머신러닝을 위한 통계학

확률론과 통계이론

7기 교육부 홍승기

확률론

확률

확률 공리

- 표본공간(sample space) : 가능한 모든 관측 결과들의 집합
- 사건(event) : 표본공간의 부분집합인 특정한 결과들의 집합
- 확률(probability) : 각 사건의 가능성을 수치화한 것
- 확률 공리(axioms of probability) : 다음의 세 조건을 확률 공리라고 한다.
- (확률의 범위) 각 사건 A 에 대해 $P(A) \geq 0$
- (전체 확률) 표본공간 S 에 대해 $P(S) = 1$
- (가산가법성; countable additivity) 사건 열 $\{A_n\}_{n \geq 1}$ 에 대해 $A_i \cap A_j = \emptyset$ 이면 $P(\cup_{n=1}^{\infty} A_n) = \Sigma_{n=1}^{\infty} P(A_n)$

확률

확률의 성질

- 각 사건 A 에 대해 $0 = P(\emptyset) \leq P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- (단조성; monotonicity) $A \subseteq B$ 이면 $P(A) \leq P(B)$
- (반가법성; subadditivity) $A \subseteq \cup_{n=1}^{\infty} A_n$ 이면 $P(A) \leq \Sigma_{n=1}^{\infty} P(A_n)$
- (연속성; continuity)

  - $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$ 이면 $P(\cup_{n=1}^{\infty} A_n) = \lim_{n \to \infty} P(A_n)$
  - $B_1 \supseteq B_2 \supseteq \cdots \supseteq B_n \supseteq \cdots$ 이면 $P(\cap_{n=1}^{\infty} B_n) = \lim_{n \to \infty} P(B_n)$

확률

조건부확률

- 한 사건 A 가 일어났다는 전제하에서 사건 B 가 일어날 가능성을 생각하는 경우의 확률을 조건부 확률(conditional probability)라 하며, 다음과 같이 정의한다:
- $$P(B\,|\,A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$
- 한 편, 조건부확률은 다음과 같은 성질들을 가지고 있다:
- (곱셈공식) $P(A) > 0,\ P(B) > 0$ 이면 $P(A \cap B) = P(B\,|\,A)P(A) = P(A\,|\,B)P(B)$
- (전확률공식) 사건 열 $\{A_n\}_{n \geq 1}$ 이 표본공간 S 를 공통부분이 없게 분할할 때, 즉 $A_i \cap A_j = \emptyset\ (i \neq j),\ \cup_{n=1}^{\infty} A_n = S$ 일 때, $P(A_i) > 0$ 이면 $P(B) = \Sigma_{n=1}^{\infty} P(B\,|\,A_n)P(A_n)$

확률

베이즈 정리(Bayes' Theorem)

- 사건 열 $\{A_n\}_{n\geq 1}$ 이 표본공간 S 를 공통부분이 없게 분할하고 $P(A_i) > 0$ 일 때, $P(B) > 0$ 이면 다음 비례식이 성립한다:
- $P(A_j \mid B) \propto P(B \mid A_j)P(A_j) \quad (j = 1,2,\cdots)$
- 베이즈 정리는 베이지안 추론의 근본이 되는 정리로, $P(A_j)$ 는 여러 모형의 가능성을 뜻하고, $P(A_j \mid B)$ 는 실험 결과를 뜻하는 B 의 관측 후에 각 모형의 가능성을 뜻한다.
- 이러한 이유에서 $P(A_j)$, $P(A_j \mid B)$ 를 각각 사전(prior), 사후(posterior) 확률이라 부르고 베이즈 정리는 이들 사이의 관계를 나타낸 것이다.

확률

사건의 독립성

- 사건 A 의 관측 여부가 사건 B 가 일어날 가능성에 아무런 영향을 주지 않는 것을 $P(B \mid A) = P(B)$ 와 같이 나타낼 수 있고, 이러한 관계식은 곱셈공식으로부터$P(A \cap B) = P(A)P(B)$ 와 같은 것임을 알 수 있다.
- 이러한 경우에는, A 의 관측 여부와 독립적으로 B 의 관측 가능성이 정해진다는 의미에서 두 사건이 서로 독립(independent)이라고 한다.
- 일반적으로, n 개의 사건이 서로 독립(mutually independent)인 경우엔 다음과 같이 정의한다:
- $P(A_i \cap A_j) = P(A_i)P(A_j)\ (1 \leq i < j \leq n)$
- $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)\ (1 \leq i < j < k \leq n)$
- $\cdots$
- $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$

확률변수와 확률분포

확률변수의 정의

- 여러 가지의 결과가 가능하고 그 가능성을 확률로 나타낼 수 있는 실험을 랜덤한 실험(random experiment)이라 부르고, 이러한 실험의 모든 가능한 결과의 집합인 표본공간에서 정의된 실수 값 함수를 확률변수(random variable)라고 한다. (정확한 정의는 참고자료의 Probability and Measure 참고)

- 확률변수는 대문자를 사용하여 X, Y 등으로 표기하고, 확률변수의 값에 관한 사건을 $\{X = 1\}, \{X \leq 1\}$ 와 같이 나타낸다. 이러한 사건들은 표본공간에서의 대응하는 실험결과들의 집합으로서 확률을 갖게 된다.

확률변수와 확률분포

확률밀도함수의 정의

- 확률밀도함수(probability density function; pdf)는 확률변수의 형태에 따라 다르게 정의된다.
- 확률변수가 가질 수 있는 값들의 집합이 $\{x_1, x_2, \cdots\}$ 와 같을 때 이산형(discrete type)이라 하고 각각의 값에 그 값을 가질 확률을 대응시키는 함수, 즉 $f(x_k) = P(X = x_k)\ \forall k$ 를 X 의 확률밀도함수라 한다.
- 이산형 확률밀도함수는 다음의 성질들을 만족한다:
- $f(x) \geq 0\ \forall x \in \mathbb{R}, \quad f(x) = 0\ \forall x : x \neq x_k, k = 1,2,\cdots$
- $\Sigma_x f(x) = \Sigma_{k=1}^{\infty} f(x_k) = 1$
- $\Sigma_{x:a \leq x \leq b} f(x) = P(a \leq X \leq b)$
- 확률변수가 실수 구간의 값들을 가질 수 있고 그에 관한 확률이 적분으로 주어질 때 연속형(continuous type) 이라 하고 확률을 정해주는 함수, 즉 $\int_a^b f(x)dx = P(a \leq X \leq b)$ 인 함수 f 를 X 의 확률밀도함수라 한다.
- 연속형 확률밀도함수는 다음의 성질들을 만족한다:
- $f(x) \geq 0\ \forall x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = 1, \int_a^b f(x)dx = P(a \leq X \leq b)\ (-\infty < a < b < \infty)$

확률분포의 특성치

확률분포의 기댓값과 분산

- 확률변수 X 의 확률밀도함수가 f 일 때, 실수 값 함수 $g(x)$ 의 기댓값(expectation) 은 다음과 같이 정의한다:
- $$E[g(X)] = \begin{cases} \sum_x g(x)f(x), & X \text{ 가 이산형일 때} \\ \int_{-\infty}^{\infty} g(x)f(x)dx, & X \text{ 가 연속형일 때} \end{cases}$$
- 단, 위의 연산이 실수로 정의될 때만 유효하다.
- 만약 $g(x) = x$ 라면, 위의 기댓값은 곧 확률변수 X 의 평균을 의미한다.
- 한 편, 확률변수 X 의 평균을 μ 라고 할 때, X 의 분산(variance)는 다음과 같이 정의한다:
- $$Var(X) = E\left[(X-\mu)^2\right]$$

확률분포의 특성치

확률분포의 기댓값과 분산의 성질

- 기댓값의 성질은 다음과 같다:
- (선형성; linearity) $E[c_1 g_1(X) + c_2 g_2(X)] = c_1 E[g_1(X)] + c_2 E[g_2(X)]$, c_1, c_2 는 상수
- (단조성; monotonicity) $g_1(X) \leq g_2(X)$ 이면 $E[g_1(X)] \leq E[g_2(X)]$
- 분산의 성질은 다음과 같다:
- $Var(aX + b) = a^2 Var(X)$, a 는 상수
- $Var(X) = E[X^2] - E[X]^2$

누적분포함수

누적분포함수의 정의와 성질

- 확률변수의 누적분포함수(cumulative distribution function; cdf) 는 다음과 같이 정의된다:
- $$F(x) = P(X \leq x) = \begin{cases} \sum_{t:t \leq x} f(t), & X \text{ 가 이산형일 때} \\ \int_{-\infty}^{x} f(t)dt, & X \text{ 가 연속형일 때} \end{cases}$$
- 누적분포함수의 성질은 다음과 같다:
- $x_1 < x_2$ 이면 $F(x_1) \leq F(x_2)$
- $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$
- $\lim_{h \downarrow 0} F(x+h) = F(x)$

여러가지 부등식

젠센, 리아푸노프, 마코프, 체비셰프 부등식

- (젠센 부등식; Jensen's inequality) : 수직선 위의 구간 I 에서의 값을 갖는 확률변수 X 의 기댓값이 존재하면, 구간 I 에서 볼록한 함수 ϕ 에 대해 다음이 성립한다:
- $\phi(E[X]) \leq E[\phi(X)]$
- (리아푸노프 부등식; Liapounov's inequality) : 확률변수 X 에 대해 $E[\,|X|^s] < \infty$, $0 < r < s$ 이면, 다음이 성립한다:
- $E[\,|X|^r]^{\frac{1}{r}} \leq E[\,|X|^s]^{\frac{1}{s}}$
- (마코프 부등식; Markov's inequality) : 확률변수 X 에 대해 $E[\,|X|^r] < \infty$, $r > 0$ 이면, 임의의 양수 $\epsilon > 0$ 에 대해 다음이 성립한다:
- $P(\,|X| \geq \epsilon) \leq \dfrac{E[\,|X|^r]}{\epsilon^r}$
- (체비셰프 부등식; Chebyshev's inequality) : 확률변수 X 에 대해 $Var(X) < \infty$ 이면, 임의의 양수 $\epsilon > 0$ 에 대해 다음이 성립한다:
- $P(\,|X - E[X]\,| \geq \epsilon) \leq \dfrac{Var(X)}{\epsilon^2}$

다차원 확률변수의 분포

확률벡터와 결합확률밀도함수

- 확률변수 $X_1, \cdots, X_k$ 를 각 성분으로 하는 벡터 $(X_1, \cdots, X_k)^T$ 를 k 차원 확률변수 또는 k 변량(k-variate) 확률벡터(random vector)라 한다.
- 1차원의 경우와 마찬가지로, 확률벡터는 이산형, 연속형으로 나뉘며, 이 경우에 확률밀도함수에 대응되는 개념으로 결합확률밀도함수(joint probability density function)를 정의할 수 있다.
- 이산형의 경우:
- $f(x_1, \cdots, x_k) \geq 0, \forall x_i \in \mathbb{R}, i = 1, \cdots, k$
  $f(x_1, \cdots, x_k) = 0, \forall x \notin \{x_{i1}, x_{i2}, \cdots\}, i = 1, \cdots, k$
- $\sum_{x_1} \cdots \sum_{x_k} f(x_1, \cdots, x_k) = 1$
- $\sum_{x_1 : a_1 \leq x_1 \leq b_1} \cdots \sum_{a_k \leq x_k \leq b_k} f(x_1, \cdots, x_k) = P(a_1 \leq X \leq b_1, \cdots, a_k \leq X_k \leq b_k)$
- 연속형의 경우 또한 비슷하게 정의할 수 있다.

다차원 확률변수의 분포

주변확률밀도함수

- 확률벡터 $(X_1, \cdots, X_k)^T$ 의 결합확률밀도함수가 $f(x_1, \cdots, x_k)$ 일 때, X_1 과 $(X_1, X_2)^T$ 의 주변확률밀도함수(marginal probability density distribution) $f_1(x)$와 $f_{1,2}(x, y)$는 다음과 같이 주어진다:

- $$f_1(x) = \begin{cases} \sum_{x_2} \cdots \sum_{x_k} f(x, x_2, \cdots, x_k) & \text{(이산형인 경우)} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x, x_2, \cdots, x_k) dx_k \cdots dx_2 & \text{(이산형인 경우)} \end{cases}$$

- $$f_{1,2}(x, y) = \begin{cases} \sum_{x_3} \cdots \sum_{x_k} f(x, y, x_3, \cdots, x_k) & \text{(이산형인 경우)} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x, y, x_3, \cdots, x_k) dx_k \cdots dx_3 & \text{(이산형인 경우)} \end{cases}$$

다차원 확률변수의 분포

확률벡터 함수의 기댓값

- 확률벡터 $(X_1, \cdots, X_k)^T$ 의 결합확률밀도함수가 $f(x_1, \cdots, x_k)$ 일 때 실수 값 함수 $g(x_1, \cdots, x_k)$ 의 기댓값(expectation) 은 다음과 같이 정의한다:
- $$E[g(X_1, \cdots, X_k)] = \begin{cases} \sum_{x_1} \cdots \sum_{x_k} g(x_1, \cdots, x_k) f(x_1, \cdots, x_k) & \text{(이산형인 경우)} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_k) f(x_1, \cdots, x_k) dx_k \cdots dx_1 & \text{(이산형인 경우)} \end{cases}$$
- 단, 위의 연산이 실수로 정의될 때만 유효하다.
- 일변량일 때와 마찬가지로, 선형성과 단조성의 성질을 가지고 있다.

다차원 확률변수의 분포

확률벡터의 평균, 분산행렬, 공분산행렬

- 확률벡터 $(X_1, \cdots, X_k)^T$ 의 성분들인 $X_1, \cdots, X_k$ 의 평균, 분산, 공분산인 $\mu_i = E[X_i]$, $\sigma_{i,j} = Cov(X_i, X_j)$ 를 대응하는 원소로 갖는 벡터와 행렬을 각각 X 의 평균벡터 (mean vector), 분산·공분산 행렬(variance-covariance matrix) 또는 간략히 X 의 평균, 분산행렬 이라고 하며, 기호로는 다음과 같이 정의한다:

- $E[X] = (\mu_1, \cdots, \mu_k)^T = (E[X_1], \cdots, E[X_k])^T$, $Var(X) = (\sigma_{i,j})_{1 \le i,j \le k} = (Cov(X_i, X_j))_{1 \le i,j \le k}$

- 한 편 $X = (X_1, \cdots, X_k)^T$, $Y = (X_1, \cdots, X_l)^T$ 의 평균을 각각 $\mu = (\mu_1, \cdots, \mu_k)^T$, $\eta = (\eta_1, \cdots, \eta_l)^T$ 라 하면 X 의 분산행렬과 X 와 Y 의 공분산행렬을 각각 다음과 같이 정의된다:

- $Var(X) = E[(X - \mu)(X - \mu)^T]$, $\quad Cov(X, Y) = E[(X - \mu)(Y - \eta)^T]$

다차원 확률변수의 분포

확률벡터의 평균, 분산행렬, 공분산행렬

- A, C 가 상수의 행렬, b, d 가 상수의 벡터일 때, 확률벡터 X, Y, Z, W 에 대하여 다음의 성질들이 만족한다:
- $E[AX + b] = AE[X] + b$
- $Var(AX + b) = AVar(X)A^T$
- $Cov(AX + b, CY + d) = ACov(X, Y)C^T$
- $Cov(X + Y, Z + W) = Cov(X, Z) + Cov(Y, Z) + Cov(X, W) + Cov(Y, W)$
- $Cov(Y, X) = (Cov(X, Y))^T, \quad Var(X) = Cov(X, X)$
- $Var(X + Y) = Var(X) + Var(Y) + Cov(X, Y) + Cov(Y, X)$
- 이 때, X 의 분산행렬 $Var(X)$ 는 위의 성질들과 분산의 정의로부터 SPSD 행렬임을 알 수 있다.

다차원 확률변수의 분포

다차원 확률변수에 대한 조건부 확률밀도함수와 조건부기댓값

- 확률벡터 $X = (X_1, \cdots, X_k)^T$, $Y = (X_1, \cdots, X_l)^T$ 에 대해 X 와 Y 의 결합확률밀도함수가 $f_{X,Y}(x, y)$ 이고 X 의 주변확률밀도함수가 $f_X(x)$ 일 때 $X = x$ 인 조건에서 Y 의 조건부 확률밀도함수(conditional probability density function)는 다음과 같이 정의한다:
- $f_{Y|X}(y \,|\, x) = \dfrac{f_{X,Y}(x, y)}{f_X(x)}, \quad \forall x : f_X(x) > 0$
- 한 편, $X = x$ 인 조건에서 실수값 함수 $g(X, Y)$ 의 조건부기댓값(conditional expectation)은 다음과 같이 정의한다:
- $$E[g(X, Y) \,|\, X = x] = \begin{cases} \sum_{y_1} \cdots \sum_{y_l} g(x, y_1, \cdots, y_l) f_{Y|X}(y_1, \cdots, y_l \,|\, x) & \text{(이산형)} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x, y_1, \cdots, y_l) f_{Y|X}(y_1, \cdots, y_l \,|\, x) dy_l \cdots dy_1 & \text{(연속형)} \end{cases}$$
- 이 때 조건부 평균과 조건부 분산행렬은 다음과 같이 정의한다:
- $E[Y \,|\, X] = (E[Y_1 \,|\, X], \cdots, E[Y_l \,|\, X])^T, \quad Var(Y \,|\, X) = (Cov(Y_i, Y_j \,|\, X))_{1 \le i,j \le l}$

다차원 확률변수의 분포

조건부기댓값의 성질

- 조건부기댓값은 다음과 같은 성질을 가지고 있다:
- $E[E[Y|X]] = E[Y]$
- $Cov(Y - E[Y|X], v(X)) = 0, \quad \forall v(X)$
- 두 번째 성질은 선형대수의 직교여공간 개념과 비슷하게 이해할 수 있다. 즉, X 만으로 표현할 수 있는 모든 확률변수는 Y 에서 X 의 영향력을 뺀 것과 상관관계가 전혀 없으며, 이는 곧 조건부기댓값 $E[Y|X]$를 Y 에 있는 X 의 영향력을 의미하는 것으로 바라볼 수 있음을 시사한다.
- 한 편, 분산행렬은 다음과 같이 분해할 수 있다:
- $Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$
- 이러한 분산의 분해는 추정량 또는 예측값의 분산, 즉 정확도를 비교할 때에 유용하게 이용된다.
- 마지막으로, 확률벡터 X 의 벡터값 함수 $u(X) = (u_1(X), \cdots, u_k(X))^T$ 에 대해 항상 다음이 성립한다:
- $E[\|Y - E[Y|X]\|_2^2] \leq E[\|Y - u(X)\|_2^2], \quad \forall u(X)$
- 즉, L^2-norm 관점에서 X 를 이용하여 Y 를 가장 잘 설명하는 값은 $E[Y|X]$ 이다. 이는 데이터로 결과를 예측, 추정하는 데에 가장 좋은 값은 조건부기댓값임을 시사한다.

다차원 확률변수의 분포

다차원 확률변수의 독립성

- 여러 개의 확률변수 $X_1, \cdots, X_n$ 에 대하여 다음이 성립할 때, $X_1, \cdots, X_n$ 이 서로 독립이라고 한다:
- $P(X_1 \in A_1, \cdots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n) \quad \forall A_i$
- 한 편, 확률변수 X_i 의 확률밀도함수를 $f_i(x_i)$ 라 할 때, 위의 정의는 다음과 동치이다:
- $f_{1,\cdots,n}(x_1, \cdots, x_n) = f_1(x_1) \cdots f_n(x_n)$
- $X_1, \cdots, X_n$ 이 서로 독립이면 다음의 성질들을 만족한다:
- 각각의 함수인 $g(X_1), \cdots, g(X_n)$ 도 서로 독립이다.
- $E[g_1(X_1) \cdots g_n(X_n)] = E[g_1(X_1)] \cdots E[g_n(X_n)]$
- $Cov(X_i, X_j) = 0, \quad i \neq j$
- $Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n)$

여러가지 확률분포

확률분포의 예시

- 베르누이분포(Bernoulli) : $X \sim Bern(p) \Leftrightarrow f(x) = p^x(1-p)^{1-x},\ x = 0,1$
- 이항분포(Binomial) : $X \sim B(n,p) \Leftrightarrow f(x) = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0,\cdots,n$
- 포아송분포(Poisson) : $X \sim Pois(\lambda) \Leftrightarrow f(x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0,\cdots$
- 정규분포(normal) : $X \sim N(\mu,\sigma^2) \Leftrightarrow f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathbb{R}$

여러가지 확률분포

정규분포의 성질

- $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$ 이고 X_1, X_2가 서로 독립이면 다음이 성립한다:
- $E[X_1] = \mu_1,\ Var(X_1) = \sigma_1^2$
- $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- 상수 a, b 에 대해 $aX_1 + b \sim N(a\mu_1 + b, a^2\sigma_1^2)$
- $X_1 \sim N(\mu_1, \sigma_1^2) \Leftrightarrow X_1 = \sigma_1 Z + \mu_1,\ Z \sim N(0,1)$

여러가지 확률분포

다변량 정규분포의 정의

- $Z_1, \cdots, Z_n \overset{iid}{\sim} N(0,1)$ (iid: identically and independently distributed의 약자)이고 상수 행렬과 벡터 $A = (a_{ij})_{1 \le i,j \le n}, \mu = (\mu_1, \cdots, \mu_n)^T$ 에 대해 확률변수 $X = (X_1, \cdots, X_n)^T$ 를 다음과 같이 정의하자:
- $X = AZ + \mu,\ Z = (Z_1, \cdots, Z_n)^T$
- 이 때 A 의 역행렬이 존재한다면, X 의 확률밀도함수는 다음과 같다:
- $f(x) = |2\pi\Sigma|^{-1} \exp\left[\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right], \quad x \in \mathbb{R}^n, \Sigma = AA^T$
- 이러한 분포를 다변량 정규분포(multivariate normal distribution)라 하며, 기호로는 다음과 같이 표기한다:
- $X \sim N(\mu, \Sigma)$
- 위의 정의에 따르면 $Z = (Z_1, \cdots, Z_n)^T$ 의 분포는 자연스럽게 $Z \sim N(0, I_n)$ 이 된다.
- 따라서 다변량 정규분포는 다음과 같이 정의할 수도 있다:
- $X \sim N(\mu, \Sigma) \Leftrightarrow X = AZ + \mu,\ Z \sim N(0, I_n),\ AA^T = \Sigma$

여러가지 확률분포

다변량 정규분포의 성질

- $X \sim N(\mu, \Sigma)$ 이고 상수의 행렬과 벡터 M, b 에 대해 다음이 성립한다:
- $E[X] = \mu,\ Var(X) = \Sigma$
- 분산행렬 Σ 에 대하여 $AA^T = \Sigma, A = A^T$ 인 행렬 A 를 $\Sigma^{1/2}$ 이라 하면(∵ SPSD 행렬의 대각화정리) 다음이 성립한다:
- $X \sim N(\mu, \Sigma) \Leftrightarrow X = \Sigma^{1/2}Z + \mu,\ Z \sim N(0, I_n)$
- $MX + b \sim N(M\mu + b, M\Sigma M^T)$
- M, N 이 상수의 행렬일 때, $Cov(MX, NX) = M\Sigma N^T = 0$ 이면 MX, NX 는 서로 독립이다.
- $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$ 일 때 다음이 성립한다:
- $Cov(X_1, X_2) = \Sigma_{12} = 0$ 이면 X_1, X_2 는 서로 독립이다.
- $X_1 \sim N(\mu_1, \Sigma_1)$
- 분산행렬의 역행렬이 존재한다면, $X_{2|X_1=x} \sim N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$

여러가지 확률분포

다변량 정규분포의 활용(1)

- (카이제곱분포) $Z \sim N(0, I_n)$ 일 때, 카이제곱분포(χ^2- distribution; central χ^2 - distribution)는 다음과 같이 정의한다:
- $V = Z^T Z = Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n)$
- 이 때 카이제곱분포의 모수 n을 자유도(degree of freedom)라 한다.
- 카이제곱분포와 다변량 정규분포는 밀접한 관계를 가지고 있다. 특히 다음과 같은 성질은 분산분석(ANOVA)의 이론적 근거이다:
- $X \sim N_k(\mu, \Sigma)$ 이고 Σ 가 정칙행렬이면, $(X-\mu)^T \Sigma^{-1} (X-\mu) \sim \chi^2(k)$
- $Z \sim N(0, I)$ 이고 $A^2 = A$ 이면 $Z^T A Z \sim \chi^2(r)$ 이고, $r = rank(A)$
- (t 분포) $Z \sim N(0,1)$ 이고 $V \sim \chi^2(k)$ 이며 Z, V 가 서로 독립일 때, t 분포는 다음과 같이 정의한다:
- $t = \dfrac{Z}{\sqrt{V/k}} \sim t(k)$
- (F 분포) $V_1 \sim \chi^2(k_1), V_2 \sim \chi^2(k_2)$ 이고 V_1, V_2 가 서로 독립일 때, F 분포는 다음과 같이 정의한다:
- $F = \dfrac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2)$

여러가지 확률분포

다변량 정규분포의 활용(2)

- 선형회귀모형은 다음과 같이 일반화시킬 수 있다: $X : n \times p$ 행렬, $\beta \in \mathbb{R}^p$ 라 하면
- $$\begin{cases} Y = X\beta + \epsilon \\ \epsilon \sim N_n(0, \sigma^2 I) \end{cases}$$
- 보통 X 를 설계행렬(design matrix), β 를 회귀계수(regression coefficient)라 한다.
- 여기서 우리의 목적이 β, σ^2을 적절히 추정하는 것이라 할 때, 일반적으로 표본회귀계수, 표본분산은 다음과 같이 추정한다:
- $\hat{\beta} = (X^T X)^{-1} X^T Y$, $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|_2^2/(n-p)$, 단, $rank(X) = rank(X^T X) = p$
- 이 때, 다변량 정규분포와 카이제곱분포의 성질로 인해 다음이 성립한다:
- $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$
- $\hat{\beta}, \hat{\sigma}^2$ 은 서로 독립
- $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$

극한분포

확률수렴과 분포수렴의 정의

- 확률수렴(convergence in probability) : 확률변수 X_n 이 상수 c 로 확률수렴함은 다음과 같이 정의한다:
- $X_n \xrightarrow{p} c \Leftrightarrow \lim_{n\to\infty} P(|X_n - c| \geq \epsilon) = 0 \; \forall \epsilon > 0$
- 분포수렴(convergence in distribution) : 확률변수 X_n이 확률변수 Z 로 분포수렴함은 다음과 같이 정의한다:
- $X_n \xrightarrow{d} Z \Leftrightarrow \lim_{n\to\infty} P(X_n \leq x) = P(Z \leq x), \forall x : Z$ 의 누적분포함수에서 연속

극한분포

대수의 법칙과 중심극한정리

- 대수의 법칙(Weak Law of Large Number; WLLN) : 확률변수 $X_1, \cdots, X_n$ 이 서로 독립이고 동일한 분포를 따르며 $E|X_1| < \infty$ 이면 다음이 성립한다.
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E[X_1]$
- 중심극한정리(Central Limit Theorem; CLT) : 확률변수 $X_1, \cdots, X_n$ 이 서로 독립이고 동일한 분포를 따르며 $Var(X_1) < \infty$ 일 때 $E[X_1] = \mu, Var(X_1) = \sigma^2$ 이라 하면 다음이 성립한다.
- $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z, Z \sim N(0,1)$
- 두 정리 모두 확률벡터의 경우로 자연스럽게 확장이 되며, 특히 다차원의 경우에서의 중심극한정리는 다음과 같이 서술된다:
- k 차원 확률벡터 $X_1, \cdots, X_n$ 가 서로 독립이고 동일한 분포를 따르며 분산행렬이 존재할 때 $E[X_1] = \mu, Var(X_1) = \Sigma$ 라 하면 다음이 성립한다.
- $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} Z, Z \sim N_k(0,\Sigma)$

극한분포

극한분포의 활용

- 대수의 법칙의 경우 몬테카를로(Monte Carlo) 방법과 부트스트랩(bootstrapping), 교차검증(Cross Validation)과 같은 수많은 머신러닝 방법론에 이론적 정당성을 부여해주는 역할을 한다.

- 중심극한정리의 경우 데이터의 수가 충분히 많으나 데이터의 분포를 모를 때, 데이터의 표본평균이 근사적으로 정규분포를 따름을 의미한다. R이나 Python 에 있는 대부분의 가설검정관련 함수들은 실제 데이터의 분포가 무엇인지 모를 경우, 중심극한정리를 이용하여 검정을 수행한다.

- 대수의 법칙, 중심극한정리 이외에도 수많은 극한분포 관련 이론들이 있으며, 이러한 이론들은 예측값의 점근적 성질을 연구하는 데에 주요하게 쓰인다.

통계이론

추정

통계적 추론과 랜덤표본

- 통계적 추론 방법의 성질을 연구할 때에는 흔히 복원추출을 개념화하여 동일한 모집단을 독립적으로 관측하는 것을 전제로 한다.
- 특히, 표본 수가 증가할수록 비복원추출의 경우 또한 복원추출의 경우로 근사적으로 생각할 수 있으며, 현실적으로도 이러한 가정이 유의미하다.
- 이러한 전제 하에, 랜덤표본(random sample)을 서로 독립이고 동일한 분포를 따르는 확률변수라 정의한다.
- 이 때 모형 설정에 사용되는 매개변수를 모수(population parameter)라 하며, 전체 가능한 모수의 집합을 모수공간(parameter)라 한다.
- 형식적으로 다음과 같이 서술할 수 있다: 모수공간을 Ω, 모수를 $\theta \in \Omega$ 라 할 때, 확률분포 $f(x;\theta)$ 에서 추출한 랜덤표본 $X_1, \cdots, X_n$ 은 다음과 같이 표기한다.
- $X_1, \cdots, X_n \overset{iid}{\sim} f(x;\theta), \theta \in \Omega$
- 이 때 통계량이란 미지의 모집단 분포 또는 모수에 대한 추측을 목적으로 사용하는 랜덤표본의 관측 가능한 함수를 의미한다.

추정

적률이용 추정법

- 랜덤표본을 이용하여 모평균 $\mu = E[X_1]$에 관한 추측을 하는 경우에 표본평균 $\bar{X}_n$ 를 이용한다.
- 이와 마찬가지로 r 차 적률 $m_r = E[X_1^r]$에 관한 추측을 할 때에는 $\hat{m}_r = (X_1^r + \cdots + X_n^r)/n$ 을 이용하는 것이 자연스럽다.
- 모집단 적률(moment)의 함수로 표현되는 모수 $\eta = g(m_1, \cdots, m_k)$ 에 관한 추측을 할 때 대응하는 함수 $\hat{\eta} = g(\hat{m}_1, \cdots, \hat{m}_k)$ 를 이용하는 방법을 적률이용법이라 한다.
- 모수 η 의 추측에 사용하는 통계량 $\hat{\eta}$ 를 적률이용추정량(method of moments estimator; MME)라 한다.
- 이러한 추정량은 대수의 법칙과 확률수렴의 성질(참고자료 1,7,8,9,10 참고)에 의해 유의미한 추정 정확도를 보인다.

추정

최대가능도 추정법

- 고정된 관측 결과 $x = (x_1, \cdots, x_n)^T$ 에 대해 가능도함수(likelihood function)는 다음과 같이 정의한다:
- $L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta),\ \theta \in \Omega$
- 가능도함수는 모수가 θ 일 때 주어진 관측결과가 나올 가능성을 나타내준다고 볼 수 있다. 이러한 관점에서 가능도함수가 더 높은 값을 가지는 모수를 찾아내는 것이 좋다. 따라서 관측 결과가 x 일 때 가능도함수가 최대가 되는 모수를 찾는 것이 목적이 될 수 있다.
- 최대가능도 추정량(maximum likelihood estimator : MLE) 의 정의는 다음과 같다:
- $X_1, \cdots, X_n \overset{iid}{\sim} f(x; \theta), \theta \in \Omega$ 일 때, $\hat{\theta}^{MLE} = \arg\max_{\theta \in \Omega} L(\theta, X) = \arg\max_{\theta \in \Omega} \prod_{i=1}^{n} f(X_i, \theta)$
- 일반적으로 최대가능도 추정량을 구하기 위해 미분을 이용하기 때문에, 가능도함수에 로그를 취해 이용하는 경우가 많다. 이를 로그가능도함수(log-likelihood function)이라 하고, 다음과 같이 정의한다:
- $l(\theta; x) = \log L(\theta; x) = \sum_{i=1}^{n} \log f(x_i; \theta),\ \theta \in \Omega$
- 최대가능도 추정법은 머신러닝, 특히 분류 문제나 분포 가정이 있는 경우의 예측 문제의 목적함수로 자주 쓰인다.

추정

최대가능도 추정량의 성질

- 모수 θ 의 최대가능도 추정량이 존재하고 $\eta = (\eta_1, \eta_2)^T = (g_1(\theta), g_2(\theta))^T$ 가 일대일 변환일 때, $\eta_1 = g_1(\theta)$ 의 최대가능도 추정량은 다음과 같다:
- $\hat{\eta}_1{}^{MLE} = g_1(\hat{\theta}^{MLE})$
- 또한, 지수족(exponential family)이라는 확률밀도함수의 형태를 데이터들이 따를 때, 최대가능도 추정량을 쉽게 구하는 방법은 알려져 있다.
- 특히 지수족에서의 최대가능도 추정량의 성질과 이진 분류(binary classification)에서 주로 쓰이는 로지스틱 회귀(logistic regression)는 큰 관련이 있으며, 지수족에서의 최대가능도 추정량의 성질들은 이러한 일반화 선형모형(generalized linear model; GLM)들의 이론적 기반이 된다.
- 지수족의 대표적인 분포들은 다음과 같다.
- 이산형 분포 : 베르누이 분포, 이항분포, 다항분포, 기하분포, 음이항분포, 초기하분포, 포아송분포 등등
- 연속형 분포 : 감마분포, 지수분포, 정규분포, 베타분포, 디리클레 분포, t 분포, F 분포, 카이제곱분포 등등

추정

최대가능도 추정량의 점근적 성질

- 확률밀도함수 $f(x;\theta)$ 와 모수공간 $\Omega \in \mathbb{R}^k$ 가 특정 조건들을 만족할 때, 최대가능도 추정량은 근사적으로 정규분포를 따른다. 이러한 조건들을 정규 조건(regularity conditions)이라 한다.
- 정규 조건을 만족하는 대표적인 분포들로 지수족 분포들이 있다.
- 정규 조건을 따를 때, n 개의 랜덤표본으로 추정된 최대가능도 추정량 $\hat{\theta}_n^{MLE}$ 은 다음과 같은 성질을 만족한다.
- $\sqrt{n}(\hat{\theta}_n^{MLE} - \theta) \xrightarrow{d} N(0,[I(\theta)]^{-1}), \quad I(\theta) = Var(\dot{l}_n(\theta)) = E[-\ddot{l}_n(\theta)]$
- 단, $\dot{l}_n(\theta), \ddot{l}_n(\theta)$ 는 로그가능도함수 $l_n(\theta) = \Sigma_{i=1}^n \log f(X_i;\theta)$ 의 일차, 이차 편도함수.
- 또한, 실수값 모수 $\eta = \eta(\theta)$ 의 추정량 $\hat{\eta}_n = \hat{\eta}(X_1, \cdots, X_n)$ 의 분산에 대하여 다음이 성립한다.
- 정보량 부등식(information inequality) : $Var(\hat{\eta}_n) \geq \left(\frac{\partial}{\partial\theta}E[\hat{\eta}_n]\right)^T [nI(\theta)]^{-1}\left(\frac{\partial}{\partial\theta}E[\hat{\eta}_n]\right), \quad \forall\theta \in \Omega$
- $I(\theta)$ 가 만약 충분히 크다면 최대가능도 추정량의 극한분포의 분산은 0에 가까워질 것이며, 이는 곧 최대가능도 추정량이 실제 모수값과 큰 차이가 없음을 의미한다.
- 또한 정보량 부등식은 어떠한 모수 추정량도 $I(\theta)$ 의 역수만큼에 해당하는 분산을 극복할 수 없음을 의미한다.
- 이러한 의미에서 $I(\theta)$ 를 정보 또는 정보행렬(Fisher's information)이라 부른다.

추정

최소제곱 추정법

- 최소제곱 추정법(least square estimation;LSE)은 최대가능도 추정법과 더불어 수많은 예측 문제에 사용되는 방법으로, 해석의 직관성 및 계산 편의성이 좋아 널리 이용된다.
- 최소제곱 추정법은 어떠한 모형을 가정했을 때, 그 모형으로 생기는 오차제곱합이 최소가 되는 모수를 찾는 것을 목표로 한다.
- 선형회귀모형 $Y = X\beta + \epsilon$ 에 적용시켜보면 다음과 같이 수식화시킬 수 있다:
- $\hat{\beta}^{LSE} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 = (X^T X)^{-1} X^T Y,$ 단, $rank(X) = rank(X^T X) = p$
- 특히 모형 오차항이 정규분포를 따를 경우, 최소제곱 추정법을 이용하는 것은 최대가능도 추정법을 이용하는 것과 근본적으로 같다.
- 최소제곱 추정법은 모형 가정을 하기 어렵거나, 최대가능도 추정과 같은 방법을 계산상의 문제로 이용하기 어려운 신경망 모형(neural network model), 시계열 모형(time series model) 등에서 주로 이용된다.
- 그러나 근본적으로 특이치(outlier)에 대한 비강건성(unrobustness) 문제에서 매우 취약할 수 밖에 없는 구조를 가지고 있으며, 이는 과적합 문제로 직결된다.
- 특히 오차의 등분산성(homoscedasticity)과 정규분포 가정이 만족되지 않으면 최적해가 아닐 가능성이 높다.

검정

검정의 형식논리(1)

- 랜덤한 현상의 모형을 탐색하는 과정에서, 경험이나 지식을 근거로 모형을 설정하고 관측 결과가 설정된 모형과 얼마나 부합하는지 또는 얼마나 괴리가 있는지 살펴볼 것이다.
- 특히 설정된 모형에서는 일어나기 매우 어려운 관측 결과를 얻었다면, 그 설정된 모형을 부정하고 대신할 수 있는 다른 모형을 찾아보게 될 것이다.
- 이러한 논리에 따라 설정된 모형을 채택할 것인지 또는 기각하고 대안을 모색할 것인지를 판단하는 것을 통계적 가설검정(statistical hypothesis testing)이라 한다.
- 일반적으로 관측 결과로부터 뚜렷한 반증이 있을 때에 부정하고자 하는 가설을 귀무가설 (null hypothesis; H_0)이라 하고, 이러한 경우에 대안으로 제시되는 가설을 대립가설 (alternative hypothesis; H_1) 이라 한다.

검정

검정의 형식논리(2)

- 한 편, 검정에서 귀무가설을 기각하거나 채택하는 어느 경우에도 잘못 판단하는 오류가 있다. 특히, 귀무가설이 맞지만 기각하는 경우를 1종 오류(type I error), 반대로 대립가설이 맞지만 귀무가설을 채택하는 경우를 2종 오류(type II error)라 한다.
- 이러한 오류를 범할 확률을 가능한 작게 해주는 것이 바람직할 것이다. 특히 뚜렷한 반증이 있을 때에 기각하고자 하는 가설이 귀무가설이므로 1종 오류를 범할 확률이 미리 지정한 작은 값 이하인 검정을 사용하도록 한다.
- 이러한 경우에 제 1종 오류를 범할 확률의 최대 허용한계로 제시된 값 α 를 유의수준(significance level)이라 하며, 수식적 정의는 다음과 같다:
- $P(\text{기각} \mid H_0) \leq \alpha$
- 이러한 검정을 유의수준 α 의 검정이라 한다.

검정

최대가능도비 검정법(likelihood ratio test)

- 모집단 분포가 확률밀도함수 $f(x;\theta), \theta \in \Omega$ 중 하나라고 모형을 설정한 경우에 이 모집단으로부터의 랜덤표본 $X_1, \cdots, X_n$ 을 이용하여 가설을 검정하는 방법을 알아보자.
- $H_0 : \theta \in \Omega_0 \;\; vs \;\; H_1 : \theta \in \Omega_1$ $(\Omega_0 \cap \Omega_1 = \emptyset, \Omega_0 \cup \Omega_1 = \Omega)$ 로 주어진 경우에 각 가설 하에서 관측 결과 $x = (x_1, \cdots, x_n)^T$ 가 생성되었을 가능성이 가장 큰 경우의 가능도를 비교하자는 것이 검정에서의 최대가능도를 이용하는 방법이다.
- 즉, 가능도함수를 $L(\theta; x)$ 라 하면, 최대가능도비 $\max_{\theta \in \Omega_1} L(\theta; x) / \max_{\theta \in \Omega_0} L(\theta; x)$ 가 크면 대립가설에 대한 증거로서 귀무가설에 대한 반증이 뚜렷하므로 귀무가설을 기각하자는 것이다.
- 이때 위의 가능도비를 이용하는 것은 $\max_{\theta \in \Omega} L(\theta; x) / \max_{\theta \in \Omega_0} L(\theta; x)$ 를 이용하는 것과 동치이므로, 최대가능도비 검정은 일반적으로 전체모수공간과 귀무가설 하의 모수공간에서의 가능도비를 이용한다.
- 따라서 가능도비와 로그 가능도비를 이용한 검정의 기각역(rejection region) 은 다음과 같다:
- $\max_{\theta \in \Omega_1} L(\theta; x) / \max_{\theta \in \Omega_0} L(\theta; x) \geq \lambda \Leftrightarrow 2(l(\hat{\theta}^{MLE}; x) - l(\hat{\theta}^{MLE_0}; x)) \geq c$

검정

최대가능도 검정법의 근사

- 만약 확률모형이 정규 조건을 만족한다면, 가능도비의 극한분포가 존재함이 알려져 있다.
- 특히 단순귀무가설의 경우엔 다음이 성립한다:
- $H_0 : \theta = \theta_0 \;\; vs \;\; H_1 : \theta = \theta_1, \theta \in \Omega \subset \mathbb{R}^k$ 인 가설을 생각하자. 이 때
- H_0 가 사실이면, 다음 성질들이 만족한다.
- $2[l(\hat{\theta}^{MLE}; x) - l(\theta_0; x)] \xrightarrow{d} \chi^2(k)$
- (Wald test) $n(\hat{\theta}^{MLE} - \theta_0)[I(\theta_0)]^{-1}(\hat{\theta}^{MLE} - \theta_0) \xrightarrow{d} \chi^2(k)$
- (Rao test) $\dot{l}(\theta_0)^T[nI(\theta_0)]^{-1}\dot{l}(\theta_0) \xrightarrow{d} \chi^2(k)$
- 따라서 근사적으로 기각역이 $\phi(x, \theta_0) \geq \chi^2_\alpha(k)$ 의 꼴으로 주어진다.

의사결정이론

베이지안 추론과 사전분포, 사후분포

- 지금까지는 모수 θ 가 정해진 수치(deterministic)라 생각했지만, 한 편으로는 어떠한 확률구조로부터 랜덤하게 생성된 값(stochastic)이라고 생각할 수 있다.
- 이러한 관점에서, 모수 θ 에 확률밀도함수 $\pi(\theta)$ 를 부여할 수 있다. 이는 모수에 대한 사전에 가지고 있는 정보라고 생각할 수 있기 때문에 자료를 관측하기 전의 정보로 해석할 수 있다.
- 이는 베이지안 추론(Bayesian inference)의 시발점이지만, 모수에 확률구조를 부여하는 것이 의미가 있는가에 대한 논쟁이 있다.
- 그러나 이러한 해석이 현대통계학과 머신러닝 방법론의 발전에 큰 기여를 한 것은 부정할 수 없으며, 따라서 다양한 관점의 통계적 추론법을 알 필요가 있다.
- 이러한 관점에서, 베이즈 정리를 다음과 같이 다시 서술한다:
- 확률밀도함수가 $f(x \mid \theta)$ 인 모집단에서의 랜덤표본이 $X_1, \cdots, X_n$ 이고 θ 의 확률밀도함수가 $\pi(\theta)$ 이면 x 가 주어진 경우의 θ 의 확률밀도함수는 다음을 만족한다:
- $\pi(\theta \mid x) \propto L(\theta; x)\pi(\theta)$
- 이 때 $\pi(\theta)$ 를 사전분포(prior), $\pi(\theta \mid x)$ 를 정보가 반영되었다는 의미에서 사후분포(posterior)라 부른다.

의사결정이론

베이지안 추론에서의 추정, 검정

- 베이지안 추론에서는 사후분포를 근본으로 하여 추론의 목적에 부합하도록 사후분포를 요약하여 사용한다.
- 베이지안 관점에서의 추정값으로는 사후평균(posterior mean), 사후중앙값(posterior median), 사후최빈값(posterior mode) 등을 사용하며, 특히 사후최빈값은 사후 최대 확률(maximum a posteriori;MAP)이란 이름으로 최대가능도에 대응된다.
- 베이지안 관점에서 가설 $H_0 : \theta \in \Omega_0 \;\; vs \;\; H_1 : \theta \in \Omega_1$ 을 검정할 때에는 각 가설의 사후확률을 비교하여 판단한다. 이를 수식적으로 표현하면 다음과 같다.
- $$\frac{P(\theta \in \Omega_0 \,|\, X = x)}{P(\theta \in \Omega_1 \,|\, X = x)} = \frac{P(\theta \in \Omega_0 \,|\, X = x)}{1 - P(\theta \in \Omega_0 \,|\, X = x)}$$
- 이를 가설 H_0 의 사후승산비(posterior odd ratio)라 부른다. 가능도비 검정과 대응된다고 생각할 수 있다.

의사결정이론

손실함수, 의사결정함수

- 손실함수(loss function)는 어떠한 추정이 잘못되었을 때, 이를 손실로 보는 함수이다.
- 예를 들어 분류 문제의 경우, 잘 분류하면 0, 그렇지 않으면 1인 0-1 loss function을 생각할 수 있다.
- 일반적인 예측 문제의 경우, 예측값과 실제값 차이의 절대값이나 제곱을 생각할 수 있다.
- 따라서 손실함수는 다음과 같이 정의할 수 있다:
- $l : \Omega \times \mathscr{A} \rightarrow \mathbb{R}_+, l = l(\theta, a) \in \mathbb{R}_+$. 단, $\mathscr{A}$ 는 행동공간(action space)으로, 모수에 대한 모든 추정을 나타내주는 공간.
- 한 편, 의사결정함수(decision fuction, decision rule)는 어떤 문제에 대한 우리의 의사결정을 나타내는 함수이다.
- 예를 들어 분류 문제의 경우, 로지스틱 회귀함수를 이용한 함수를 생각할 수 있다.
- 선형회귀 문제의 경우, 최소제곱추정량을 이용한 예측값을 함수로 생각할 수 있다.
- 따라서 의사결정함수는 다음과 같이 정의할 수 있다:
- $\delta : \mathscr{X} \rightarrow \mathscr{A}, \delta = \delta(x) \in \mathscr{A}$. 단, $\mathscr{X}$ 는 관측값들의 범위를 나타내주는 공간.

의사결정이론

위험함수

- 어떠한 의사결정을 할 때, 의사결정이 좋지 않은 결과를 내놓을 가능성을 최대한 줄이는 것이 좋을 것이다. 이러한 관점에서, 위험함수(risk function, risk)는 다음과 같이 정의한다:
- $R(\theta, \delta) = E[l(\theta, \delta(X)) \,|\, \theta]$
- 즉, 모수가 주어져있는 상황에서, 의사결정으로 인한 손실의 기댓값을 위험이라 보는 것이다.
- 위험함수의 대표적인 예시로 평균제곱오차(mean square error; MSE)를 들 수 있다. 즉, $l(\theta, a) = \|\theta - a\|_2^2$ 인 경우이다.
- 한 편, 앞서 살펴본 베이지안 추론에 따라 모수에 확률분포를 부여할 수 있다. 이러한 관점에서 베이즈 위험함수(Bayes risk function, Bayes risk)는 다음과 같이 정의한다:
- $r(\pi, \delta) = E[R(\theta, \delta)]$
- 일반적으로 대부분의 의사결정문제는 위험함수나 베이즈 위험함수를 최소화하는 것을 목적으로 한다.
- 최선의 예측 또는 분류를 찾는 것이 목적인 머신러닝에서도 마찬가지로 이러한 논리 하에서 알고리즘이 만들어진다.

의사결정이론

베이즈규칙과 최소최대규칙

- 베이즈규칙(Bayes rule)은 다음과 같이 정의한다:
- $\delta^{\pi} = \arg\min_{\delta} r(\pi, \delta) \Leftrightarrow \delta^{\pi}(x) = \arg\min_{\delta} E[l(\theta, \delta(X)) \,|\, X = x]$
- 즉, 베이즈 위험함수를 최소화하는 의사결정을 의미한다.
- 한 편, 모수의 분포가정이 없는 경우에 최소최대규칙(minimax rule)은 다음과 같이 정의한다:
- $\delta^{minimax} = \arg\min_{\delta} \max_{\theta \in \Omega} R(\theta, \delta)$
- 즉, 최악의 상황에서의 최선의 의사결정을 의미한다. 이는 게임이론의 관점에서 자연과 의사결정자 간의 게임으로 해석할 수도 있다.
- 이러한 방법은 베이즈 위험함수에도 적용할 수 있다: 만약 집합 Π 가 모수 θ 에 대한 사전분포들을 모은 집합이라 하면, 베이즈 위험함수 관점에서의 최소최대규칙은 다음과 같다:
- $\delta^{minimax} = \arg\min_{\delta} \max_{\pi \in \Pi} r(\pi, \delta)$
- 특정 조건들이 만족되면 분포 가정이 있는 경우와 없는 경우의 최소최대규칙이 같음이 알려져 있다.
- 최소최대규칙은 매우 방어적인 의사결정방법이지만, 만약 최소최대규칙의 정확성이 어느정도 보장된다면 다른 방법보다 훨씬 좋은 결과를 도출할 수 있다.
- 이러한 방법론은 고차원 저표본 문제를 다루거나 강화학습의 영역에서 주로 쓰인다.

고생하셨습니다!

참고자료

- [1] 수리통계학, 김우철(2012), 민영사
- [2] 고급통계적방법론 강의자료, 정성규(2020), 서울대학교 통계적학습이론 연구실
- [3] 통계이론 1 강의자료, 박병욱(2020), 서울대학교 비모수추론 연구실
- [4] Foundations of Machine Learning, M. Mohri, et al.(2018), MIT Press
- [5] The Elements of Statistical Learning, T. Hastie, et al.(2017), Springer
- [6] Linear Models in Statistics, A. C. Rencher, et al.(2008), John Wiley & Sons Inc
- [7] Probability : Theory and Examples, R. Durrett(2019), Cambridge University Press
- [8] Probability and Measure, P. Billingsley(1986), John Wiley & Sons Inc
- [9] Mathematical Statistics : Basic Ideas and Selected Topics, P. J. Bickel, K. A. Doksum(2015), CRC Press
- [10] Theory of Point Estimation, E. L. Lehmann, G. Casella(1998), Springer
- [11] Testing Statistical Hypotheses, E. L. Lehmann, J. P. Romano(2005), Springer