예측 데이터: 연속형->수치예측(Regression)

범주예측 데이터: 범주형->분류(Classification)

1. **머신러닝 모델 학습 프로세스**

다변량 데이터 : x(원인) , y(결과) / x변수들의 어떻게 조합하여 y를 표현할까??

X변수 와 Y변수 사이의 관계 – 확정적 관계(X변수만으로 Y를 100% 표현) / 확률적 관계 (100% 표현X)

Y=x로 설명되는 부분 + 그렇지 않은 부분 (오차)

오차 y-f(x) : 손실함수 / 손실함수를 다 더해주면 비용함수

비용함수를 최소화 하는 모델의 파라미터(모수,매개변수)를 찾는 과정임

다중선형회귀 모델은 Least square estimation algorithm 사용

로지스틱회귀 모델은 Conjugate gradient algorithm 사용

뉴럴네트워크 모델은 Backpropagation algorithm 사용

1. **선형회귀 모델 (개요,모델과정)**

Made by Galton

:출력변수 Y를 입력변수 X들의 선형결합으로 표현한 모델

:입력변수(x)와 출력변수(y) 평균과의 관계를 설명하는 선형식의 찾는 것

단순 vs 다중 / 선형 vs 비선형 -> 4가지

선형회귀 모델의 가정 : 확률오차에 대한 가정 -> 오차가 정규분포를 따름. / 평균 0 / 분산이 시그마 제곱임

1. **선형회귀 모델 (최소제곱법)**

비용함수를 최소화 시키는 파라미터를 찾자! -> 찾는 방법이 알고리즘임

이차함수 형태 전역 최적해 존재 -> 기울기가 0이 되는 경우 찾음

각 변수에 대해 편미분해서 0이되는 경우를 구하면 된다.

잔차(Residual) : 잔차e 는 확률오차가 실제로 구현된 값임

1. **선형회귀 모델 (파라미터 구간추정,검정)**

최소제곱법으로 구한 파라미터를 Least square estimator(추정량) 라고 함

추정량 : 샘플의 함수

추정량의 용도: 알려지지 않은 파라미터를 추정

추정량의 종류 : 점 추정 / 구간추정

점 추정

Gauss-Markov Theorm

Blue(best linear unbiased estimator) 의 조건 1) unbiased 2) variance 작음

최소제곱법 추정량 성질 : Gauss-Markov Theorm에 따라 좋음

구간추정

구간으로 추정하여 보다 유연한 정보 제공

점 추정값 – 상수값\*점 추정값의 표준편차<< 점 추정값 +상수값\*점 추정값의 표준편차

신뢰구간 구하자 = 구간 추정 하자

상수값: 유의수준 1-알파 하에서 자유가 n-2인 t분포의 값

기울기에 대한 가설검정 : 알려지지 않은 파라미터에 대한 가설을 세우고 이를 검증

귀무가설(기울기가 0이다) / 대립가설(아니다 기울기 0 아님) -> t(검정통계량) 검정 OR P Value 구하기

1. **선형회귀모델(R2,ANOVA)**

결정계수 -> R^2

SSE(에러에 의해) / SSR(X변수에 의해) / SST(SSE+SSR) (Y변수에 의해)/ 결정계수 : SSR 나누기 SST

SSE=0->R^2=1 -> 확정적 관계, 모든 관측치가 회귀직선 위에 있다 / SSR=0->R^2=0 ->X변수 가지고 Y 설명에 전혀 도움 안됨 / 평균으로 해도 차이 X

특징 0과1사이

정의 : 사용하고 X변수가 Y변수의 분산을 얼마나 줄였는지 정도 or 단순히 y의 평균값을 사용했을 때 대비 X 정보를 사용함으로써 얻는 성능향상 정도

수정 결정계수 (Adjusted R^2)

R^2 는 유의하지 않은 변수가 추가되어도 항상 증가하는 문제점 있음 -> 수정 R^2는 앞에 특정 계수를 곱해 줌으로써 유의하지 않은 변수가 추가 될 경우 증가 X / 항상 R^2 보다는 작거나 같음

ANOVA (분산분석) – analysis of variance

궁극적으로 가설검정을 행하는 용도로 사용됨

사실 SST,SSR,SSE -> 분산임

SSR 나누기 SSE

1보다 크면 : X변수가 Y예측에 유의미한 영향 미침 / X변수의 계수(기울기)가 0이 아님

0과 1사이: 영향 X / 기울기가 0임

위 비율의 분포는 F분포를 따름

단순선형 회귀모델의 경우

SSR- 카이제곱 분포 (자유도=1) SSE-카이제곱 분포 (자유도 n-2)

F 값이 크면 귀무가설에서 주장한 기울기 값이 0이라는 주장 기각

1. **로지스틱 회귀모델(로지스틱함수,승산)**

선형회귀 모델 y값이 연속적인 형태 -> y값이 범주형 데이터일 경우 사용 (범주예측/분류)

y값이 베르누이 확률의 기대 값으로 표현이 되는 모델임 / y값이 1을 가질 확률.

S curve 모형 -> 로지스틱 함수 : 결과값이 0과 1사이임 / large input->small output / input에 대한 단조증가(or 단조감소) / 미분결과를 output의 함수로 표현 가능 -> Gradient learning method에 유용

로지스틱 함수에서는 파라미터의 해석이 직관적이지 X

승산 : 성공 확률 / 실패 확률 ex) 월드컵 배당금

로지스틱 회귀모델의 승산 : 범주 1에 속할 확률 / 범주 0에 속할 확률 -> 이 승산에 로그를 취하면 단순한 선형 결합 형태로 바뀜. (이렇게 odds를 구하고 로그를 취하는 변환을 로짓 변환이라고 함 )

이렇게 되면 파라미터인 B는 x가 한단위 증가 했을 떄 log(odds)의 증가량으로 정의됨

1. **로지스틱 회귀모델(파라미터 추정, 결과 및 해석)**

로지스틱 회귀모델 파라미터 추정하는 방법 : 최대 우도 추정법 (MLE)

가장 먼저 log likelihood 함수를 만든다 . -> 목적은 loglikelihood 함수가 최대가 되도록 하는 파라미터 찾자 -> B는 비선형이므로 수치 최적화 알고리즘 사용 (iterative reweight least square, conjugate gradient, Newton’s Method)

cross entropy : 두 확률분포 p(x),q(x) 의 차이 / 음의 loglikelihood function 기대값

즉 Loglikelihood 함수 최대 = cross entropy 최소

이진 분류를 위한 기준값 설정

* 일반적으로 0.5
* 성공 범주의 비중이 낮을 때 0.3

승산 비율(Odds Ratio 해석 가능) : 나머지 입력변수는 모든 고정시킨 상태에서 한 변수를 1단위 증가 시켰을떄 변화하는 Odds의 비율 / 1보다 크면 확률이 증가하는 방향. Coefficient >0 -> Odds Ration 증가 ( 즉 성공확률 증가)

1. **의사결정나무모델(모델개요,예측나무)**

데이터에 내재되어 있는 패턴을 변수의 조합으로 나타내는 예측/분류 모델을 나무의 형태로 만드는 것 ex) 스무고개

장점: 이해하기 쉬움 / 이산.범주 변수 둘다 가능

단점: 한 번에 변수 하나씩 사용 -> 변수들간 관계 파악 x

개요 : Data->Algorithm->Model

Data: 다변량 데이터

Algorithm: 데이터를 2개 혹은 그 이상의 부분집합으로 균일해지도록 분할

Model: 나무를 거꾸로 세운 형태

이진 분할 : 끝 마디 개수 = 최종 부분집합 개수

-예측나무 (Regression Tree) : 이진함수(조건 True 시 1 / False 시 0) 를 이용해 새로 들어온 값이 어디 위치에 있는지 표현 가능

모델링 프로세스

최상의 분할은 비용함수를 최소로. -> 각 분할에 속해 있는 y값들의 평균으로 예측했을 때 오류가 최소

분할변수(j)와 분할점(s)은 어떻게 결정 -> 가능한 경우를 대해보고 비용함수 최소인 경우 선택

1. **의사결정나무모델(분류나무, Information gain)**

-분류나무

각 관측치 마다 반응변수 값, 즉 K개의 클래스(범주) 존재 / Rm: 끝 노드 m에 해당하며 Nm 관측지 개수를 가지고 있음 / ^pmk: 끝 노드 m에서 k클래스에 속해 있는 관측치의 비율

최종적으로 끝 노드 m으로 분류된 관측치는 k(m) 클래스로 분류

모델링 프로세스

비용함수 형태

1) Misclassification rate : 1-p

2) Gini Index : 시그마 p(1-p)

3) Cross-entropy : -시그마 p\*log(p)

분할변수(j)와 분할점(s)은 어떻게 결정 -> 가능한 경우를 대해보고 비용함수 최소인 경우 선택 / 불순도의 감소가 최대가 되도록 선택

Information Gain: 특정 변수를 사용했을 때 entropy 감소량 -> 크면 중요한 변수임

CART -> 불순도 계산법 by Gini index / by Deviance

개별 트리 모델의 단점

계층적 구조로 중간에 에러발생시 다음단계로 전파 / 미세한 변동에도 최종결과 영향 / 노이즈 영향 큼 / 최종노드 개수를 늘리면 과적합 위험 큼

* 해결 방안 : Pruning(가지치기) / 랜덤 포레스트

1. **랜덤 포레스트 모델**

랜덤 포레스트 배경 -앙상블

앙상블: 여러 Base 모델들의 예측을 다수결 법칙 또는 평균을 이용해 통합하여 예측 정확성 향상시키는 방법

앙상블 모델의 우수성 조건 : base 모델들이 독립적 / base 모델들이 무작위 예측(ex-binary는 0.5) 수행하는 모델 보다 성능이 좋아야 함

랜덤 포레스트는 base 모델로 의사결정나무를 사용하는 것 !! – 활용도가 높음 why? 데이터의 크기가 방대한 경우에도 모델 빨리 구축 가능 / 데이터 분포에 대한 전제가 필요 X

핵심 아이디어 : Diversity, Random 확보

1. 여러 개의 Training data를 생성하여 각 데이터마다 개별 의사결정나무모델 구축 -> Bagging
2. 의사결정 나무 모델 구축 시 변수 무작위로 선택 -> Random subspace

Bagging(Bootstrap Aggregating)

Bootstrap->샘플링 기법 (복원추출+원래 데이터 수만큼 샘플링)

Aggregating->합쳐 주는 것 ex)Majority / Weighted / 확률

2.

Random subspace : 의사결정 나무의 분기점을 탐색할 때,원래 변수의 수보다 적은 변수를 임의로 선택하여 변수들만의 고려 대상으로 함

랜덤 포레스트 특성

각각의 개별 tree는 과적합 될수 있음 -> Random forest는 tree수가 충분히 많아 큰수의 법칙에 따라 과적합 되지 않고 그 에러는 limiting value에 수렴됨

선형회귀모델/로지스틱 회귀모델과 달리 개별변수가 통계적으로 얼마나 유의한지에 대한 정보 제공 X -> 간접적인 방식으로 변수의 중요도를 결정

1. 원래 데이터 집합에 대해서 Out of bag(OOB) Error를 구함

-OOB : Bootstrap에 포함되지 않은 데이터

2. 특정 변수의 값을 임의로 뒤섞은 데이터 집합에 대해서 OOB Error를 구함

3. 변수의 중요도는 1,2단계의 OOB Error 차이의 평균이 클수록 중요함

랜덤 포레스트 – 하이퍼(user들이 결정) 파라미터

1. Decision Tree 수 (2000개?)
2. 2. 노드 분할 시 무작위로 선택되는 변수의 수 (분류: 변수의수의 제곱근 / 예측:변수의수/3 )