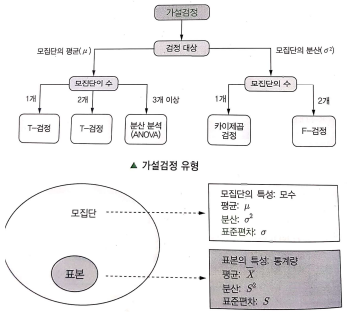
Ⅳ 빅데이터 결과 해석

CH.1 분석모형 평가 및 개선

1. 분석모형평가
    1. 평가지표
        1. 회귀모형 평가지표 [SSE,SST,SSR,AE,MAE,RMSE,MAPE,MPE]
           [R^2, Radj, Cp]
        2. 분류모형 평가지표 [혼동행렬(TP,TN/FP,FN) - 정확도, 오차비율, 민감도, 특이도, 거짓긍정률, 정밀도, F-Measure, 카파통계량]
           [ROC 곡선, AUC 0.5~1, 1에가까울수록 좋은거]
           [이익도표 (Gain Chart, Lift Curve)]
    2. 분석모형진단
        1. 데이터 분석 모형의 오류 [일반화 오류, 학습 오류]
        2. 데이터 분석 모형 검증 [홀드아웃 교차검증, 다중 교차 검증]
        3. 분석 모형 시각화 [구조화 > 시각화 > 시각표현]
        4. 분석모형 진단 [선형성(단순회귀), 독립성, 등분산성, 정상성]
    3. 교차검증
        1. 교차검증 개념
        2. 교차검증 종류
            1. 홀드아웃 교차검증 [학습데이터(분류기), 검증데이터(매개변수), 평가데이터]
            2. 랜덤 서브샘플링 [홀드아웃반복]
            3. K-Fold Cross Validation [K개로 나누고, 1/k는 평가데이터, 나머지는 학습데이터로선정, k번반복]
            4. Leave-One-Out Cross Validation (LOOCV) [k-fold랑 같은데, 1개 평가데이터 n번반복]
            5. Leave-p-Out Cross Validation (LpOCV) [똑같은데 p개 샘플을 평가데이터]
            6. Repeated Learning-Testing (RLT) [랜덤, 비복원추출]
            7. 부트스트랩 [랜덤, 복원추출]
        3. 주요교차검증간의 비교
    4. 모수 유의성 검정
        1. 모집단과 모수 관계
        2. 모집단의 평균에 대한 유의성 검정
            1. Z-검정 귀무가설,정규분포로 통계검정
            2. T-검정 T-분포, 정규성, 등분산성, 독립성
            3. 분산검정(ANOVA) F-분포 이용, 일원/이원
        3. 모집단의 분산에 대한 유의성 검정
            1. 카이제곱검정 $x^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$
            2. F-검정 $F = \frac{s_1^2}{s_2^2}$ 분산간의 비율에대한 검정



    5. 적합도검정
        1. 적합도검정 개념
        2. 적합도검정 기법 유형
            1. 카이제곱검정 [p-값이 0.05보다 크면 가정된 확률을 따른다.]
            2. 정규성검정
                1. 샤피로-윌크검정 [보통 데이터적을 때]
                2. 콜모고로프-스키르노프 적합성 검정 (K-S검정) [보통2000개이상]
                3. Q-Q Plot (Quantile-Quantile Plot) [시각화를 통한, 보조용]

2. 분석모형개선
    1. 과대적합방지
        1. 과대적합개념 (Over-fitting) [모델파라미터가 많거나, 학습용 데이터세트가 부족한 경우]
        2. 과대적합방지하기
            1. 데이터증강 [양이적으면 노이즈까지 분석되니까 양을늘림]
            2. 모델의 복잡도 감소 [인공신경망복잡도는 은닉층수나 모델수용력등으로 결정되니 낮추면됨]
            3. 가중치 규제 적용 [가중치규제, 규제강도정하는 하이퍼파라미터]
            4. 드롭아웃 [일부 신경망을 쓰니 않음, 앙상블하는 것 같은 효과]
    2. 매개변수 최적화
        1. 매개변수의 개념 (Parameter)
        2. 매개변수 최적화의 개념 (Parameter Optimization) [차이는 손실함수로 표현, 이 값을 최소화 하는 매개변수를 찾는 것]
        3. 매개변수 종류 [y= ax + b (a가 가중치, b가 편향)]
        4. 매개변수 최적화 과정
        5. 매개변수 최적화 기법
            1. 확률적 경사 하강법 SGD (Stochastic Gradient Descent) [함수따라서 최소가 되는 변곡점 찾는거]
            2. 모멘텀 Momentum [1번에 속도개념 적용. 미적분으로 구하는 듯?]
            3. AdaGrad [처음엔 크게, 가까워질수록 작게, 효율적으로]
            4. Adam [2번, 3번 장점 합친거]
    3. 분석모형 융합
        1. 취합 방법론 (Aggregation)
            1. 다수결 Voting [직접투표, 간접투표]
            2. 배깅 Bagging [중복허용, 세트나누는 기법으로 복원추출]
            3. 페이스팅 Pasting [비복원추출]
            4. 랜덤 서브스페이스 Random Subspaces [학습데이터 다 사용하고 특성만 샘플링]
            5. 랜덤 패치 Random Patches [학습데이터, 특성 모두 샘플링]
            6. 랜덤 포레스트 Random Forests [Decision Tree 개별모형 결합, 랜덤으로 뽑고 변수선택]
        2. 부스팅 방법론
            1. 에이다 부스트 AdaBoost [약한모형을 순차적으로 적용, 오샘플 가중치 높임]
            2. 그레디언트 부스트 Gradient Boost
    4. 최종모형 선정
        1. 최종모형평가기준 선정
        2. 최종모형 분석 결과검토
        3. 알고리즘별 결과비교

CH. 2 분석결과 해석 및 활용

1. 분석결과 해석
    1. 분석모형 해석
        1. 데이터 시각화의 개념 (Data Visualization)
        2. 데이터 시각화의 기능 [설명, 탐색, 표현]
        3. 데이터 시각화의 목적
        4. 데이터 시각화의 유형
        5. 빅데이터 시각화 도구
            1. 태블로 [클라우드기반]
            2. 인포그램
            3. 차트블록 [웹기반]
            4. 데이터래퍼
        6. 데이터 시각화 절차 (구조화 - 시각화 - 시각표현)
        7. 시각화 분석을 위한 데이터 유형 [범주,비율 / 추세,패턴 / 관계,연결]
    2. 비즈니스 기여도 평가
        1. 비즈니스 기여도 평가의 개념
        2. 비즈니스 기여도 평가지표 [TCO, ROI, NPV, IRR, PP]
        3. 비즈니스 기여도 평가 고려사항
2. 분석결과 시각화
    1. 시공간 시각화
        1. 시간 시각화 [선 Line Graph, 영역 Area Chart, 계단식 Step Line Graph]
        2. 공간 시각화 [등치지역도, 등치선도, 도트맵, 버블플롯맵, 카토그램]
    2. 관계시각화 [산점도, 산점도 행렬, 버블차트, 히스토그램]
    3. 비교시각화 [플로팅 바 차트, 히트맵, 체르노프 페이스, 스타차트, 평행좌표 그래프]
    4. 인포그래픽 [지도형, 도표형, 스토리텔링형, 타임라인형, 비교분석형, 만화형]

3. 분석결과 활용
    1. 분석모형 전개
        1. 빅데이터 모형 운영 시스템 적용방안
            [분석목적정의→가설검토→데이터 준비 및 처리→ 모델링 및 분석 → 정확도 및 성능평가 → 운영]
            [→분석모형적용 모듈 적용, 분석모형 통합결정 및 구현]
        2. 빅데이터 모형의 운영 및 개선방안 수립 [예측오차를 가지고 추적신호를 계산하며 개선]
            [TS(Tracking Sigmal):예측 오차의 합을 평균 절대 편차로 나눈 값]
    2. 분석결과 활용 시나리오 개발
        1. 분석 결과에 따른 활용분야 분류
        2. 분류 결과를 토대로 적용 가능한 서비스 영역 도출
        3. 분류 결과를 토대로 적합한 신규 서비스 모형 도출[서브퀄모형: 반응성,공감성,확신성,유형성,신뢰성]
        4. 서비스 모형에 따른 활용 방안 제시
    3. 분석 모형 모니터링
        1. 분석 모형 모니터링 개념 [배치스케줄러가 정상적으로 실행되는지, 성과가 예상수준인지 모니터링]
        2. 분석 모형 모니터링 솔루션 [샤이니; ui.R / server.R]
        3. 분석 모형 성능 모니터링 [응답시간, 사용률, 가용성, 정확성]
        4. 분석 모형 모니터링 고려사항 [시뮬레이션, 최적화]
    4. 분석 모형 리모델링
        1. 분석 모형 리모델링 개념 [데이터마이닝:분기별, 시뮬레이션:반기(주요변경시), 최적화: 1년]
        2. 분석모형 리모델링 절차