2018년 10월 6일 기계공학과 김현빈

[ML_15] Probability & Information Theory

Probability

- experiment : 모든 관찰의 과정
  - outcomes : experiment의 결과
    - deterministic : 동일한 조건에서 반복하면 항상 동일한 결과 O
    - random : 동일한 조건에서 반복해도 항상 동일한 결과 X
  - sample space : possible outcome들의 집합
  - event space : possible event들의 집합
  - event : sample space들의 부분집합

Axioms of probability

1. $P(A) \geq 0$ 2. $P(S) = 1$ 3. $P(A \cup B) = P(A) + P(B)$ $(P(A \cap B) = \emptyset)$

Properties of probability

1. $P(A^c) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. $P(A) \leq P(B)$ (if $A \subset B$)
4. $P(A) \leq 1$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional Probability : event A given event B

B가 일어났을 때 A의 확률

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

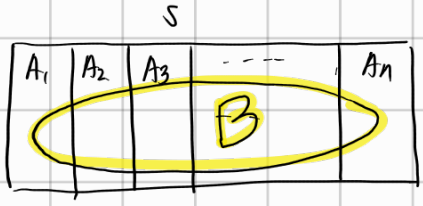
Bayes' Rule

$$\rightarrow P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

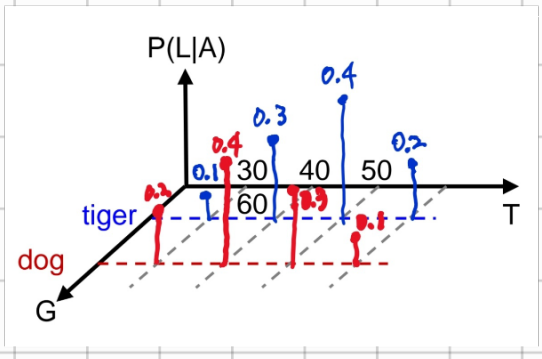
Bayes' Theorem

suppose the events $A_1, A_2, \cdots A_n$: partition of S



$$\rightarrow P(B) = \sum_{\text{모두}}^{n} P(B \cap A_i) = \sum_{\text{모두}}^{n} P(B|A_i)P(A_i)$$

Example)



L=50 일때 dog일지 tiger일지

ML test

$\rightarrow P(L=50|dog) = 0.3 < P(L=50|tiger) = 0.4$

↳ dog 중에서 L=50일 확률 ↳ tiger 중에서 L=50일 확률

MAP test

$\rightarrow P(L=50|dog)P(dog) = 0.24 > P(L=50|tiger)P(tiger) = 0.08$

⇒ 즉, tiger와 dog 각각에서 L=50일 확률은 tiger가 높지만,

tiger와 dog의 개체수에 대한 확률이 적용되면

L=50일때 전체에서 tiger와 dog일 확률은 dog가 더 높음.

ML classification

$\rightarrow k^* = \arg\max P(1_{new}|C_k)$

MAP classification

$\rightarrow k^* = \arg\max P(1_{new}|C_k)P(C_k)$

Independence

→ $P(A \cap B) = P(A)\,P(B)$

$P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$

Independent rv

→ $P(X=x_i, Y=y) = P(X=x)\,P(Y=y)$

Discrete : $P_{XY}(x_i, y_j) = P_X(x_i)\,P_Y(y_j)$

Continuous : $f_{XY}(x,y) = f_X(x)\,f_Y(y)$

Naive Bayes classifiers

$$P(\underline{x}|C_k) = \prod_{d=1}^{D} P(x_d|C_k)$$

↳ 각 확률변수 (x)들의 발생확률은 통상 독립을 가정한다.

But 확률변수를 가정하는 표본들의 독립은 보장할 수 없다.

→ Naive Bayes는 그 한계들로 독립이라고 가정한 것.

Expectations

$$\mu_X = E(X) = \begin{cases} \sum_k x_k P_X(x_k) \\ \int x f_X(x) \end{cases}$$

$$\sigma_x^2 = Var(X) = E[(X-E(X))^2]$$

$$= \begin{cases} \sum (x-\mu_x)^2 P_X(x) \\ \int (x-\mu_x)^2 f_X(x) \end{cases}$$

$$Var(X) = E(X^2) - (E(X))^2$$

Conditional Expectation

$$E(Y|x) = \begin{cases} \sum_k y_k P(y_k|x) \\ \int y f(y|x)\,dy \end{cases}$$

$$E(g(x)) = \begin{cases} \sum g(x_k) P(x_k) \\ \int y f(y)\,dy \end{cases}$$

Correlation & Covariance → Independent ⇄ uncorrelated

- Correlation : $E(XY)$

[orthogonal : $E(XY) = 0$

[uncorrelated : $E(XY) = E(X)E(Y)$

- Covariance : $Cov(X,Y) = \sigma_{XY} = E[(X-E(X))(Y-E(Y))]$

공분산 함수명

$= E(XY) - E(X)E(Y)$

uncorrelated $\sigma_{XY} = 0$

- Correlation coefficient : a normalized covariance

→ $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ $\quad |\rho_{XY}| \le 1$

- correlation coefficient & Linear Dependence

let $Y = aX+b$

⇒ ① $E(XY) = E(X(aX+b)) = aE(X^2) + bE(X)$

$E(Y) = E(aX+b) = aE(X) + b$, $Var[Y] = a^2 Var(X)$

$\sigma_Y = |a|\sigma_X$

$cov(X,Y) = \sigma_{XY} = E(XY) - E(X)E(Y)$

$= aE(X^2) + bE(X) - a(E(X))^2 - bE(X)$

$= a\{E(X^2) - (E(X))^2\} = a\sigma_X^2$

② $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{a\sigma_X^2}{\sigma_X |a| \sigma_X} = \begin{cases} 1 & a>0 \\ -1 & a<0 \end{cases}$

Gaussian Distribution

$$f_X(\underline{x}) = \frac{1}{(2\pi)^{n/2} |\det K|^{1/2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})^T K^{-1} (\underline{x}-\underline{\mu})\right]$$

$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ $\quad \underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix}$ $\quad K = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$

$\sigma_{ij} = cov(X_i, X_j)$

Information Theory

Information

→ 정보이론의 기본원리 : 확률의 작은수록 (불확실성이 큰수록) 많은 정보

자기정보 : 사건 x_i의 정보량 → 불확실성을 계량화.

$$I(x_i) = \log_b \frac{1}{P(x_i)} = -\log_b P(x_i)$$

property : ① $I(x_i) = 0$ for $P(x_i) = 1$

② $I(x_i) \geq 0$

③ $I(x_i) > I(x_j)$ if $P(x_i) < P(x_j)$

④ $I(x_i, x_j) = I(x_i) + I(x_j)$ if x_i and x_j are Independent

Entropy (평균정보량)

$$H(X) = E[I(x_i)] = \sum_{i=1}^{m} P(x_i) I(x_i)$$

$$= -\sum_{i=1}^{m} P(x_i) \log_2 P(x_i)$$

Property : $0 \leq H(X) \leq \log_2 m$

⇒ $H(X)$는 각 symbol이 일어날 확률이 동일할때 최대

각 symbol이 일어날 확률이 동일하다는 것은 어느 한쪽의 확률이 크게보다 불확실성이 크다는 것

1) 동전 앞 : 70%, 뒤 : 30%

2) 동전 앞 : 50% 뒤 : 50%

동전을 던졌을때 앞이나올지 뒤가 나올지 더 예측불가능한건 2번

⇒ 즉 확률이 동일할때 불확실성↑ ⇒ $I(X_i)$↑ ⇒ $H(X)$↑

Source Coding Theorem

- Average code length

$$L = \sum_{i=1}^{m} P(x_i) \underbrace{n_i}_{\text{i번째 symbol의 길이}}$$

$$L \geq H(X) = -\sum_{i=1}^{m} P(x_i) \log_2 P(x_i)$$

Cross-Entropy, KL Divergence

$P(x)$: 실제확률분포, $Q(x)$: 추정확률분포

Pattern 을 위한 교차엔트로피

$$H(P,Q) = E_P[I_Q(x)] = E_P[-\log(Q(x))]$$

$$= -\sum_{i=1}^{m} P(x_i) \log(Q(x_i))$$

Q에서 P의 KL Divergence (상대엔트로피)

$$D_{KL}(P||Q) = \sum_{i=1}^{m} P(x_i) \log\left(\frac{P(x_i)}{Q(x_i)}\right) = H(P,Q) - H(P,P) \geq 0$$

$$\Rightarrow H(P,Q) = H(P,P) + D_{KL}(P||Q)$$

Cross Entropy Loss

$y \in \{0,1\}$: true
$h(x)$: predicted

$$\rightarrow H(y_i, h(x_i)) = -\sum_{i=1}^{k} y_i \log(h(x_i))$$