데이터 통계 용어 정리

정형화된 데이터의 요소

- 수치형 : 숫자를 이용해 표현할 수 있는 데이터
- 연속형 : 일정 범위 안에서 어떤 값이든 취할 수 있는 데이터 (구간형, 실수형, 수치형 데이터)
- 이산(Discrete): 횟수와 같은 정수 값만 취할 수 있는 데이터 (유의어 : 정수형, 횟수 데이터)
- 범주형 (Categorical): 가능한 범주 안의 값만을 취하는 데이터 (목록, 열거 요인, 명목, 다항형 데이터)
- 이진 (Binary) : 두 개의 값 (0/1 혹은 참/거짓)만을 갖는 범주형 데이터의 특수한 경우(이항적, 논리형, 지표 , 불리언 데이터)
- 순서형 (Ordinal):값들 사이에 분명한 순위가 있는 범주형 데이토(정렬된 요인 데이터)

테이블 데이터

- 데이터 프레임 : 통계와 머신러닝 모델에서 가장 기본이 되는 테이블 형태의 데이터 구조
- 피처(Feature) : 일반적으로 테이블의 각 열이 하나의 피처를 의미한다 (특징, 속성, 입력 , 예측변수, 변수)
- 결과 : 실험이나 연구에서 결과를 예}측하기 위해 피처를 사용함
- 레코드 일바적으로 테이블의 각 행은 하나의 레코드를 의미함 (기록값, 사건, 사례, 예제, 관측값, 패턴 ,샘플)

위치 추정

- 평균 (mean) : 모든 값의 총합을 개수로 나눈 값
- 가중 평균 (Weighted mean) : 가중치를 곱한 값의 총합을 가중치의 총합으로 나눈 값

- 중간값(Median) : 데이터에서 가장 가운데 위치한 값 (유의어: 50번째 백분위 수)
- 백분위수 (Percentile): 전체 데이터의 P% 를 아래에 두는 값 (유의어 : 분위수)
- 가중 중간값(Weighted median): 데이터를 정렬한 후, 각 가중치 값을 위에서부터 더할 때, 총합의 중간이 위치하는 데이터 값
- 절사평균(Trimmed mean) : 정해진 개수의 극단값 을 제외한 나머지 값들의 평균
- 로버스트 하다 : 극단값들에 민감하지 않다는 것을 의미
- 특잇값(Outlier) 대부분의 값과 매우 다른 데이터 값

변이 추정

- 편차 (Deviation): 관측값과 위치 추정값 사이의 차이 (유의어 : 오차, 잔차)
- 분산 (Variance) : 평균과의 편차를 제곱한 값들의 합을 n-1로 나눈 값. n은 데이터 개수 (유의어 : 평균 제곱오차)
- 표준편차 (Standard Deviation) : 분산의 제곱근
- 평균절대편차 (Mean absolute Deviation) 평균과의 편차의 절댓값의 평균
- 중간값의 중위절대편차 (MAD) : 중간값과의 편차의 절댓값의 중간값
- 범위 : 데이터의 최댓값과 최솟값의 차이
- 순서통계량 : 최소에서 최대까지 정렬된 데이터 값에 따른 계량형 (유의어 : 순위)
- 백분위수 (Percentile): 어떤 값들의 P퍼센트가 이 값 혹은 더 작은 값을 갖고, (100-P). 퍼센트가 이 값 혹은 더 큰 값을 갖도록 하는 값 (유의어 : 분위수)
- 사분위범위(IQR) : 75번째 백분위수와 25번째 백분위수 사이의 차이

데이터 분포 탐색

- 상자 그림 (Boxplot) : 투키가 데이터의 분포를 시각화하기 위한 간단한 방법으로 소개한 그림 (상자수염도)
- 도수분포표 : 어떤 구간 (bin)에 해당하는 수치 데이터 값들의 빈도를 나타내는 기록

- 히스토그램 : x축은 구간들을, y축은 빈도수를 나타내는 도수 테이블의 그림, 시각적으로 비슷하지만 막대 그래프를 히스토그램과 혼동해서는 안됨
- 밀도 그림 : 히스토그램을 부드러운 곡선으로 나타낸 그림 , 커널 밀도 추정을 주로 사용함

이진 데이터와 범주 데이터 탐색

- 최빈값 (mode) : 데이터에서 가장 자주 등장하는 범주 혹은 값
- 기댓값(Expected value): 범주에 해당하는 어떤 수치가 있을 때, 범주의 출현 확률에 따른 평균
- 막대도표 : 각 범주의 빈도수 혹은 비율을 막대로 나타낸 그림
- 파이그림: 각 범주의 빈도수 혹은 비율을 원의 부채꼴 모양으로 나타낸 그림

상관관계

- 상관계수 : 수치적 변수들 간에 어떤 관계가 있는지를 나타내기 위해 사용되는 측정량 (-1 ~ 1 까지 범위)
- 산관행렬 : 행과 열이 변수들을 의미하는 표를 말하며, 각 셀은 그 행과 열에 해당하는 변수들 간의 상관관계를 의미한다.
- 산점도 x 축과 y 축이 서로 다른 두개의 변수를 나타내는 도표

두 개 이상의 변수 탐색하기

- 분할표 : 두 가지 이상의 범주형 변수의 빈도수를 기록한 표
- 육각형 구간 : 두 변수를 육각형 모양의 구간으로 나눈 그림
- 등고 도표 : 지도상에 같은 높이의 지점을 등고선으로 나타내는 것처럼, 두 변수의 밀도를 등고선으로 표시한 도표
- 바이올린 도표 : 상자그림과 비슷하지만 밀도추정을 함께 보여줌

임의표본추출과 표본편향

- 표본 : 더 큰 데이터 집합으로부터 얻은 부분집합
- 모집단 : 어떤 데이터 집합을 구성하는 전체 대상 혹은 전체 집합
- N(n) : 모집단(표본)의 크기
- 임의표본추출(임의표집, 랜덤표본추출) : 무작위로 표본 추출하는 것
- 층화표본추출(층화표집) : 모집단을 층으로 나눈 뒤, 각 층에서 무작위로 표본을 추출하는 것
- 계층 : 공통된 특징을 가진 모집단의 동종 하위 그룹
- 단순임의표본(단순랜덤표본) : 모집단 층화 없이 임의표본추출 얻은 표본
- 편향 : 계통상의 오류
- 표본편향 : 모집단을 잘못 대표하는 표본

선택편향

- 선택편향 : 관측 데이터를 선택하는 방식 때문에 생기는 편향
- 데이터 스누핑 : 뭔가 흥미로운 것을 찾아 광범위하게 데이터를 살피는 것
- 방대한 검색 효과 : 중복 데이터 모델링이나 너무 많은 예측변수를 고려하는 모델리에서 비롯되는 편향 혹은 비재현성

통계학에서 표본분표

- 표본통계량 : 더 큰 모집단에서 추출된 표본 데이터들로부터 얻은 측정 지표
- 데이터 분포 : 어떤 데이터 집합에서의 각 개별 값의 도수분포
- 표본분포 : 여러 표본들 혹은 재표본들로부터 얻은 표본통계량의 도수 분포
- 중심극한 정리 : 표본크기가 커질수록 표본분포가 정규분포를 따르는 경향
- 표준오차 : 여러 표본들로부터 얻은 표본통계량의 변량 (개별 데이터 값들의 변량을 뜻하는 표준편차와 혼동 금지)

부트스트랩

- 부트스트랩 표본 : 관측 데이터 집합으로부터 얻은 복원추출 표본

- 재표본추출(재표집,리샘플링) : 관측 데이터로부터 반복해서 표본추출하는 과정, 부트스트랩과 순열(셔플링)과정을 포함한다.

신뢰구간

- 신뢰수준 : 같은 모집단으로부터 같은 방식으로 얻은, 관심 통계량을 포함할 것으로 예상되는 신뢰구간의 백분율
- 구간끝점 : 신뢰구간의 최상위, 최상위 끝점

정규분포

- 오차 : 데이터 포인트와 예측값 혹은 평균 사이의 차이
- 표준화(정규화)하다 : 평균을 빼고 표준편차로 나눈다.
- z 점수 : 개별 데이터 포인트를 정규화한 결과
- 표준정규분포 : 평균=0 , 표준편차=1인 정규분포
- QQ 그림 : 표본분포가 특정분포에 얼마나 가까운지를 보여주는 그림

긴 꼬리 분포

- 꼬리 : 적은 수의 극단값이 주로 존재하는 , 도수분포의 길고 좁은 분포
- 왜도 : 분포의 한쪽 꼬리가 반대쪽 다른 꼬리보다 긴 정도

스튜던트 t 분포

- n : 표본크기
- 자유도 : 다른 표본크기, 통계량, 그룹의 수에 따라 t 분포를 조절하는 변수

이항분포

- 시행 : 독립된 결과를 가져오는 하나의 사건 (예: 동전 던지기)
- 성공 : 시행에 대한 관심의 결과 (유의어 :1)

- 이항식 : 두가지 결과를 가짐
- 이항시행 : 두 가지 결과를 가져오는 시행
- 이항분포 : n 번 시행에서 성공한 횟수에 대한 분포 (유의어 : 베르누이 분포)

푸아송 분포와 그 외 관련 분포들

- 람다 : 단위 시간이나 단위 면적당 사건이 발생하는 비율
- 푸아송 분포 : 표집된 단위 시간 혹은 단위 공간에서 발생한 사건의 도수분포
- 지수분포 : 한 사건에서 그다음 사건까지의 시간이나 거리에 대한 도수분포
- 베이불 분포 : 사건 발생률이 시간에 따라 변화하는, 지수분포의 일반화된 버전

A/B 검정

- 처리: 어떤 대상에 주어지는 특별한 환경이나 조건
- 처리군(처리 그룹) : 특정 처리에 노출된 대상들의 집단
- 대조군(대조 그룹) : 어떤 처리도 하지 않은 대상들의 집단
- 임의화(랜덤화): 처리를 적용할 대상을 임의로 결정하는 과정
- 대상: 처리를 적용할 개체 대상 (유의어: 피실험자)
- 검정통계량 : 처리 효과를 측정하기 위한 지표

가설검정

- 귀무가설 : 우연 때문이라는 가설 (유의어 : 영가설)
- 대립가설 : 귀무가설과의 대조 (증명하고자 하는 가설)
- 일원 검정 : 한 방향으로만 우연히 일어날 확률을 계산하는 가설검정
- 이원검정 : 양방향으로 우연히 일어날 확률을 계산하는 가설검정

재표본추출

- 순열검정 : 두 개 이상의 표본을 함께 결합하여 관측값들을 무작위로 재표본으로 추출하는 과정 (임의화 검정 , 임의순열검정, 정확검정)
- 재표본추출 : 관측 데이터로부터 반복해서 표본추출하는 과정
- 복원/비복원 : 표본을 추출할 때, 이미 한번 뽑은 데이터를 다음번 추출을 위해 다시 제자리에 돌려놓거나 / 다음 추출에서 제외하는 표본추출 방법

통계적 유의성과 p값

- p 값 : 귀무가설을 구체화한 기회 모델이 주어졌을 때 관측된 결과와 같이 특이하거나 극단적인 결과를 얻을 확률
- 알파 : 실제 결과가 통계적으로 의미 있는것으로 간주되기 위해, 우연에 의한 결과가 능가해야 하는 “비정상적인” 가능성의 임계 확률
- 제1종 오류 : 우연에 의한 효과를 실제 효과라고 잘못 결론 내리는 것
- 제2종 오류 : 실제 효과를 우연에 의한 효과라고 잘못 결론 내리는 것

다중검정

- 제 1종 오류 : 어떤 혀과가 통계적으로 유의미하다고 잘못 결론을 내린다.
- 거짓 발견 비율(FDR) : 다중검정에서 1종 오류가 발생하는 비율
- 알파 인플레이션 : 1종 오류를 만들 확률인 알파가 더 많은 테스트를 수행할수록 증가하는 다중검정 현상
- p 값 조정 : 동일한 데이터에 대해 다중검정을 수행하는 경우에 필요
- 과대적합(오버피팅) : 잡음까지 피팅

자유도

- 표본크기 n : 해당 데이터에서 관측값의 개수 (행 혹은 기록값의 개수와 같은 의미
- d.f : 자유도

분산분석 (ANOVA)

- 쌍별 비교 : 여러 그룹 중 두 그룹 간의 (예를 들면 평균에 대한) 가설검정

- 총괄검정 : 여러 그룹 평균들의 전체 분산에 관한 단일 가설검정
- 분산분해 : 구성 요소 분리. 예를 들면 전체 평균, 처리 평균 , 잔차 오차로부터 겨별값들에 대한 기여를 뜻함
- F 통계량 : 그룹 평균 간의 차이가 랜덤 모델에서 예상되는 것에서 벗어나는 정도를 측정하는 표준화된 통계량
- SS : 어떤 평균으로부터의 편차들의 제곱합

카이제곱검정

- 카이제곱통계량 : 기댓값으로부터 어떤 관찰값까지의 거리를 나타내는 측정치
- 기댓값 : 어떤 가정(보통 귀무가설)으로부터 데이터가 발생할 때, 그에 대해 기대하는 정도
- d.f : 자유도

멀티암드 뱃딧 알고리즘

- 멀티암드 밴딧(MAB) : 고객이 선택할 수 있는 손잡이가 여러 개인 가상의 슬롯머신을 말하며, 각 손잡이는 각기 다른 수익을 가져다준다. 다중 처리 실험에 대한 비유
- 손잡이 : 실험에서 어떤 하나의 처리를 말함 (예를 들면 웹테스트에서 헤드라인 A)
- 상금(수익) : 슬롯머신으로 딴 상금에 대한 실험적 비유 (예를 들면” 고객들의 링크 클릭수”)

검정력과 표본크기

- 효과크기 : 클릭률의 20% 향상과 같이 통계 검정을 판달할 수 있는 효과의 최소 크기
- 검정력 : 주어진 표본크기로 주어진 효과크기를 알아낼 확률
- 유의수준 : 검증 시 사용할 통계 유의수준

단순선형회귀

- 응답변수(반응변수) : 예측하고자 하는 변수(유의어 : 종속변수 , 변수 Y , 목표 , 출력)
- 독립변수 : 응답치를 예측하기 위해 사용되는 변수 (유의어 : 예측변수 , 변수 X , 피처 , 속성)
- 레코드 : 한 특정 경우에 대한 입력과 출력을 담고 있는 벡터 (유의어 : 행, 사건, 예시 , 예제)
- 절편 : 회귀직선의 절편 (즉 , X = 0 일 때 예측값)
- 회귀계수 : 회귀직선 기울기
- 적합값 : 회귀선으로부터 얻은 추정치
- 잔차 : 관측값과 적합값의 차이 (유의어 : 오차)
- 최소제곱 : 잔차의 제곱합을 최소화하여 회귀를 피팅하는 방법 (유의어 : 보통최소제곱 , OLS)

회귀를 이용한 예측

- 예측구간 : 개별 예측값 주위의 불확실한 구간
- 외삽법 : 모델링에 사용된 데이터 범위를 벗어난 부분까지 모델링을 확장하는 것

회귀방정식 해석

- 변수 간 상관 : 변수들이 같은 방향으로 움직이려는 경향을 가짐

(예를 들어 한 변수가 올라갈 때 다른 변수도 올라가고 그 반대 경우에도 동일함, 부정적인 상관관계일 때는 한 변수가 올라갈 대 다른 변수는 반대로 내려감) . 예측변수끼리 서로 높은 상관성을 가질 때는 개별 계수를 해석하는 것이 어려움

- 다중공산성 : 예측변수들이 완벽하거나 거의 완벽에 가까운 상관성을 갖는다고 할 때, 회귀는 불안정하며 계산이 불가능함
- 교란변수 : 중요한 예측변수이지만 회귀방정식에 누락되어 결과를 잘못되게 이끄는 변수
- 주효과 : 다른 변수들과 독립된, 하나의 예측변수와 결과변수 사이의 관계
- 상호작용 : 둘 이상의 예측변수와 응답변수 사이의 상호 의존적인 관계

다항회귀와 스플라인 회귀

- 다항회귀 : 회귀모형에 다항식(제곱, 세제곱등) 항을 추가한 방식
- 스플라인 회귀 : 다항 구간들을 부드러운 곡선 형태로 피팅한다.
- 매듭 : 스플라인 구간을 구분하는 값들
- 일반화가법모형(GAM) : 자동으로 구간을 결정하는 스플라인 모델
-

나이브 베이즈

- 나이브 베이즈 알고리즘 : 주어진 결과에 대해 예측변숫값을 관찰할 확률을 사용하여 예측변수가 주어졌을대 , 결과 Y = i 를 관찰할 확률 , 즉 정말 관심 있는 것을 추정
- 조건부 확률 : 어떤 사건 (Y=i)이 주어졌을 때, 해당 사건(X= i)을 관찰할 확률 P(Xi | Yi)
- 사후확률 : 예측 정보를 통합한 후 결과의 확률 (이와 달리, 사전확률에서는 예측변수에 대한 정보를 고려하지 않음)

판별분석

- 공분산 : 하나의 변수가 다른 변수와 함께 변화하는 정도(유사한 크기와 방향)를 측정하는 지표
- 판별함수 : 예측변수에 적용했을 때, 클래스 구분을 최대화 하는 함수
- 판별 가중치 : 판별함수를 적용하여 얻은 점수를 말하며, 어떤 클래스에 속할 확률을 추정하는 데 사용된다.

로지스틱 회귀

- 로짓 : (0~1이 아니라) +- 무한의 범위에서 어떤 클래스에 속할 확률을 결정하는 함수 (유의어: 로그오즈)
- 오즈 : “실패”(0)에 대한 “성공”(1)의 비율
- 로그 오즈 : 변환 모델(선형)의 응답변수, 이 값을 통해 확률을 구한다.

불균형 데이터 다루기

- 과소표본 : 분류 모델에서 개수가 많은 클래스 데이터 중 일부 소수만을 사용하는 것
- 과잉표본: 분류 모델에서 회귀 클래스 데이터를 중복하여, 필요하면 부트스트랩해서 사용하는것(유의어: 업샘플)
- 상향 가중치 or 하향 가중치 : 모델에서 회귀(혹은 다수) 클래스에 높은(혹은 낮은) 가중치를 주는것
- 데이터 생성 : 부트스트랩과 비슷하게 다시 샘플링한 레코드를 빼고 원래 원본과 살짝 다르게 데이터를 생성하는 것
- z 점수 : 표준화 결과
- k : 최근접 이웃 알고리즘에서 이웃들의 개수

k- 최근접 이웃

- 이웃 : 예측변수에서 값들이 유사한 레코드
- 거리 지표 : 각 레코드 사이가 얼마나 멀리 떨어져 있는지를 나타내는 단일 값
- 표준화 : 평균을 뺀 후에 표준편차로 나누는 일 (유의어: 정규화)
- z 점수 : 표준화를 통해 얻은 값
- k : 최근접 이웃을 계산하는 데 사용되는 이웃의 개수

트리모델

- 재귀분활 :마지막 분할 영역에 해당하는 출력이 최대한 비슷한 결과를 보이도록 데이터를 반복적으로 분할하는 것
- 분할값 : 분할값을 기준으로 예측변수를 그 값보다 작은 영역과 큰 영역으로 나눈다 .
- 마디(노드) : 의사 결정 트리와 같은 가지치기 형태로 구성된 규칙들의 집합에서, 노드는 분할 규칙의 시각적인 표시라고 할 수 있다.
- 잎 : if-then 규칙의 마지막 부분, 혹은 트리의 마지막 가지 부분을 의미한다. 트리 모델에서 잎 노드는 어떤 레코드에 적용할 최종적인 분류 규칙을 의미한다.

- 손실 : 분류하는 과정에서 발생하는 오분류의 수, 손실이 클수록 불순도가 높다고 할 수 있다.
- 불순도 : 데이터를 분할한 집합에서 서로 다른 클래스의 데이터가 얼마나 섞여 있는지를 나타낸다. 더 많이 섞여있을수록 불순도가 높다고 할 수 있다 .
- 가지치기 : 학습이 끝난 트리 모델에서 오버피팅을 줄이기 위해 가지들을 하나씩 잘라내는 과정

주성분분석

- 주성분 : 예측변수들의 선형결합
- 부하: 예측변수들을 성분으로 변형할 때 사용되는 가중치
- 스크리그래프 : 성분들의 변동을 표시한 그림, 설명된 분산 혹은 분산 혹은 설명된 분산의 비율을 이용하여 성분들의 상대적인 중요도를 보여줌

k-평균 클러스트링

- 클러스터(군집) : 서로 유사한 레코드들의 집합
- 클러스터 평균 : 한 클러스터 안에 속한 레코드들의 평균 벡터 변수
- k : 클러스터의 개수

스케일링과 범주형 변수

- 스케일링 : 데이터의 범위를 늘리거나 줄이는 방식으로 여러 변수들이 같은 스케일에 오도록 하는것
- 정규화 : 원래 변수 값에서 평균을 뺀 후에 표준펴차로 나누는 방법으로, 스케일링의 일종
- 고워 거리 : 수치형과 범주형 데이터가 섞여 있는 경우에 모든 변수가 0~1 사이로 오도록 하는 스케일링 방법