비즈니스 애널리틱스 과정 과제-1: 세그멘테이션과 타겟마케팅

1. 3차원에서 주어진 세개의 관측치에 대해 서로 간의 거리가 주어졌다. 다차원척도법을 적용할 때 1차원 공간상으로 표현할 수 있는 예제와 없는 예제를 만드시오.

2. $x, y \in \mathbb{R}^p$에 대해,

   a. $d(x,y) \geq 0$이고 $d(x,y) = 0 \Leftrightarrow x = y$

   b. $d(x,y) = d(y,x)$

   c. 어떤 $z \in \mathbb{R}^p$에 대해 $d(x,y) \leq d(x,z) + d(z,y)$

   을 만족하면 $d(x,y)$은 두 점 x와 y의 거리(metric)라고 정의한다. 이 때 유클리드 거리 $d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$와 범주형 자료에서 불일치 항목의 수가 거리(metric)임을 증명하시오.

3. 계층적 군집분석에서 최단연결법(single linkage), 최장연결법(complete linkage)의 정의를 서술하고, 두 연결법의 성질을 비교하시오.

4. K-평균 알고리즘을 설명하고 K-평균 알고리즘과 계층적 군집분석의 계산량을 비교하여 데이터의 크기가 큰 경우 어떤 방법이 더 적합한지 서술하시오.

5. (Open probelm) 범주형자료 사이에서 사용할 수 있는 거리(metric) 3개를 서술하시오.

6. (Open problem) n개의 관측치 간의 거리만 주어진 경우 이 n개의 관측치의 평균을 정의하고 구하는 방법을 설명하여라.

7. 입력변수 $X \in \mathbb{R}^p$와 출력변수 $Y \in \mathbb{R}$라 하면, 회귀분석 문제에서 모위험(population risk) $R(f) = \mathbb{E}_{(X,Y)}|Y - f(X)|$을 최소화하는 추정량은 $\text{median}(Y|X)$임을 보이시오.

8. Y을 자료가 속하는 클래스를 나타내는 변수로 0 또는 1의 값을 갖는 이진분류의 문제를 고려하자. Y의 사전확률은 $\mathbb{P}(Y=0)$와 $\mathbb{P}(Y=1)$으로 주어지고, $\mathbb{P}(Y=0)=1-\mathbb{P}(Y=1)$을 만족한다. $Y=0$인 클래스에서 n_0개의 자료를, $Y=1$인 클래스에서 n_1개의 자료를 각각 임의로 추출하여 얻은 표본으로부터 구한 사후확률을 $\mathbb{P}^o(Y=1|X)$라 하자. 이 때, 모집단에서의 베이즈 분류기 C^*을 다음과 같이 주어진다.

$$C^* = I\Big(\mathbb{P}^o(Y=1|X) > \frac{n_1\mathbb{P}(Y=0)}{n_0\mathbb{P}(Y=1)+n_1\mathbb{P}(Y=0)}\Big)$$

이를 증명하시오.

9. (X,Y)은 $\{0,1\}^2$의 값을 가지고 확률 P을 따르는 확률벡터이다.

   a. 주어진 확률 P에 대한 오즈비를 정의하시오.

   b. 주어진 $\pi \in (0,1)$에 대해 확률분포 Q를 다음과 같이 정의하자:

$$Q(X,Y) = P(X|Y)[(1-\pi)I(Y=0)+\pi I(Y=1)].$$

   모든 $\pi \in (0,1)$에 대해서 P와 Q의 오즈비가 같음을 증명하시오.

   c. $P(Y|X)=\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$라 하자. $\exp(\beta_1)$이 P의 오즈비가 됨을 증명하시오.

   d. $P(Y|X)=\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}$인 경우 $Q(Y|X)=\frac{\exp(\beta_0^*+\beta_1 X)}{1+\exp(\beta_0^*+\beta_1 X)}$으로 주어짐을 보이고 β_0^*을 구하시오.

10. $X \in \mathbb{R}^p$, $Y \in \{0,1\}$이고 결합확률분포 P을 따르는 확률변수들이라고 할 때, 주어진 분류모형 $C : \mathbb{R}^p \to \{0,1\}$에 대해 민감도와 특이도는 다음과 같이 정의된다:

$$\begin{aligned} \text{민감도}(P) &= P(C(X)=1, Y=1)/P(Y=1) \\ \text{특이도}(P) &= P(C(X)=0, Y=0)/P(Y=0). \end{aligned} \tag{1}$$

주어진 $\pi \in (0,1)$에 대해 확률분포 Q을 다음과 같이 정의하자:

$$Q(X,Y) = P(X|Y)[(1-\pi)I(Y=0)+\pi I(Y=1)].$$

이 때, 민감도(P) = 민감도(Q), 특이도(P) = 특이도(Q)임을 보이시오.

11. ROC 커브를 설명하시오.

12. Lasso 추정량의 정의를 서술하고, lasso 추정량의 성긴 성질 (sparsity)에 대해서 설명하시오.

13. AIC와 BIC에 대해서 설명하시오.

14. a. $X \in \mathbb{R}^p$, $Y \in \{0,1\}$이고 결합확률분포 P을 따르는 확률변수들이라고 할 때, 주어진 분류함수 $C : \mathbb{R}^p \to \{0,1\}$에 대해 모위험을 $R(C) = \mathbb{E}_{(X,Y)} I(C(X) \neq Y)$ 라 하자. $R(C)$을 최소로하는 분류모형은 $I(P(Y|X) > 0.5)$임을 보이시오.

    b. 주어진 손실함수 $l : \{0,1\} \to \mathbb{R}_+$에 대해 모위험을 $R_l(C) = \mathbb{E}_{(X,Y)}[l(Y)I(C(X) \neq Y)]$로 정의하자. $R_l(C)$을 최소화하는 분류모형을 구하시오.

15. 배깅(bagging)에서 가지치기(pruning)을 하지 않는 이유를 설명하시오.

16. a. 로짓부스팅(logit boosting)을 설명하고 일반화 가법 모형(generalized additive model)과의 관계를 설명하여라.

    b. 부스팅에서 shrinkage parameter의 역할을 설명하여라.

    c. 부스팅에서 week learner들의 크기의 의미를 설명하여라.

    d. Stump를 사용하는 부스칭 모형과 coarse classification과의 관계를 설명하여라.

17. 주어진 함수 f에 대해서 $P(Y|X) = \exp(Yf(X))/(\exp(f(X)) + \exp(-f(X)))$라 하자. 여기서 Y는 1 또는 -1을 갖는 확률변수이다.

    a. $Y^* = I(Y = 1)$이라고 하자. $P(Y^*|X) = \exp(2Y^*f(X))/(1 + \exp(2f(X)))$임을 보이시오.

    b. $(X_1, Y_1), \ldots, (X_n, Y_n)$은 $P(Y|X) = \exp(Yf(X))/(\exp(f(X)) + \exp(-f(X)))$을 따르는 입력-출력 쌍이다. f에 대한 음의 로그우도함수가

$$\sum_{i=1}^{n} \log\{1 + \exp(-2Y_i f(X_i))\}$$

    임을 보이시오.

18. Empirical risk minimization(또는 M-추정량)에 대해서 설명하고 이 방법을 사용한 알고리즘의 예를 하나만 드시오.