가트너 3V

- 규모 (Volume) :
- 유형 (Variety) : 데이터 종류, 근원 확대
- 속도 (Velocity) : 데이터 수집, 처리속도 변화

+2V

- 품질 (Veracity) : 데이터 신뢰성, 정확성, 타당성 필수
- 가치 (Value) : 데이터 속 가치 창출

데이터베이스 관리 시스템(DBMS)

빅데이터 저장

· NoSQL (Not-only-SQL)

- 비관계형 DB
- 데이터 업데이트 즉시 가능
- ex) Hbase, Cassandra, MongoDB

· 분산 파일 시스템

- 네트워크로 여러 파일에 접근 가능
- 데이터 추출 및 가공 시 빠르게 처리
- ex) GFS(Google File System), HDFS(Hadoop Distributed File System), Amazon S3

· 병렬 데이터베이스 관리 시스템 (PDBMS)

- DB 처리를 동시에 수행
- 확장성 제공을 위해 트랜잭션 적용 필요
- ex) VoltDB, SAP HANA, Vertica, Greenplum, Netezza

빅데이터 처리

· 분산 병렬 컴퓨팅

- 다수의 컴퓨팅 자원을 네트워크 상에 연결하여 미들웨어로 제어하여 하나의 시스템으로 동작하게 하는 기술
- 전제 작업을 잘 쪼개어 여러 작업으로 나눠야함

· 하둡 (Hadoop)

- 분산 처리 환경에서 대용량 데이터 처리 및 분석을 지원하는 오픈 소스 소프트웨어 프레임워크
- 하둡 분산파일시스템인 **HDFS**와 분산칼럼 기반 데이터베이스인 **Hbase**, 분산컴퓨팅 지원 프레임워크인 **맵리듀스**로 구성됨

· 아파치 스파크 (Apache Spark)

- 실시간 분산형 컴퓨팅 플랫폼
- In-Memory 방식으로 처리
- 하둡보다 처리 속도 빠름
- 스칼라 언어로 개발되었지만 Java, R, Python 지원

· 맵리듀스 (MapReduce)

- 구글에서 개발한 빅데이터 신속 처리 프로그래밍 모델
- 효과적인 병렬 및 분산 처리 지원
- 런타임에서의 입력데이터 분할, 작업 스케줄링, 노드 고장, 노드 간의 데이터 전송 작업이 맵리듀스 처리 성능에 많은 영향을 미침

맵리듀스 처리단계

1단계 입력데이터 읽고 분할

2단계 분할된 데이터 할당하여 맵 작업 수행 후, 중간 데이터를 통합 및 재분할

3단계 통합 및 재분할된 중간 데이터를 셔플

4단계 셔플된 중간 데이터를 이용해 리듀스 작업 수행

5단계 출력 데이터 생성, 맵리듀스 처리 종료

K-익명성 : 주어진 데이터에서 같은 값이 적어도 K개 이상 존재하도록 하여 쉽게 다른 정보로 결합할 수 없도록 함

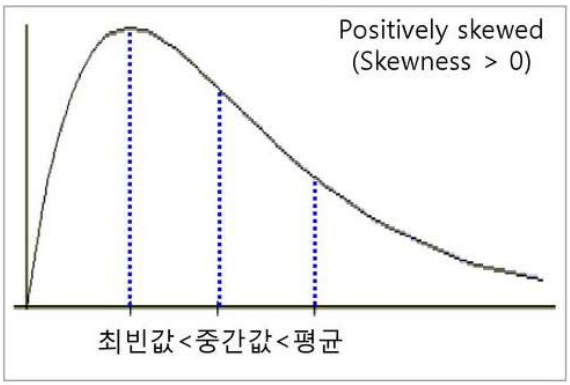
L-다양성 : 주어진 데이터에서 함께 비식별되는 레코드들은 동질 집합에서 적어도 L개의 서로 다른 정보를 가지도록 함

· 왜도(Skewness)

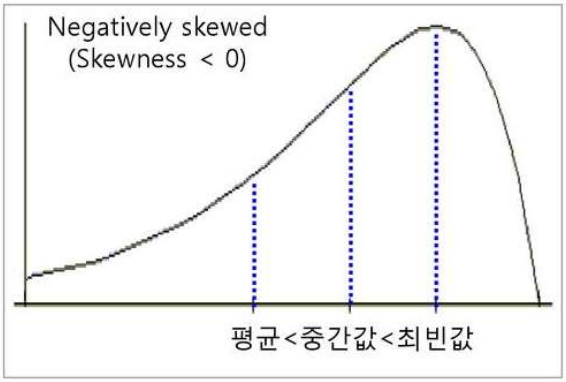
- 자료의 분포에 대한 비대칭의 정도
- 표준정규분포의 왜도 = 0

$$Skewness = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

오른쪽으로 긴 꼬리



왼쪽으로 긴 꼬리



· APRIORI 알고리즘 (연관규칙분석)

- 복잡한 계산을 줄이기 위해 최소 지지도 이상을 갖는 항목집합(빈발항목집합)을 추출
- 상품 간의 많은 연관 규칙 발견 가능
- 쉽게 이해할 수 있는 간단한 원리
- 상품 수가 많을수록 계산량이 기하급수적으로 늘어남
- 중요한 현실적 연관 규칙 발견 어려움

· MSE (Mean Squared Error) 평균 제곱오차

예측값과 실제값 차이의 제곱에 대한 평균

수치가 작을수록 정확성이 높은 것

· MAE (Mean absolute Error) 평균 절대오차

실제값과 측정값의 차이에 대한 평균

· 혼동행렬

		실제	
		Y (1)	N (0)
예측	Y (1)	True Positive (TP)	False Positive (FP)
	N (0)	False Negative (FN)	True Negative (TN)

정확도 (Accuracy)

- 정확히 예측한 수 (TP, TN)를 전체 샘플 수로 나눈 값
- Accuracy = (TP+TN)/(TP+FP+TN+FN)

정밀도 (Precision)

- 양성으로 예측된 것(TP, FP) 중 얼마나 많은 샘플이 진짜 양성(TP)인지 측정한 것
- Precision = TP/(TP+FP)

재현율 (Recall) (=민감도)

- 실제 양성 샘플(TP, FN) 중 얼마나 많은 샘플이 양성(TP)로 분류되는지 측정한 것
- Recall = TP/(TP+FN)

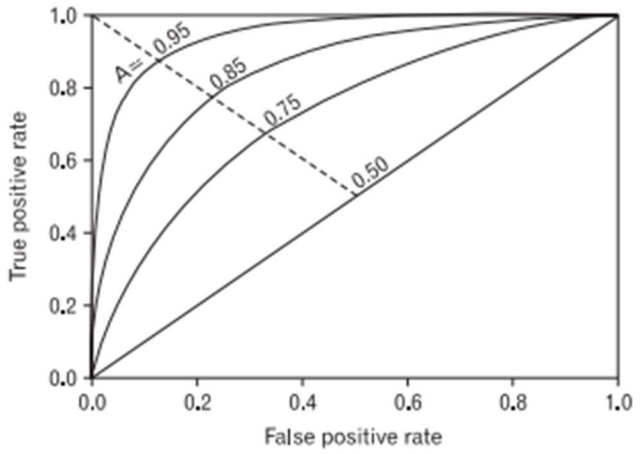
F1 Score

- 정밀도와 재현율을 결합한 조화평균 지표, 값이 클수록 모형이 정확함
- F1 = 2*(precision*recall)/(precision+recall)

특이도 (Specificity)

- 실제 음성 샘플(TN, FP) 중 검출된 음성 개수
- 재현율(민감도)와 특이도는 반비례함
- ex) 검사항목 모두 양성으로 판정 시 민감도는 1, 특이도는 0이 됨
- Specificity = TN/(TN+FP)

ROC 그래프



TPR (True Positive Rate 민감도)

- T를 T로 잘 예측할 확률
- ex) 암환자를 암이라고 진단함

FPR (False Positive Rate = 1-특이도)

- F를 T로 잘못 예측함
- ex) 암환자가 아닌데 암이라고 진단함

· 교차검증

· Hold-out 교차검증

데이터셋을 훈련/테스트 or 훈련/테스트/검증 세트로 나눔

· k-fold 교차검증

k개의 서브셋으로 분리

하나의 서브셋만 테스트에 사용하고 나머지 k-1개의 서브셋은 훈련에 사용함

이 과정을 k번 반복함

K번 측정한 성능 지표를 평균내어 최종적으로 모델의 성능을 평가함

오버피팅 위험이 크지 않음

시간이 오래 걸림

· 변수 변환

최소최대 정규화 (Min-Max Normalization)

최소값 0, 최대값 1

Z-Score 정규화

- 데이터가 평균과 일치하면 0으로 정규화
- 평균보다 작으면 음수, 크면 양수
- 표준편차에 의해 결정됨
- 이상치 처리를 잘함
- 정확히 동일한 척도로 정규화된 데이터를 생성하지 않음

로그 변환 : 데이터가 우측으로 치우쳤을 때,

로그를 취하면 정규분포에 가깝게 분포됨

역수변환 : 극단적으로 우측으로 치우쳤을 때

지수변환 : 좌측으로 치우쳤을 때

제곱근변환 : 우측으로 약간 치우쳤을 때

· 상관분석 방법

피어슨 상관계수

- -1 ~ +1 사이의 값
- +1은 양의 상관관계
- 0은 상관관계 없음
- -1은 음의 상관관계

스피어만 상관계수

- 서열자료인 경우 이용하는 상관계수
- 두 변수의 차이가 클수록 상관계수 값 커짐
- 1에 가까울수록 변수 간에 단조적 상관성을 가짐
- 0에 가까울수록 상관성 없음

· 기초통계량

$$\text{산술평균} = \frac{a+b}{2}$$

$$\text{기하평균} = \sqrt{ab}$$

$$\text{조화평균} = \frac{1}{\frac{\frac{1}{a}+\frac{1}{b}}{2}} = \frac{2ab}{a+b}$$

ex) 다기간 수익률에 대한 평균 수익률, 평균물가상승률(기하평균), 변화율(조화평균)

각 자료 동일 : 조화평균=기하평균=산술평균

다를 경우 : 조화평균<기하평균<산술평균

사분위범위 (IQR)

- 1사분위수 (Q1) : 자료의 1/4
- 3사분위수 (Q3) : 자료의 3/4
- 사분위범위 (IQR) : Q3-Q1 (자료의 50% 범위 내에 위치함)
- 최대값 = Q3 + 1.5 * IQR
- 최소값 = Q1 – 1.5 * IQR

변동계수 (CV)

- 평균을 중심으로 한 상대적인 산포의 척도를 나타내는 수치
- CV = 표준편차/평균 * 100(%)

· 확률변수

이산확률변수 : 값이 유한한 변수 ex) 동전, 주사위

연속확률변수 : 값이 무한한 변수

· 확률분포

이산확률분포

- 확률변수가 취하는 값의 수가 유한한 확률분포
- 확률질량함수 : 특정값에 대한 확률을 나타내는 함수

베르누이 분포 : 결과가 성공 or 실패 두가지

포아송 분포

- 단위시간이나 단위공간에서 무작위로 발생하는 사건의 발생횟수에 적용되는 분포
- 시행횟수 n이 크고 사건의 발생확률 p는 매우 작은 경우에 사용됨
- 아래 두가지 조건을 만족하면 발생횟수라는 확률변수는 포아송 확률분포를 따름

1) 두 구간의 길이가 같다면 발생확률이 동일함 (**일정성**)
2) 어떤 구간에서 발생하거나 발생하지 않는 사건은 다른 구간에서 발생하거나 발생하지 않는 사건과 독립임 (**독립성**)

λ = 기대값 = 분산 = 단위시간당 평균발생횟수

초기하분포

- 비복원추출에서 N개의 모집단 중 n개를 추출할 때 k번의 성공을 할 확률에 대한 분포
- 확률변수는 위 조건 하에 성공할 횟수
- 앞선 시행이 다음 시행의 확률에 영향을 줌

이항분포

복원추출을 통해 매 시행이 독립적임

기하분포

- 베르누이 시행에서 처음 성공까지 시도한 횟수를 분포화한 이산확률분포

음이항분포

- x번의 베르누이 시행에서 k번째 성공할 때까지 계속 시행하는 실험에서의 확률을 나타내는 이산확률분포

분포	확률밀도(질량) 함수	평균, 분산	적률 함수
이산형 균등분포	$P(X=x_i)=\frac{1}{n}, i=1,2,\dots,n$	$E(X)=\frac{n+1}{2}, V(X)=\frac{n^2-1}{12}$	
베르누이 분포 X ~ Ber(p)	$P(X=x)=p^x(1-p)^{n-x}$, x=0, 1	E(x) = p, Var(x) = p(1-p)	$M(t)=(1-p)+pe^t$
이항 분포 X ~ B(n, p)	$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$	E(x) = np V(x) = np(1-p)	$M(t)=[(1-p)+pe^t]^n$
초기하 분포	$P(X=x)=\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$	$E(X)=n\frac{D}{N}, Var(x)=np(1-p)\frac{N-n}{N-1}$	
음이항분포	$f(x\|\theta)=\binom{n+r-1}{x}\theta^r(1-\theta)^x$	$E(x)=\frac{r(1-\theta)}{\theta}, V(x)=\frac{r(1-\theta)}{\theta^2}$	
포아송 분포 X ~ Poisson(λ)	$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!}$	E(x) = λ V(x) = λ	$M(t)=e^{\lambda(e^t-1)}$

연속확률분포

- 확률변수가 취하는 값의 수가 무한한 확률분포
- 확률밀도함수 : 확률변수의 분포를 나타내는 함수

연속균등분포

분포가 특정범위 내에서 균등하게 나타나 있는 경우

지수분포

사건이 서로 독립일 때, 한 개의 사건이 발생할 때까지의 대기시간에 대한 확률이 따르는 분포

(*포아송분포는 횟수)

표준정규분포

평균 0, 표준편차 1

기댓값의 성질

a,b,c가 임의의 상수이고, X, Y가 확률변수,

(1) $E(c)=c$

(2) $E[cX]=cE[X]$

(3) $E[X\pm Y]=E[X]\pm E[Y]$

(4) $E[aX\pm bY]=aE[X]\pm bE[Y]$

분산의 성질

a,b,c가 임의의 상수이고, X, Y가 확률변수,

(1) $V(X)=E[(X-\mu)^2]=E[X^2]-\mu^2$

(2) $V(aX+b)=a^2V(X)$

※ X, Y의 공분산이 존재할 때,

(3) $V[X+Y]=V(X)+V(Y)+2Cov(X,Y)$

(4) $V[X-Y]=V(X)+V(Y)-2Cov(X,Y)$

(5) $V[aX+bY]=a^2V(X)+b^2V(Y)+2abCov(X,Y)$

★ X, Y가 독립이라면 $Cov(X,Y)=0$이다.

데이터 시각화

- **정보 시각화** : 카토그램, 분기도, 개념도, 계통도(덴드로그램), 트리맵, 하이퍼볼릭 트리, 네트워크 다이어그램
- **시간 시각화** : 누적, 묶은 막대그래프, 점그래프, 꺾은선 그래프, 계단그래프, 추세선
- **분포 시각화** : 히스토그램, 파이차트, 도넛차트, 트리맵, 누적연속그래프
- **관계 시각화** : 산점도, 버블차트,
- **비교 시각화** : 히트맵, 체르노프페이스, 스타차트(레이더차트), 평행좌표계, 다차원척도법
- **공간 시각화** : 단계구분도, 카토그램

확률 표본추출 기법

· 단순무작위 추출

- 편향성 제거
- 모집단에 대한 사전지식이 많지 않은 경우 시행함

· 계통추출

- 추출간격을 설정하여 간격 사이에서 무작위로 추출
- N개인 집단에서 K라는 추출 간격으로 뽑는다면 N/K 수만큼 표본이 선택됨

· 층화추출

- 모집단을 서로 겹치지 않게 동질적인 여러 층으로 나누어 분할된 층별로 표본을 단순임의 추출에 따라 추출
- 특성치의 효율적인 추정이 필요한 경우 시행
- 단순임의 추출에 비해 추정의 정도를 높일 수 있음
- ex) 모집단 남녀 성비 3:2일 때 표본 성비도 3:2로 추출할 경우

· 군집추출 (집락추출)

- 서로 인접한 데이터를 묶어 군집으로 구분하여 일부 군집을 추출
- 단순 임의추출에 비해 표본 오차가 클 수 있음

비확률 표본추출 기법

· 간편추출법

- 조사원의 주관적 판단에 따라 간편한 방법으로 표본추출
- 표본이 모집단을 잘 대표하는지 알 수 없음

- 통계적 정확성 또한 평가할 수 없음
- ex) 길거리 설문조사

· 판단추출법

- 조사자 지식과 경험에 의해 주관적으로 선정하는 방법
- 추정치의 정확성에 대해 객관적으로 평가 불가
- 표본 크기가 작을 경우 조사의 오차를 좌우하는 요인은 추정량의 분산이 될 수 있음
- ex) 전국 학생들의 평균 성적을 조사하기 위해 학생들의 성적을 대표한다고 생각되는 학교를 선택하는 경우

· 할당추출법

- 조사목적과 관련된 연령이나 성별 같은 변수에 따라 모집단을 부분집단으로 구분하고 모집단의 부분집단별 비율과 표본의 부분집단별 비율이 유사하도록 표본 선정
- 적은 비용으로 단기간에 조사 가능

· 눈덩이추출법

- 접근이 어려운 집단에 대해 조사 시 집단 일부 조사 후 집단에 속하는 다른 사람들을 소개 받아서 조사하는 방법 (소개 과정으로 표본이 눈덩이처럼 커짐)
- ex) 폭력조직원들의 약물 사용 실태 조사

학습 유형에 따른 데이터 분석 모델

· 준지도학습

- 정답이 있는 데이터와 정답이 없는 데이터를 동시에 학습에 사용하는 기법
- 비용 절감, 학습 정확도 상승
- ex) GAN, 셀프 트레이닝

· 강화학습

- 주어진 환경에서 보상을 최대화하도록 에이전트를 학습하는 기법
- 에이전트가 행동을 결정하고 환경을 통해 보상이 있으면 이전 입력값과 행동들을 긍정적으로, 벌칙이 있으면 부정적으로 학습하게 됨
- ex) Q-Learning, 정책경사(Policy Gradient)

활성화함수 : 입력 값의 총합 출력 여부와 어떤 값으로 변환하여 출력할지 결정하는 함수

- **Sigmoid 함수** : 0 ~ 1 사이로 확률 표현 (이진분류)

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

- **Tanh 함수** : -1 ~ 1 사이로 확률 표현

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

- **Softmax 함수**
  - 출력값이 N개 (다중 분류)
  - 출력값의 합은 1

$$softmax(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

- **Relu 함수**
  - **입력값이 음수일 경우 출력값과 미분값을 모두 0으로 강제하여 죽은 뉴런 회생에 어려움**
  - **구현이 단순하고 연산 속도가 빠름**

$$ReLU(x) = max(0, x)$$

손실함수

- 지도학습 시 알고리즘이 예측한 값과 실제 정답의 차이를 비교하기 위한 함수
- 학습 중에 얼마나 잘못 예측하는 정도를 확인하기 위한 함수
- 손실이 커질수록 학습이 잘 안되고 있으며, 작아질수록 학습이 잘 되고 있다고 해석 가능

경사하강법 : 손실함수의 크기를 최소화하는 가중치를 찾는 것

AdaGrad : 중요도, 크기가 다른 Feature마다 학습률을 다르게 조절함

참고 사이트 : https://heytech.tistory.com/380

회귀분석

적합도 검정 : 추정된 회귀식이 표본의 실제값을 얼마나 잘 설명하는지 확인하는 방법

변수 영향력 분석 : 독립변수들이 얼마나 영향력을 미치는지 회귀변수의 통계적 유의성 검정, 회귀 계수 추정치에 대한 표준오차와 신뢰구간 검증으로 확인

선형회귀분석의 기본적인 가정

- **선형성** : 독립변수와 종속변수가 선형적이어야 함, 산점도로 확인 가능
- **잔차 정규성** : 잔차의 기댓값은 0이며, 정규분포를 이루어야함
- **잔차 독립성** : 잔차들은 서로 독립적이어야 함
- **잔차 등분산성** : 잔차들의 분산이 일정해야 함
- **다중 공선성** : 3개 이상의 독립변수 간에 상관관계를 띄면 안됨, 다중공선성이 나타나면 설명력이 약해지고 표준오차가 증가함

승산 (odds)

- 사건A가 발생하지 않을 확률 대비 일어날 확률의 비율
- 로지스틱 회귀함수가 1에 가까울수록 발생확률 승산은 올라감

의사결정나무

- 분류나무 (이산형 목표변수)
  - 카이제곱통계량

    (실제도수-기대도수)^2/기대도수의 합
  - 지니 지수

    특정 집합에서 한 항목을 뽑아 무작위로 추정시 틀릴 확률
  - 엔트로피 지수 : 무질서 정도에 대한 측도
- 회귀나무 (연속형 목표변수)
  - 분산분석 F-통계량의 p-value

    등분산성을 검정하여 p-value가 커지면 등분산성이 있음을 뜻하고 순수도가 높아짐

    (즉, 분류가 잘되었음을 뜻함)
  - 분산의 감소량

    분산의 감소량이 최대화가 될수록 낮은 이질성, 순수도가 높아지는 방향으로 가지 분할이 진행됨
- 정지규칙

  더 이상 분리가 일어나지 않고 현재의 마디가 끝마디가 되도록 함
- 장단점
  - 장점 : 연속형, 범주형 변수 모두 가능, 데이터의 규칙 도출에 유용함
  - 단점 : 트리구조가 복잡하면 예측/해석력이 떨어짐, 데이터 변형에 민감함

과대 적합 해결 방안 - 가중치 감소

- 가중치가 클수록 패널티를 부과하여 가중치 매개변수 절대값을 감소시킴

Lasso(라쏘) : L1 규제(정규화)

- L1-norm 개념 (맨하탄 거리)로 비용함수 조정

Ridge(릿지) : L2 규제(정규화)

- L2-norm 개념(유클리드 거리)로 비용함수 조정

엘라스틱넷

- 릿지와 라쏘의 결합
- L1-norm과 L2-norm을 모두 이용하여 패널티를 부과하여 회귀모델 생성