**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.

-

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.

-

**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.

13. 과적합이 생기는 이유를 추정에러와 모형에러로 설명하시오.

-

14. 교차확인 방법에 대해서 설명하시오

- 데이터가 많지 않을 때 사용하는 방법으로

15. 최소제곱추정량의 정의를 쓰시오.

-

18. 선형회귀모형에 사용된 가정들을 설명하시오.

- 선형성, 등분산성, 정규성