숙제 1

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

1. n개의 값, $x_1, x_2, \ldots, x_n$을 하나의 값 m으로 대표하고자 한다. 대표값과 실제값들의 제곱오차의 합, $\sum_{i=1}^{n}(x_i - m)^2$을 최소화하는 m으로 대표하고자 한다. 이때, m의 값은 어떻게 주어지는가?

2. n개의 관측 데이터 $X_1, X_2, \ldots, X_n$은 독립적으로 같은 모집단에서 추출되었다. 모집단의 평균과 분산은 각각 μ, σ^2 라고 하자. 데이터의 관측된 평균을 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 라고 하자.

(1) $E[\bar{X}]$ 는 얼마인가?

(2) $E[\sum_{i=1}^{n}(X_i - \bar{X})^2]$ 는 얼마인가?

(3) $E[k \cdot \sum_{i=1}^{n}(X_i - \bar{X})^2] = \sigma^2$ 이 되도록 k를 정한다면, k의 값은 얼마인가?

(4) 위의 사실을 이용하여 표본분산의 계산에서 n이 아니라 $n-1$로 나누는 이유를 설명하시오.

3. 다음의 물음에 답하시오.

(1) 0에서 1사이에 n개의 점이 임의로 분포하고 있다. 이때, 한 점에서 가장 가까운 점까지의 평균적인 거리는 대략적으로 $1/n$에 비례함을 설명하시오. 여기서는 n이 충분히 크다고 생각한다. 엄밀하게 수학적으로 증명할 필요는 없음.

(2) $(0,1)^p$로 주어지는 p차원 공간(한 변의 길이가 1인 p차원 입방체)에 n개의 점이 임의로 분포하고 있다. 이때, 한 점에서 가장 가까운 점까지의 평균적인 거리는 얼마인가?

(3) $(0,1)^p$로 주어지는 p차원 공간에 n개의 점이 임의로 분포하고 있다. 만일, 차원이 두 배가 된다면, 가장 가까운 점까지의 평균적인 거리를 같게 만들기 위해서는 몇 개의 점이 필요한가?

(4) 위의 사실을 이용하여 차원의 저주를 설명하시오.

4. 재원이는 앞면이 나오면 이기고 뒷면이 나오면 지는 동전 던지기 게임을 하고 있다. 10번 던진 결과 앞면이 1번 나오고 뒷면이 9번 나왔다. 재원이는 이 동전이 공정하지 않은 조작된 동전이 아닐까 의심한다. 다음의 물음에 답하시오.

(1) 공정한 동전이라고 가정할 때, 앞면이 1번 나오고 뒷면이 9번 나올 확률은 얼마인가?

(2) 공정한 동전이라고 가정할 때, 지금 내가 관측한 상황과 같거나 더 극단적인 결과가 나올 확률은 얼마인가? 다시 말하면, (앞면,뒷면)이 (1,9), (9,1), (10,0), (0,10)으로 나올 확률은 얼마인가? 이 확률을 바탕으로 이 동전은 공정하지 않은 동전이라고 말할 수 있는가?

(3) 앞면 1번 뒷면 9번이라는 관측결과에 대하여, 이 동전이 공정한 동전인지를 통계적으로 확인하고자 한다. 동전이 앞면이 나올 확률을 θ라고 할 때, 귀무가설과 대립가설을 세우고 p값을 구하시오. 유의수준 0.05에서 이 가설을 검정하시오.

5. 직종(White/Blue)과 소비성향(A/B/C)에 대한 크로스테이블에 대하여 카이제곱 검정을 수행하고자 한다.

	A	B	C	Total
White	90	70	140	300
Blue	60	80	60	200
Total	150	150	200	500

다음의 물음에 답하시오.

(1) 직종과 소비성향이 완전히 독립적이라고 가정할 때, White이면서 C인 사람들이 평균적으로 관측되는 숫자는 얼마인가?

(2) 카이제곱 검정을 위한 Q값을 계산하여, 소수점이하 둘째자리까지 반올림하여 나타내시오.

6. (분석문제) data99_churn_train.csv는 통신망 가입자의 서비스탈퇴(churn) 여부에 대한 데이터이다. 데이터에서 가장 마지막 열에 있는 churn 변수가 관심 변수이고, 나머지는 이를 설명하기 위한 변수이다. 이 데이터를 이용하여 아래의 질문에 답하시오. 답변에 필요한 코드를 첨부하여 제출하시오.

(1) 이 데이터는 모두 몇 개의 변수와 몇 개의 표본으로 이루어져 있는가?

(2) 각 변수에 대하여 변수의 이름을 이용하여 변수의 의미를 설명하고, 각 변수가 수치형인지 범주형인지 확인하시오.

(3) 결측치를 확인하시오. 총 몇 개의 결측치가 존재하는가? 결측치를 갖고 있는 변수는 몇 개인가? 결측치를 포함하는 표본은 몇 개인가? 결측치를 갖는 표본을 제거하시오.

(4) 각 변수에 대하여 요약하시오. 수치형 변수는 평균과 분산을 제공하시오. 범주형 변수에 대해서는 도수분포를 제공하시오. 수치형 변수에 대해서는 히스토그램을 제공하시오.

(5) (2)와 (4)의 결과를 바탕으로 customerID를 후속 분석에서 제외해야하는 이유를 설명하시오.

(6) Churn와 다른 모든 변수(customerID 제외) 사이에서, 수치-범주의 경우에는 boxplot을 그리고, 범주-범주인 경우에는 크로스테이블을 구하시오.

(7) Churn와 다른 모든 변수(customerID 제외) 사이에서 적절한 관계검정을 수행하고 p값을 구하시오. P값 기준으로 가장 중요한 변수 3개와 가장 중요하지 않은 변수 3개를 각각 선정하시오.

답안

1.

$L = \sum_{i=1}^{n}(x_i - m)^2$ 일 때, $\frac{dL}{dm} = 2\sum_{i=1}^{n}(x_i - m) = 0$ 을 풀어 최소가 되는 지점을 구하면 $m = \frac{1}{n}\sum_{i=1}^{n} x_i$, 즉 평균으로 주어짐.

2.

(1)

$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$ 이고 iid에서 $E[X_i] = \mu$ 이므로, $E[\bar{X}] = \mu$

(2)

모든 X_i 에 대하여 iid이므로 $E[X_i] = \mu$ 이고 $Var[X_i] = E[X_i^2] - \mu^2 = \sigma^2$ 에서 $E[X_i^2] = \sigma^2 + \mu^2$.

$E[\sum_{i=1}^{n}(X_i - \bar{X})^2] = \sum_{i=1}^{n} E[X_i^2 - 2X_i\bar{X} + \bar{X}^2] = \sum_{i=1}^{n} E[X_i^2] - 2\sum_{i=1}^{n} E[X_i\bar{X}] + \sum_{i=1}^{n} E[\bar{X}^2]$

i) $\sum_{i=1}^{n} E[X_i^2] = n(\sigma^2 + \mu^2)$

ii) $\sum_{i=1}^{n} E[X_i\bar{X}] = \sum_{i=1}^{n} E\left[X_i \cdot \frac{1}{n}(X_1 + \cdots + X_n)\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_iX_1 + \cdots + X_iX_n] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iX_j]$. 여기 $\sum_{j=1}^{n} E[X_iX_j]$ 에서 $E[X_iX_j]$를 구하면, $i = j$인 1가지의 경우 $E[X_iX_j] = E[X_i^2] = \sigma^2 + \mu^2$로 주어지고, $i \neq j$인 $n-1$가지의 경우 독립이므로 $E[X_iX_j] = E[X_i]E[X_j] = \mu^2$. 결국, $\sum_{j=1}^{n} E[X_iX_j] = \sigma^2 + n\mu^2$. 그러므로 $\sum_{i=1}^{n} E[X_i\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iX_j] = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + n\mu^2) = \sigma^2 + n\mu^2$.

iii) $\sum_{i=1}^{n} E[\bar{X}^2] = nE[\bar{X}^2] = \frac{1}{n}E[(X_1 + \cdots + X_n)(X_1 + \cdots + X_n)] = \frac{1}{n}\sum_{1\le i,j\le n} E[X_iX_j]$. $\sum_{1\le i,j\le n} E[X_iX_j]$ 는 1과 n 사이의 모든 i와 j의 조합에 대하여 $E[X_iX_j]$을 더하는 것으로 총 n^2 개의 항을 더하게 됨. 이 중에서 $i = j$ 인 n 가지의 경우 $E[X_iX_j] = E[X_i^2] = \sigma^2 + \mu^2$, $i \neq j$ 인 $n^2 - n$ 가지의 경우 $E[X_iX_j] = E[X_i]E[X_j] = \mu^2$. 종합하면, $\sum_{i=1}^{n} E[\bar{X}^2] = \frac{1}{n}\{n(\sigma^2 + \mu^2) + (n^2 - n)\mu^2\} = \sigma^2 + n\mu^2$.

종합하면, $E[\sum_{i=1}^{n}(X_i - \bar{X})^2] = n(\sigma^2 + \mu^2) - 2(\sigma^2 + n\mu^2) + \sigma^2 + n\mu^2 = (n-1)\sigma^2$

(3) $E[k \cdot \sum_{i=1}^{n}(X_i - \bar{X})^2] = \sigma^2$ 이기 위해서 $k = \frac{1}{n-1}$.

(4) 표본분산을 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 로 정의하는 경우 $E[S^2] = \sigma^2$ 이 되기 때문에 표본분산의 기대값이 실제 분산과 같아진다. (이것을 우리는 unbiased라고 이야기한다.) 하지만 직관적으로 n으로 나눈다면 $E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n}\sigma^2$ 이기 때문에, 실제 분산과 달라 bias가 생긴다.

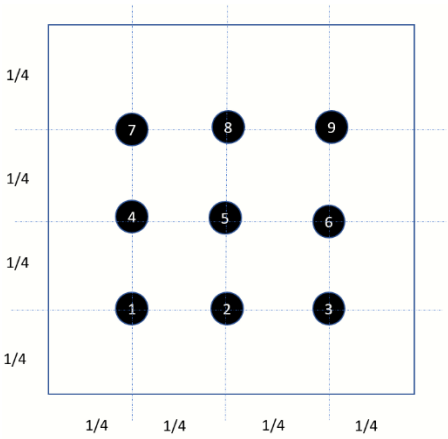
3.

(1)

0과 1 사이에 n개의 점이 균일하게 분포하는 경우를 생각해보면 각 점 사이의 거리는 대략 $1/n$로 주어진다. 그러므로 양 옆에 있는 두 점 중 더 가까운 점의 거리도 $1/n$ 에 비례할 것으로 생각할 수 있다.

(2)

$p = 2$ 일 때를 생각해보면, 매우 균일하게 점이 분포하는 경우, 아래와 같이 가로축으로 $\sqrt{n}$ 개가 세로축으로 $\sqrt{n}$ 개가 균일하게 분포하는 것을 생각해 볼 수 있다. 이 경우 가장 가까운 점까지의 거리는 대략 $1/\sqrt{n}$ 임을 알 수 있다. 이것을 일반적인 p 차원으로 확장해 보면, 대략적인 가장 가까운 점까지의 거리는 $\frac{1}{n^{1/p}}$ 로 주어짐을 생각할 수 있다. 이는 n이 클 때의 대략적인 결과이고 정확한 결과는 훨씬 복잡하게 주어진다.



(3) $\frac{1}{n^{1/p}} = \frac{1}{m^{1/2p}}$이기 위해서는 $m = n^2$.

(4) n개의 표본과 p개의 변수로 주어지는 데이터 행렬이 있다고 하자. p차원의 공간을 n개의 포인트가 채우고 있는 형태이다. 이때, 포인트(표본)의 밀도가 높을수록 데이터가 많아지는 것이기 때문에 분석이 쉬워진다. 표본의 밀도는 다양한 방식으로 측정할 수 있지만 가장 가까운 점까지의 거리로도 측정이 가능하다. 위의 결과와 같이, 만일 변수의 개수가 2배로 늘어난다면 비슷한 정도로 공간을 채우기 위해서 필요한 표본의 수는 n^2 이다. 결국, 변수의 수가 2배가 되면 같은 정도의 분석력을 갖기 위해 필요한 표본의 수는 제곱으로 늘어난다. 이는 차원의 증가 속도를 표본의 증가 속도가 따라갈 수 없다는 의미로, 곧 차원의 저주를 의미한다.

4.

(1) 1번 앞면 9번 뒷면이 나오는 경우의 수 10회. 각 경우에 대한 확률은 1/1024. 확률은 10/1024.

(2) 모두 앞면이 나올 확률 = 1/1024, 모두 뒤면이 나올 확률 = 1/1024, 앞면 1번 뒷면 9번 확률 =10/1024, 앞면 9번 뒷면 1번 확률 = 10/1024. 모든 확률 = 22/1024 = 0.02. 이 동전이 공정한 동전일 때, 재원이가 관측한 경우와 같거나 더 극단적인 경우를 볼 확률은 0.02에 불과하다. 단 한 번 실험을 했을 때 일반적으로 우리가 관측할 수 있는 확률이 아니기 때문에, 원래 가정인 공정한 동전이라는 것을 의심할 수 있다.

(3) 귀무가설: $\theta = 0.5$, 대립가설: $\theta \neq 0.5$. 앞선 계산과 같이 $p = 0.02$. 이 값은 유의수준인 0.05보다 작기 때문에, $\theta = 0.5$이라는 귀무가설은 통계적으로 기각된다. 다시 말하면 실제 앞면이 나올 확률은 통계적으로 유의미하게 0.5와 다르다. 곧, 공정한 동전이 아니다.

5.

(1) 기본적으로 직종이 White일 확률은 300/500=3/5. 마찬가지로 소비성향이 C일 확률은 200/500=2/5. 직종과 소비성향이 독립이라면 White이면서 C일 확률은 3/5 * 2/5 = 6/25 = 0.24. 그러므로 예상되는 관측값은 0.24 * 500 = 120.

(2) 예상되는 관측값은

	A	B	C	Total
White	90	90	120	300
Blue	60	60	80	200
Total	150	150	200	500

$Q = (90-90)^2/90 + (60-60)^2/60 + (70-90)^2/90 + (80-60)^2/60 + (140-120)^2/120 + (60-80)^2/80 = 19.44$

6.

(1) (4696, 21)의 shape으로 21개의 변수와 4696개의 행으로 이루어진 표본개수 98616로 이루어짐.

(2)

변수 이름	설명	변수 형태
Customor ID'	통신망 가입자 ID	범주형
gender'	통신망 가입자 성별	범주형
seniorCitizen	통신망 가입자의 시니어 존재 여부	범주형
Partner	Partner 여부	범주형
Dependents	자녀 여부	범주형
tenure	가입기간	수치형
PhoneService	핸드폰 서비스 여부	범주형
MultipleLines	Multiple Lines 여부	범주형
Internet Service	인터넷 서비스 여부	범주형
OnlineSecurity	online security 여부	범주형
OnlineBackup	Online Backup 여부	범주형
DeviceProtection	Device Protection 여부	범주형
TechSupport	Tech Support 여부	범주형
StreamingTV	Streaming TV 여부	범주형
StreamingMovies	Streaming Movies 여부	범주형
Contract	계약 기간	범주형
PaperlessBilling	Paperless Billing 여부	범주형
PaymentMethod	결제 방식	범주형
MonthlyCharges	월 결제 금액	수치형
TotalCharges	총 결제 금액	수치형
Churn	가입자 이탈 여부	범주형

(3) 결측치 수 : 4, 결측치를 갖고 있는 변수의 수 : 1 (TotalCharges), 결측치 포함 표본의 수 : 4

(4) 코드 참조

(5) customer ID의 경우 모든 샘플에 대해 다른 값을 갖는 무의미한 값이기 때문에 churn과의 관련성은 없다고 볼 수 있다.

(6) 코드 참조

(7)

가장 중요한 변수 : 'Contract', 'tenure', 'OnlineSecurity'

가장 중요하지 않은 변수 : 'PhoneService', 'gender', 'MultipleLines'