나이브 베이즈 분류 모형 (Naive Bayes Classification Model)

- Logistic Regression ⇒ $P(y|x) = f(x)$
- QDA | LDA ⇒ $P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(x|y)P(y)}{P(x)}$

- QDA

⇓

$y=0 \rightarrow N(\mu_0, \Sigma_0)$

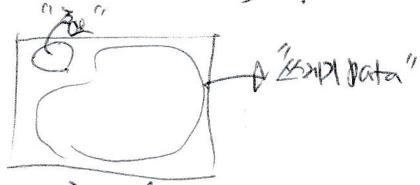
$y=1 \rightarrow N(\mu_1, \Sigma_1)$ $\quad Q(D^2)$

⇓

- LDA 가정은

$\underbrace{\Sigma_0 = \Sigma_1 = \cdots = \Sigma_N}_{(\text{차원수})^2 \text{ 만큼 계산}}$ ⇒ 계산량이 많아지면, 중요한 Data의 갯수보다 필요없는 Data 갯수가 많아짐.

따라서 나이브 가정을 이용함



나이브 가정 ⇒ "클래스 하나만 정해놓고 독립이라고 가정"

→ "모든 차원의 개별 독립변수 요소들이 서로 조건부 독립이라는 가정"

- 이러한 "나이브 가정"을 베이즈 분류 모형에 적용한 모형을 "나이브 베이즈 분류 모형"
  - x 벡터의 결합확률이, 개별원소 x_d의 확률의 곱으로 나타난다.
  - $P(x_1, \dots, x_D | y=C_k) = \prod_{d=1}^{D} P(x_d | y=C_k)$
  - $P(y=C_k | x) \underbrace{\propto}_{\text{비례한다}} \prod_{d=1}^{D} P(x_d | y=C_k) P(y=C_k)$

① 가우시안 정규분포 Likelihood 모형

- x 벡터의 원소가 모두 실수이고, 클래스마다 특정한 값 주변에서 발생한다고 하면, 사용한다.
(독립변수들이 서로 조건부 독립이라고 가정한다)

- $P(x_d | y=C_k) = \frac{1}{\sqrt{2\pi\sigma^2_{d,k}}} \exp\left(-\frac{(x_d - \mu_{d,k})^2}{2\sigma^2_{d,k}}\right)$