5강 (GDA
Generative vs Discriminative.
Naive Bayes

→ Logistic은 x x x
Regression o o 선을 update 해가면서 경계를 찾는 것.

이런 분리와는 다른 접근방식 Generative learning Algorithms.

각각의 class에 대해 model
구축한 후 각 model과 비교

Discriminative	Generative
learn $p(y \mid x)$	learn $p(x \mid y)$
(or learn $h_\theta(x) = \begin{matrix}0\\1\end{matrix}$	feature class
$x \rightarrow y$	$p(y)$: class prior.

Bayes Rule

이렇게하여 계산가능 (환자 1명이 종양이 양성일 확률은 !)

$$P(y=1 \mid x) = \frac{P(x \mid y=1)\,P(y=1)}{P(x)}$$

$$= P(x \mid y=1)P(y=1) + P(x \mid y=0)P(y=0)$$

Generative Learning Algorithm.

Gaussian Discriminant Analysis.

Suppose $x \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

Assume $p(x \mid y)$ is Gaussian

* 다변량 가우시안 분포

$z \sim N(\vec{\mu}, \Sigma)$

parameter : $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$\vec{\mu}$: 중심의 위치를 결정

$\begin{bmatrix}1\\0\end{bmatrix}$

Σ : 분산 정도를 결정

$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

* $\vec{\mu}, \Sigma$ 바꾸어가며 그림 그려봄.

GDA model. $p(x|y)$를 계산해야함!

$p(x|y=0) = \text{Gaussian}$, $p(x|y=1) = \text{Gaussian}$

일반적으로 두개의 class에 같은 Σ 사용하며 파라미터 μ_0, μ_1

$p(y) = \text{Bernoulli} = \phi^{y}(1-\phi)^{1-y}$ $(p(y=1) = \phi)$

total Parameters : $\underset{\mathbb{R}^n}{\mu_0, \mu_1}, \underset{\mathbb{R}^{n\times n}}{\Sigma}, \underset{\mathbb{R}}{\phi}$

이제 Parameter 찾는 것에 대해서 알아보자

* 이를 위해서 Max Joint likelihood

$$L(\phi, \mu_0, \mu_1, \Sigma) = \prod_{i=1}^{m} p(x^{(i)} | y^{(i)})\, p(y^{(i)})$$

→ 최대화하려는 비용함수가 X, Y의 p인 joint likelihood라는 것이 Discriminative Conditional likelihood와 차이.

→ likelihood를 Max 하는 $\phi = \dfrac{\sum_{i=1}^{m} y^{(i)}}{m}$

indicator function $= \dfrac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{m}$ $1\{\text{true}\}=1$

$1\{\text{false}\}=0$

$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\}\, x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}}$$

← y=0 인 vector들 합

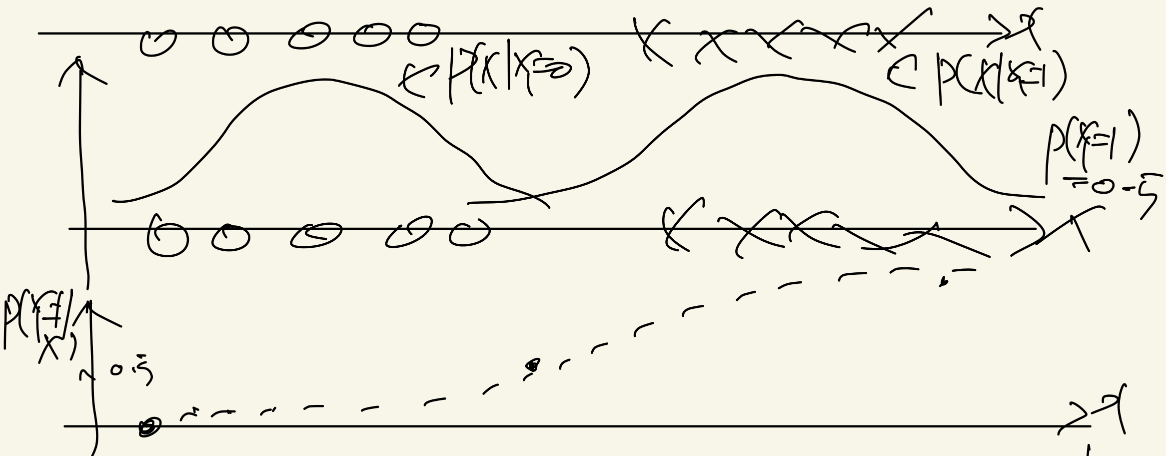
← y=0 인 개수

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^T$$

이제 parameter 얻었으니까 prediction 해야 함.

$$\arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)\,p(y)}{p(x)} \leftarrow \text{상수} \qquad = \arg\max_y p(x|y)\,p(y)$$

Compare GDA with Logistic Regression



→ 그림에서 확인할 수 있는 것처럼 $p(y=1|x;\phi,\Sigma,\mu_0,\mu_1) = \frac{1}{1+\exp(-\theta^T x)}$

GDA 모델에서 y=1 인 것은 $P(y=1|x)$를 modeling 하는 Logistic 함수 형식임.

→ 2차의 함수 모두 sigmoid 함수로 귀결됨.

$$\text{GDA} \xrightarrow{\ O\ } p(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$$
$$\xleftarrow{\ X\ }$$

따라서 GDA는 더 강한 모델링 가정을 가지므로 일반적으로 적은 data set에 더 적합함.

Naive Bayes

GDA는 연속적인 x, Naive Bayes (Generative Learning Algorithm)

ex) 이메일 스팸분류 $n = 10{,}000$

Feature vector x? / 각 단어가 등장하면: 1, 등장X: 0

$$x \in \{0,1\}^n$$

$x_i = 1\{\text{word } i \text{ appears in email}\}$

want to model $p(x|y)$ ← 이대로 하면 파라미터 수가 2^n으로 너무 많아서 불가능

따라서

Assume x_i's are conditionally independent given y

$$= p(x_1, x_2, \ldots, x_{10000}|y) = p(x_1|y)\,p(x_2|x_1, y)\,p(x_3|x_1, x_2, y) \cdots = \prod_{i=1}^{n} p(x_i|y)$$