**Data Mining Homework #1**

4조 (김상규, 손민우, 최희영)

**1. 상관관계이지만 인과관계가 아닌 예제를 만드시오.**

- 우리 동네 소아과를 방문한 아이 수와 멀리 떨어진 동네 소아과를 방문한 아이 수는 상관관계가 높으나 인과관계가 있는 것이 아니다.

감기, 수족구 등과 같은 질병이 유행하기 때문에 우리 동네와 멀리 떨어진 동네 소아과를 방문하는 아이 수가 많아 지는 것이다.

여름철 아이스크림 판매량과 전력량

**2. 교략효과(Confounding effect)에 대해 설명하시오.**

-

**3. 데이터마이닝의 5단계에 대해서 설명하시오.**

Sampling 🡪 Exploration 🡪 Modification 🡪 Modeling 🡪 Assessment

1) Sampling

적절한 양의 표본을 원자료로부터 추출하는 단계

시간과 비용의 절약과 함께 효율적인 모형의 구축에 필수적이다.

2) Exploration

여러 가지 자료의 탐색을 통해 기초통계자료, 평균, 분산 등의 기본적인 정보를 획득하는 단계

3) Modification

데이터의 효율적인 사용을 위한 변수의 변환, 수량화, 그룹화 등을 통하여 데이터를 변환하는 단계

4) Modeling

데이터 마이닝의 핵심 단계로 분석 목적에 따라 적절한 기법을 사용하여 예측 모형을 만드는 단계

5) Assessment

모형화의 결과에 대해 신뢰성, 유용성 등을 평가하는 단계

**4. 입력변수가 X = (X1,...,Xp) ∈Rp이고 출력변수가 Y ∈R이라고 할 때, 회귀모형의 정의를 서술하시오.**

**5. 베이즈위험의 정의를 서술하시오.**

- 베이즈 규칙이 갖는 오분류율

**6. 베이즈 규칙의 정의를 서술하시오.**

-

7. a. X ∈ Rp, Y ∈ {0,1}이고 결합확률분포 P을 따르는 확률변수들이라고 할 때, 주어진 분류함수 C : Rp → {0,1}에 대해 모위험을 R(C) = E(X,Y )I(C(X) 6= Y ) 라 하자. R(C)을 최소로하는 분류모형은 I(P(Y|X) > 0.5)임을 보이시오.

b. 주어진 손실함수 l : {0,1}→R+에 대해 모위험을 Rl(C) = E(X,Y )[l(Y )I(C(X) 6= Y )] 로 정의하자. Rl(C)을 최소화하는 분류모형을 구하시오.

**8. 층화추출법을 사용하는 예를 드시오.**

기업의 연평균 매출액을 조사하는 경우 단순랜덤추출시 기업의 대부분을 차지하는 소규모 기업이 추출될 것이므로 매출액 추정이 실제보다 낮게 평가될 수 있다.

따라서 기업의 규모별로 대기업/중견기업/중소기업 등의 층으로 나누어

각 층에서 표본을 랜덤하게 추출하여 매출액을 조사해야 한다.

**9. 사후추출법을 사용하는 예를 드시오.**

가족의 흡연이 암에 미치는 영향에 대한 연구

암의 발생확률이 매우 낮아 단순랜덤추출시 대부분의 표본이 암에 걸리지 않을 확률이 높다.

따라서 암환자와 건강한 사람을 동일하게 추출하여 가족에 의한 간접 흡연 유무를 조사해야 한다.

**11. 모형의 복잡도와 학습에러와의 관계에 대해 설명하시오.**

일반적으로 낮은 복잡도를 갖는 단순 모형은 학습 에러가 큰 반면, 높은 복잡도를 갖는 복잡 모형은 학습 에러가 작다. 하지만 학습 에러가 낮다고 해서 예측 에러가 항상 작은 것은 아니다.

**12. 모형의 복잡도와 예측에러와의 관계에 대해 설명하시오.**

일반적으로 복잡한 모형을 사용하여 학습 에러를 줄이고 예측 에러를 줄일 수 있다. 하지만 학습 에러가 낮다고 해서 항상 예측 에러가 낮은 것은 아니다. 데이터에 따라 모형의 복잡도를 조절하여 예측 에러를 줄일 수 있다.

**13. 과적합이 생기는 이유를 추정에러와 모형에러로 설명하시오.**

주어진 data에 대해 복잡한 모형을 적용하면 모형 에러를 매우 낮은 수준으로 낮출 수 있다. 하지만 이런 경우는 주어진 data에만 잘 맞도록 적합이 된 것으로 새로운 데이터를 추정할 때는 모형 에러보다 훨씬 높은 추정 에러를 발생시킬 수 있다. 예를 들어 y = x을 일정 오차를 가지고 분포된 점들이 있을 때 y = x라는 모형을 적용하는 것보다 높은 차수를 갖는 polynomial fitting을 하면 모형 에러를 낮출 수 있으나, 새로운 데이터에 대해서는 매우 높은 추정 에러를 발생시킨다.

**14. 교차확인 방법에 대해서 설명하시오**

일반적으로 데이터가 많지 않을 때 overfitting을 피하기 위해 많이 사용되는 방법으로 데이터를 k개로 분할하여 k-1개의 분할된 데이터로부터 모형을 만들고 나머지 1개 분할된 데이터를 가지고 모형을 에러를 구한다. 이런 방법으로 1부터 k번째 분할 데이터의 모형 에러들을 통해 모형을 평가한다.

**15. 최소제곱추정량의 정의를 쓰시오.**

-

**16. 회귀모형에서 결정계수 R2의 정의를 쓰고, 의미를 설명하시오.**

**17. 최소제곱추정량의 기하학적 의미를 설명하시오.**

**18. 선형회귀모형에 사용된 가정들을 설명하시오.**

선형성, 등분산성, 정규성, 독립성