숙제 2

숙제에 대한 답을 pdf로 변환하여 블랙보드에 제출하시오. 답변에 필요한 코드를 포함하여 하나의 pdf 파일로 만들어 제출하시오. (각 문제 10점, 분석문제 50점)

1. $X = [x_1, x_2, \dots, x_n]$와 $Y = [y_1, y_2, \dots, y_n]$ 로 주어지는 n개의 표본에 대하여, $Y \approx \beta_0 + \beta_1 X$ 로 주어지는 단순 선형 회귀 모델을 훈련하고자 한다. 최소제곱법을 이용하여 파라메터를 추정할 때, 아래와 같이 추정됨을 보이시오.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2. 빨간공(R)과 검은공(B)이 섞여 있는 항아리에서 빨간공의 비율(p)을 최대우도법으로 추정하고자 한다. 하나의 공을 꺼냈을 때, 빨간공일 확률은 p로 검은공일 확률은 $1-p$로 모델링 된다. p를 추정하기 위해서, 실제로 공을 꺼내는 실험을 통해 데이터를 수집하였다. 5번 꺼내기를 반복한 결과, 빨간공-검은공-검은공-빨간공-검은공 의 데이터를 얻었다. 다음의 순서에 따라, p를 추정하시오.

(1) 이 경우에 우도는 얼마인가? 우도를 최대화하는 p는 얼마인가?

(2) 로그우도는 얼마인가? 로그우도를 최대화하는 p는 얼마인가?

3. 로지스틱 회귀로 A, B, C개의 분류에 대하여 멀티클래스 분류를 하는 경우, 아래와 같은 두 개의 이진분류 문제로 나누어 풀 수 있다. 이때, 세 분류에 속할 확률 $\Pr[Y = A|X]$, $\Pr[Y = B|X]$, $\Pr[Y = C|X]$ 각각을 계수와 입력변수를 이용해 표현하시오.

$$\log\left(\frac{\Pr[Y = A|X]}{\Pr[Y = C|X]}\right) = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2$$

$$\log\left(\frac{\Pr[Y = B|X]}{\Pr[Y = C|X]}\right) = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2$$

$$\Pr[Y = A|X] + \Pr[Y = B|X] + \Pr[Y = C|X] = 1$$

4. ROC 커브는 항상 단조 증가함을 증명하시오. 다시 말하면, false positive rate이 증가함에 따라서 true positive rate은 항상 같거나 증가함을 증명하시오.

5. 총 p개의 변수를 갖는 선형 모델에서 변수 선택 기법을 적용하고자 한다. 다음의 경우에 실제로 조사를 해야할 모델을 수를 계산하시오.

(1) 최적의 해법 (best selection)

(2) Forward Stepwise Selection

(3) Backward Stepwise Selection

6. (분석문제) data99_churn_train.csv를 이용하여 통신망 가입자의 서비스탈퇴(churn) 여부에 대한 예측 모델을 훈련하고, data99_churn_test.csv를 이용하여 평가하시오.

(1) 훈련 데이터를 이용하여 데이터에 대한 전처리를 수행하시오. 분석에서 제외할 변수와 샘플이 존재하는가? 결측치나 이상치가 존재하는가? 존재한다면 어떤 방식으로 처리하였는가?

(2) 각 변수는 변환이 필요한가? 필요하다면 어떤 식으로 변환되었는가? 변수의 변환은 가변수화, 정규화, 분포변환 등을 포함한다.

(3) 모든 변수를 이용하여 KNN 예측 모델을 만들고 교차검증을 통하여 K를 튜닝하시오. 최적의 K 값은 얼마인가?

(4) PCA를 통하여 주성분 분석을 수행하고 처음 5개의 주성분만을 이용하여 다시 KNN 모델을 만드시오. 이 때 교차검증을 통해 얻어진 최적의 K 값은 얼마인가?

(5) Logistic Regression을 이용하여 (3)과 (4)를 반복하시오. 이때는 C값을 튜닝하시오.

(6) (3)~(5)를 통해 살펴본 4개의 모델 중에서 평가 데이터에서 가장 성능이 좋을 것으로 예상되는 모델을 하나 선정하시오. 해당 모델에 대해서 평가 데이터에서의 성능을 측정하시오.