Introduction

lecture 13

Note $P_\theta(X,Y) = P(X,Y;\theta)$

$(x_i, y_i) \sim P(X,Y)$ — assume it can access

② $P(X,Y;\theta) \leftarrow$ learn $- \forall_x \forall_y P(X=x \wedge Y=y)$

classifier $h(x_{test}) = \underset{y}{\text{argmax}}\ P(Y=y \mid X=x_{test}; \theta)$ — Bayes

Need to learn $P(y|X)$ **

① optimal classifier $h(x_t) = \underset{y}{\text{argmax}}\ P(Y=y \mid X=x_t)$; pobs ที่ $X=x_t$ แล้ว Y เป็นของคลาส y (เช่น salmon, mackarel)

pobs ของ salmon เมื่อ $X=x_{test}$

pobs ของ mackel เมื่อ $X=x_{test}$

Note: คือ (x_i, y_i) ใน pobability distribution X, Y ($P(X,Y)$) ถ้าอยากทำนาง X_{test} ต้องหาจาก $P(x,y)$ ซึ่งไม่สามารถเข้าถึงได้ จึงต้องสร้าง modeling distribution ม. $P(x,y;\theta)$

③ $P(y|X)$

$P(Y=y \mid X=\vec{x})$

$\vec{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$ $X = \begin{bmatrix} [X]_1 \\ [X]_2 \\ \vdots \\ [X]_d \end{bmatrix}$ r.v ที่อยู่ใน d=1

$P(Y=y \mid X=\vec{x}) = P(Y=y \mid [X]_1 = x_1, [X]_2 = x_2, \ldots, [X]_d = x_d)$ way 1

Bayes rule: $P(Y=y \mid X=x) = \dfrac{P(X=x \mid Y=y) \cdot P(Y=y)}{P(X=x)}$ way 2 → Navie Bayes use.

estimate $P(X=x \mid Y=y)$

: $P([X]_1 = x_1, [X]_2 = x_2, \ldots, [X]_d = x_d \mid Y=y)$

Navie Bayes Assumption

lecture 14

: Assumes all feature values are independent give the label.

$P(X=x \mid Y=y) : \prod_{i=1}^{d} P([X]_i = x_i \mid Y=y)$

- เลือกการเรียน Bayes classifier

Bayes Classifier

$h(x) = \underset{y}{\text{argmax}}\ P_\theta(Y=y \mid X=x)$

Goal: Estimate $P(Y|X)$ estimate $\forall_x \forall_y P(Y=y \mid X=x)$

Chain Rule: $P(Y=y \mid X=x) = \dfrac{P(Y=y \wedge X=x)}{P(X=x)}$

Job: (I) Estimate $\forall_x \forall_y P(Y=y \wedge X=x) \approx P(X,Y)$

(II) Estimate $P(X=x)\ \forall_x$ $P(X)$

Sinario use $P_\theta(Y=y \wedge X=x) = Bin(n, \theta)$??

Sinario use $P_\theta(Y=y \wedge X=x) = Bin(n,\theta)$ Way:1

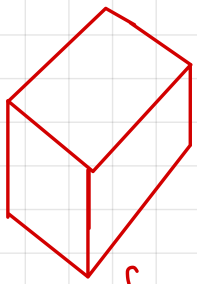
Coin tossing : n times

n_H : ตัวแปรสุ่มที่โยนได้ head.

$$n_H \sim Bin(n,\theta)$$

$$MLE = \theta = \frac{n_H}{n}$$

Dice n_3 : number of times that we obtian face 3



i x j faces

X ที่เป็นไปได้ $X \in \{1,2,\dots,i\}$

$Y \in \{1,2,\dots,j\}$

→ โยนลูกเต๋าที่หน้าออกเป็น y,x ใน Y,X

$I(E)$: indicator random varible

Assign ตัวแปรสุ่มที่มีค่า 1,0

$$I(E) = \begin{cases} 1 & \text{if } E \text{ occor} \\ 0 & \text{other wise} \end{cases}$$

(I)

$$MLE\ \theta = \frac{n_{y,x}}{n} = \frac{\sum_{i=1}^{n} I(Y_i=y \wedge X_i=x)}{n}$$

ใช้แทน ↓ corresponse

$P_\theta(Y=y \wedge X=x)$

$$P_\gamma(X=x) = Bin(n,\gamma)$$

(II)

$$MLE \Rightarrow \gamma = \frac{\sum_{i=1}^{n} I(X_i=x)}{n}$$

Assume X, Y follow Binominal distribution

$$MLE: P_\theta(Y=y \mid X=x) = \frac{P_\theta(Y=y \wedge X=x)}{P_\gamma(X=x)}$$

$$= \frac{\dfrac{\sum_{i=1}^{n} I(Y_i=y \wedge X_i=x)}{\cancel{n}}}{\dfrac{\sum_{i=1}^{n} I(X_i=x)}{\cancel{n}}}$$

$$= \frac{\sum_{i=1}^{n} I(Y_i=y \wedge X_i=x)}{\sum_{i=1}^{n} I(X_i=x)}$$

Problem with $P_\theta(Y=y \mid X=x)$

$$P_\theta(Y=y \mid X=x) = \frac{\sum_{i=1}^{n} I(Y_i = y \wedge X_i = x)}{\sum_{i=1}^{n} I(X_i = x)}$$

- X เล็กที่หลายdimension => vector (X ตัวนี้ถูกเขียนเป็น vector)

$$P_\theta(Y=y \mid \vec{X}=\vec{x}) \Rightarrow P(Y=y \mid [X]_1 = x^1 \wedge [X]_2 = x^2 \wedge \ldots \wedge [X]_d = x^d)$$

$$\vec{X} = \begin{bmatrix} [X]_1 \\ [X]_2 \\ \vdots \\ [X]_d \end{bmatrix} \quad \vec{x} = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^d \end{bmatrix}$$

$[X]_d$ ↓ r.v. ; x^1 ↑ fixed

r.v X = fixed X on each d

$$\frac{\sum_{i=1}^{n} I(Y_i = y \wedge [X_i]_1 = x^1 \wedge [X_i]_2 = x^2 \wedge \ldots \wedge [X_i]_d = x^d)}{\sum_{i=1}^{n} I([X_i]_1 = x^1 \wedge \ldots \wedge [X_i]_d = x^d)}$$

Note: When $d \gg 0$ and $n \to +\infty$

$$\Rightarrow P_\theta(Y=y \wedge X=x) = \frac{1}{n} = 0$$

$$\Rightarrow P_\theta(X=x) = \frac{1}{n} = 0$$

So: $P_\theta(Y=y \mid X=x) = \frac{0}{0}$ undifind

Apply Bayes' rule to Bayes Classifer ($P_\theta(Y=y \mid X=x)$) way: 2

Bayes' rule: $P(Y=y \mid X=x) = \frac{P(X=x \mid Y=y)P(Y=y)}{P(X=x)}$

Bayes classifier

$$= h(x) = \underset{y}{\operatorname{argmax}}\ P(Y=y \mid X=x)$$

$$= \underset{y}{\operatorname{argmax}}\ \frac{P(X=x \mid Y=y) \cdot P(Y=y)}{\cancel{P(X=x)}} \quad \text{ไม่เกี่ยวกับ } y$$

$$= \underset{y}{\operatorname{argmax}}\ P(X=x \mid Y=y) \cdot P(Y=y)$$

Estimate $P(Y=y)$ → binomial (binary, multiclass classification)

$$P_\theta(Y=y) = \sum_{i=1}^{n} \frac{I(Y_i=y)}{n}$$

- Estimate $P(\vec{X}=\vec{x} \mid Y=y)$

$= P([\vec{X}]_1 = \vec{x}^1 \wedge [\vec{X}]_2 = \vec{x}_2 \wedge \ldots \wedge [\vec{X}]_d = \vec{x}_d \mid Y=y)$ Can't estimate direct

So use Naive Bayes assumption.

Assume X, Y follow Binomial distributions

$$\Rightarrow P_\theta(Y=y) = \frac{\sum_{i=1}^{n} I(Y_i=y)}{n}$$

$$\Rightarrow P_\theta(X=x \mid Y=y) = ?$$

Naive Bayes assumption:

- All feature values are Independent

$$p(\vec{X}=\vec{x} \mid Y=y) = \prod_{i=1}^{d} P([X]_i = x^i \mid Y=y)$$; Probs ที่ Y given y ที่ a=ได้ $[X]_i = x^i$ a=อิสระต่อกัน

Naive Bayes classifier

$$h(X) = \underset{y}{\operatorname{argmax}} \prod_{\alpha=1}^{d} P([\vec{X}]_\alpha = \vec{x}^\alpha \mid Y=y)\, P(Y=y)$$

How to estimate $P([\vec{x}]_\alpha \mid Y)$ 3 cases

- There are 3 notable cases:

Case1: Categorical features : Categorical Naive Bayes Classifier

$$[\vec{x}]_\alpha \in \{c_1, c_2, \dots, c_k\}$$

eg: {male, female}

: {single, widowed, married}

We model $P([\vec{X}]_\alpha = j \mid Y=y) = [\theta_{jy}]_\alpha$ parameter ที่บอกว่าค่า j ของ coordinate α

↑ The probability of feature α having value j given the label is y

MLE estimate $\Rightarrow [\theta_{jy}]_\alpha$ = # of sample with label y that has feature α with value j / # of samples with label y.

$$\Rightarrow \frac{\sum_{i=1}^{n} I(Y_i=y) \cdot I(x_i^\alpha = j)}{\sum_{i=1}^{n} I(y_i=y)}$$: count เมื่อ $Y_i=y$ และ $X_i=j$ ที่ α

Quiz 4

The following table is a result from observing the behavior of a person whether he went out or stayed home given the two weather conditions (sunny or rainy) and the two options regarding his car status (car-broken or car-working)

- $y_i \in \{go-out, stayhome\}$
- $x_i^1 \in \{sunny, rainy\}$
- $x_i^2 \in \{car-broken, car-working\}$

Categorical feature

Estimate $P([\vec{x}]_1 \mid y)$ → $P([\vec{x}]_1 = sunny \mid y = go\text{-}out)$, → $P([\vec{x}]_1 = rainy \mid y = go\text{-}out)$, → $P([\vec{x}]_1 = sunny \mid y = stay)$, → $P([\vec{x}]_1 = rainy \mid y = stay)$

$P([\vec{x}]_1 = sunny \mid y = go\text{-}out) = [\theta_{sunny, go\,out}]_1 = \frac{4}{5} = 0.8$

i	x_i^1	x_i^2	y_i
1	sunny	car-broken	go-out
2	rainy	car-working	go-out
3	sunny	car-broken	go-out
4	sunny	car-broken	go-out
5	sunny	car-broken	go-out
6	sunny	car-working	stay home
7	rainy	car-working	stay home
8	rainy	car-broken	stay home
9	sunny	car-working	stay home
10	rainy	car-working	stay home

5

Assume that we are using Binomial distribution as the modeling distribution. You are to demonstrate solutions to the following questions.

1. Estimate P(y=go-out).
2. Estimate P(y=stay home).
3. What is the estimate of P(y)?
4. What is the estimate of P(x)?
5. Estimate P(x = (rainy, car-working) and y=go-out).
6. Estimate P(y=go-out | x = (rainy, car-working)) directly.
7. Estimate P(x = (rainy, car-working) | y=go-out) using Naive Bayes assumption.
8. By using Naive Bayes assumption, what would be the return of h(x = (sunny, car-broken))?

Case 2 : Multinomial feature : Multinomial Naive Bayes Classifier

eg. text data : "An ant is animal." —Bag of word→ $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_d \end{bmatrix}$ ← # of word 1 appearing in the text; x_j : its count of word in sentense. (d = cordinate)

- we estimate $P([\vec{x}]_\alpha = j \mid y = y)$ by using multinomial distribution

eg. Span filter

- $y \in \{ \underset{+1}{\text{span}}, \underset{-1}{\text{ham}} \}$
- $\vec{x}$ represent text data (B.O.W)

$$\vec{x} = \begin{bmatrix} x^1 \\ \vdots \\ i \\ \vdots \\ x^d \end{bmatrix} \begin{matrix} - w_1 \\ \\ - w_2 \\ \\ - w_d \end{matrix}$$

Estimate $P([\vec{x}]_\alpha = j \mid y = y)$; $\alpha = 5, j = 10, y = \text{spam}$ Given w_5 = Princess (เกิด 100ครั้ง)

$P([\vec{x}]_5 = 10 \mid y = \text{spAM})$; given spam ที่ตัวแปรสุ่ม cordinate 5 = 10

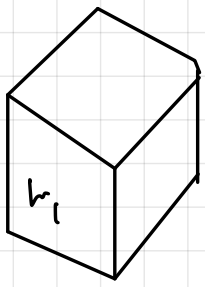
Modeling

$$P([\vec{x}]_\alpha = j \mid y = y) = \left[\binom{m}{j} \cdot P(w_\alpha \mid y = y) \right]^j$$

eg. $P[[\vec{x}]_{\underset{\text{word ที่ 2}}{2}} = 3 \mid y = \text{spam}] = \left[\binom{5}{3} \times P(w_2 \mid y = \text{spam}) \right]^3$

- $P(w_2 \mid y = y)$ is the prob. of selecting word w_α given the label is y.
- m is the number of words in total ($m = \sum_{\alpha=1}^{n} [\vec{x}]_\alpha$)

Multinomial distribution



d faces
(y = spam)

$$x = \begin{pmatrix} 1 & x^1 \\ 4 & x^2 \\ \vdots & \vdots \\ & x^d \end{pmatrix} \begin{matrix} - w_1 \\ - w_2 \\ \\ - w_d \end{matrix}$$

Assume throw and get R1:

$w_2, w_2, w_1, w_2, w_2, w_3, w_2$

m times = 7

R2:

$w_1, w_2, w_2, w_3, w_2, w_2, w_2$

n times = 7

R_1, R_2 not a same words but with power of Back of words it give same feature Vector. (Corder of word is useless)

- We model $P(w_\alpha | Y = spam) = [\theta_{spam}]_\alpha$

$P(w_\alpha | Y = ham) = [\theta_{ham}]_\alpha$

- Estimate

$$- MLE: [\theta_{spam}]_\alpha = \frac{\sum_{i=1}^{n} I(y_i = SPAM) \cdot x_i^\alpha}{\sum_{i=1}^{n} I(y_i = SPAM) \cdot (\sum_{d=1}^{\alpha} x_i^d)} = \frac{\text{จำนวนครั้งที่ } w_\alpha \text{ ปรากฏใน sample ที่เป็น spam}}{\text{จำนวนของ words รวมใน sample ที่เป็น spam}}$$

$[\theta_y]_\alpha \; \forall y \forall \alpha$

eg. $[\theta_{spam}]_{money} = P(money | y = spam)$

$$P([\vec{x}]_{money} = 2 \mid y = spam) = \binom{m}{2} \cdot [P(money | y = spam)]^2$$

∴ ใน mail มีคำว่า money 2 ครั้ง
เราไม่รู้อยู่ตรงไหนบ้าง เลยใช้ $\binom{m}{2}$ ด้วย

Summary

Bayes Rule → $h(x) = \arg\max_y P(X = x | Y = y) \cdot P(y)$

Naive Bayes classifier → $= \arg\max_y \prod_{\alpha=1}^{d} P([x]_\alpha = x^\alpha | Y = y) \cdot P(y)$

Spam filter (text classification)

"An ant is an animal" ; 5 words

Bag of word $\rightarrow \vec{x} = \begin{bmatrix} 0 \\ 2 \\ \vdots \end{bmatrix}$ $w_1 = a$, $w_2 = an$, $\vdots$, $w_d =$

eg: $P([X]_2 = 2 \mid y = spam) = \binom{5}{2}[P(w_2 \mid y = spam)]$

Estimate

$$P(w_2 \mid y = spam) = [\theta_{spam}]_2$$

MLE

$$\Rightarrow [\theta_{spam}]_2 = \frac{\sum_{i=1}^{n} I(y_i = spam) \cdot x_i^2}{\sum_{i=1}^{n} I(y_i = spam)\left(\sum_{b=1}^{d} x_i^b\right)}$$

In general

$$P([X]_\alpha = j \mid y = spam) = \binom{m}{j}\left(P(w_\alpha \mid y = spam)\right)^j$$

$$[\theta_y]_\alpha = \frac{\sum_{i=1}^{n} I(y_i = y) \cdot x_i^\alpha}{\sum_{i=1}^{n} I(y_i = y)\left(\sum_{b=1}^{d} x_i^b\right)}$$

→ จน. ครั้งแต่ละ word ปรากฏอยู่ใน y

→ จน. ของ word ทั้งหมดที่อยู่ใน y

$$h(x) = \arg\max_{y} \boxed{\prod_{\alpha=1}^{d} P([x]_\alpha = x^\alpha \mid y=y)} \cdot P(y)$$

Note: $m = \sum_{\alpha=1}^{d} x^\alpha$

$$= P([x]_1 = x^1 \mid y=y) \times \dots \times P([x]_\alpha = x^\alpha \mid y=y)$$

$$= \left(\binom{m}{x^1} \times \left([\theta_y]_1\right)^{x^1}\right) \cdot \left(\binom{m-x^1}{x^1} \times \left([\theta_y]_2\right)^{x^2}\right)$$

$$\times \dots \times \left(\binom{m-x_1-\dots-x^{d-1}}{x^d} \left([\theta_y]_d\right)^{x^d}\right)$$

$$\Rightarrow \left(\frac{m!}{(m-x^1)!\,x^1!} \times \frac{(m-x^1)!}{(m-x^1-x^2)!\,x^2!} \times \dots \times \frac{(m-x^1-\dots-x^{d-1})!}{(m-x^1-\dots-x^b)!\,x^d!}\right)$$

$$\left(\prod_{\alpha=1}^{d} \left([\theta_y]_\alpha\right)^{x^\alpha}\right)$$

$$m = \sum_{i=1}^{b} x^i$$

$$\Omega!$$

$$\prod_{\alpha=1}^{d} P([X]_\alpha = x^\alpha \mid y = y) = \frac{m!}{x^1! x^2! \ldots x^\alpha!} \prod_{\alpha=1}^{d} \left([\theta_y]_\alpha\right)^{x^\alpha}$$

Binary Classification

$$\frac{P(y=\text{spam}) \times \prod_{\alpha=1}^{d} P([X]_\alpha = x^\alpha \mid y=\text{spam})}{P(y=\text{ham}) \times \prod_{\alpha=1}^{d} P([X]_\alpha = x^\alpha \mid y=\text{ham})} = \frac{\frac{m!}{x^1! x^2! \ldots x^d!} \prod_{\alpha=1}^{d} \left([\theta_{\text{spam}}]_\alpha\right)^{x^\alpha} \cdot P(y=\text{spam})}{\frac{m!}{x^1! x^2! \ldots x^d!} \prod_{\alpha=1}^{d} \left([\theta_{\text{ham}}]_\alpha\right)^{x^\alpha} P(y=\text{ham})}$$

take $\log_e$ into both sides

$$\frac{P(y=\text{spam}) \times \prod_{\alpha=1}^{d} P([X]_\alpha = x^\alpha \mid y=\text{spam})}{P(y=\text{ham}) \times \prod_{\alpha=1}^{d} P([X]_\alpha = x^\alpha \mid y=\text{ham})} \propto \frac{\left(\sum_{\alpha=1}^{d} x^\alpha \log_e\left([\theta_{\text{spam}}]_\alpha\right)\right) + \log_e(P(y=\text{spam}))}{\left(\sum_{\alpha=1}^{d} x^\alpha \log_e\left([\theta_{\text{ham}}]_\alpha\right)\right) + \log_e(P(y=\text{ham}))}$$

$P([X]_\alpha \mid y)$

3 notable class

① Categorical features ($[X]_\alpha$ เป็นประเภท(แทน word) ไม่ใช่ จำนวนครั้งของการปรากฏ) → Categorical Naive Bayes classifier

$[X]_\alpha \in \{0, 1, \ldots, k-1\}$

② multinomial feature ($[X]_\alpha$ แทนความถี่ของการปรากฏ ของแต่ละ word)

$[X]_\alpha \in \{0, 1, \ldots, m\}$ $m = \sum_{\alpha=1}^{d} x^\alpha$ เป็นความถี่ (จน.ครั้งของการปรากฏ)

↳ multinomial Naive Bayes classifier

③ Continous feature

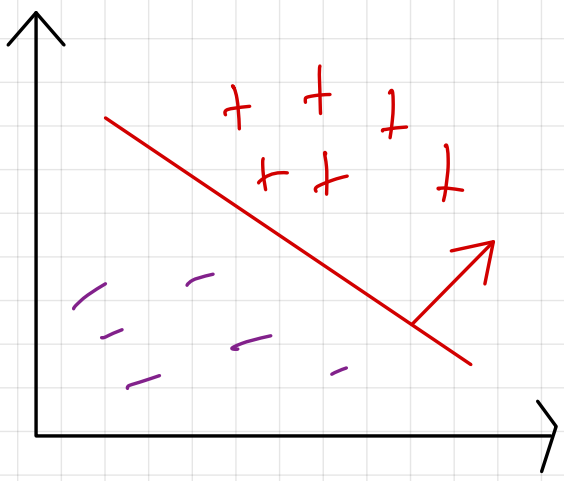
$[X]_\alpha \in \mathbb{R}$ ($[X]_\alpha$ เป็นค่าต่อเนื่อง)

↳ Guassian Naive Bayes Classifier

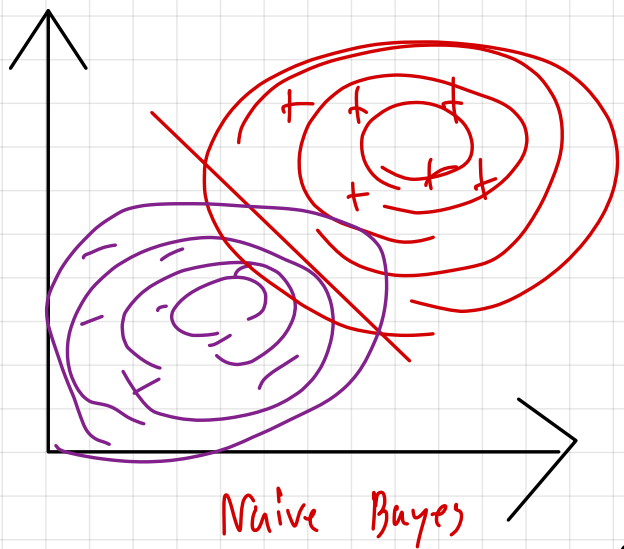
Summary of Naive Bayes

- Naive Bayes = Bayes classifier + Bayes rule + Naive Bayes assumption
- The assumption says "all feature values are Indepedent."
  - ที่จากรูป W_α เมื่อให้ label $y = spam$ $P(W_\alpha | y = spam)$
- We may have data that violates the assumption.
- If our data follows multinomial distributions (features) and our task is binary classification, the Naive Bayes gives the

linear decision boundary



Perceptron



Naive Bayes

- หาง distribution ของ probs. ใน sample

- hyperplane separate ของ data
- find $\vec{w}$ that linearly separate the data

- separate ของ distribution
- find $\vec{w}$ that separate the trained distributions

ในรูปจะได้ x

Discriminative learning : Try to model $P(y|x)$ (eg. k-NN, Perceptron)

Generative learning : try to model $P(x|y)$ and $P(y)$ to estimate $P(y|x)$

Both Base on Bayes Rule : $P(y|x) = \frac{P(x|y) \times P(y)}{P(x)}$

Prove multinomial Naive Bayes is a linear Classifier

T → T = T
T → F = F

Proof - Assume $y \in \{-1, +1\}$

- $h(x) = \pm 1$ iff $P(y=+1|x) \geq P(y=-1|x)$ T.

ด้านหลัง

iff $P(x|y=+1) \times P(y=1) \geq P(x|y=-1) \times P(y=-1)$

iff $\prod_{\alpha=1}^{d} P(X_\alpha|y=+1) \times P(y=+1) \geq \prod_{\alpha=1}^{d} P(X_\alpha|y=-1) \times P(y=-1)$

iff $\sum_{\alpha=1}^{d} \log_e P(X_\alpha|y=+1) + \log_e P(y=+1) \geq \sum_{\alpha=1}^{d} \log_e P(X_\alpha|y=1) + \log_e P(y=-1)$

$([\theta_{+1}]_\alpha)^{x_\alpha}$

iff $\sum_{\alpha=1}^{d} (\log_e P(X_\alpha|y=+1) - \log_e P(X_\alpha|y=-1)) + (\log_e P(y=+1) - \log_e P(y=-1)) \geq 0$

iff $\sum_{\alpha=1}^{d} x_\alpha (\log_e [\theta_{+1}]_\alpha - \log_e [\theta_{-1}]_\alpha) + (\log_e P(y=+1) - \log_e P(y=-1)) \geq 0$

$\vec{w}^T \vec{x}$; b

∴ เหมือนสูตร $\vec{w}^T \vec{x} + b$ ในการหา hyperplane ของ perceptron

$$\vec{w} = \begin{bmatrix} \log_e [\theta_{+1}]_1 - \log_e [\theta_{-1}]_1 \\ \vdots \\ \log_e [\theta_{+1}]_d - \log_e [\theta_{-1}]_d \end{bmatrix} ; \vec{x} = x_\alpha$$

iff $\vec{w}^T \vec{x} + b \geq 0$

Case 3 : Continuous feature: : Guassian Naive Bayes classifier

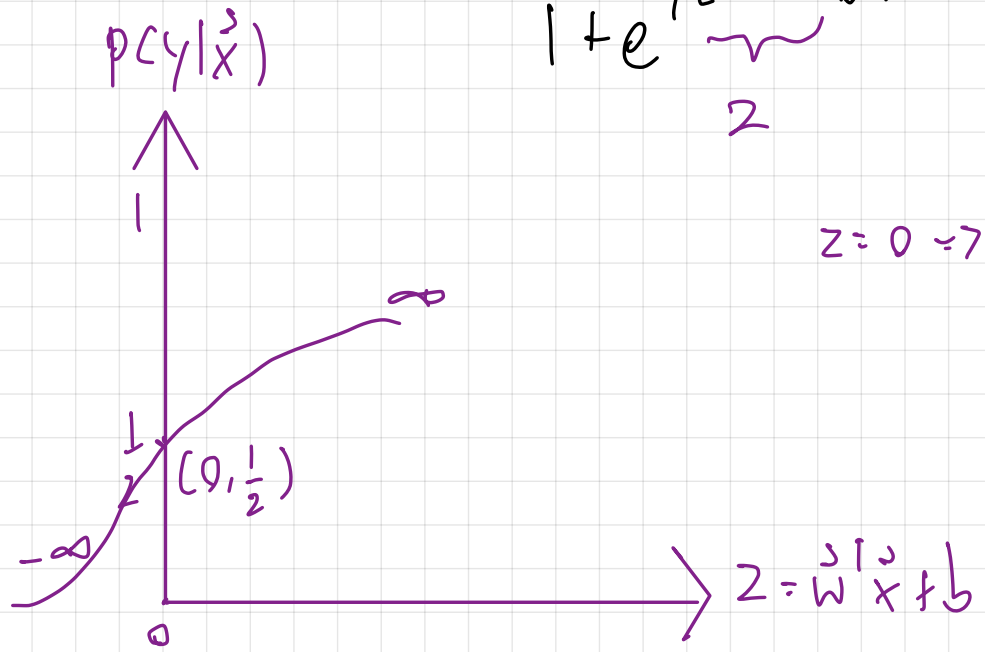
- $[X]_\alpha = \mathbb{R}$
- Model $P([X]_\alpha = j \mid Y = y)$

$$[X]_\alpha \sim N([\mu_y]_\alpha, [\sigma_y]_\alpha)$$

- MLE -> $[\mu_y] = \frac{\sum_{i=1}^{n} I(y_i = y) \cdot x_i^\alpha}{n}$ same with $\frac{\sum x}{n}$

- For Guassian Naive Bayes, we will arrive the following expression by taking the same derivation

$$P(y \mid \vec{x}) = \frac{1}{1 + e^{-y(\vec{w}^T \vec{x} + b)}}, \quad \text{for } y \in \{-1, +1\}$$



$z = 0 \Rightarrow e^{-yz} = e^{-0} = 1$

Recall before Prove

Discriminative learning : Try to model $P(y \mid x)$ (eg. k-NN, Perceptron)

Generative learning : try to model $P(x \mid y)$ and $P(y)$ to estimate $P(y \mid x)$

Both Base on Bayes Rule : $P(y \mid x) = \frac{P(x \mid y) \times P(y)}{P(x)}$

$$P(y \mid x) \propto P(x \mid y) \times P(y)$$

- Discrimative learning : Try to model $P(y \mid x)$ directly
- Generative learning : Try to model $P(x \mid y)$ and $P(y)$

eg. Perceptron is a discrimative algorithm

$$P(y \mid x) = \begin{cases} 1 & \text{if } w^T x \geq 0 \\ & \hookrightarrow \text{pobability} \\ 0 & \text{otherwise} \end{cases}$$

เมื่อ $y = +1$

∴ บ่งชี้ได้ว่า โอกาสคือ 1 ถ้าไม่ใช่ก็เป็น 0

eg. Naive Bayes is a generative algorithm

try to model distribution $\begin{cases} P(y) \\ P(x \mid y) = \prod_{\alpha=1}^{d} P(x_\alpha \mid y) \end{cases}$

linear classifier: A classifier $h(x)$ is called 'linear' if $h(x) = +1$ if and only if $\exists\, w, b$ such that

$w^T x + b \geq 0$; assume $y \in \{1, -1\}$

∴ ถ้า b รวมใน w => $w^T x \geq 0$

Note: $h(x)$ จะเป็น linear ก็ต่อเมื่อ มี w, b ซึ่ง $w^T x + b \geq 0$

eg. Perceptron, Multinomial Naive Bayes are linear classifiers

- By taking the similar derivation, we can derive the following expression for Guassian Naive Bayes

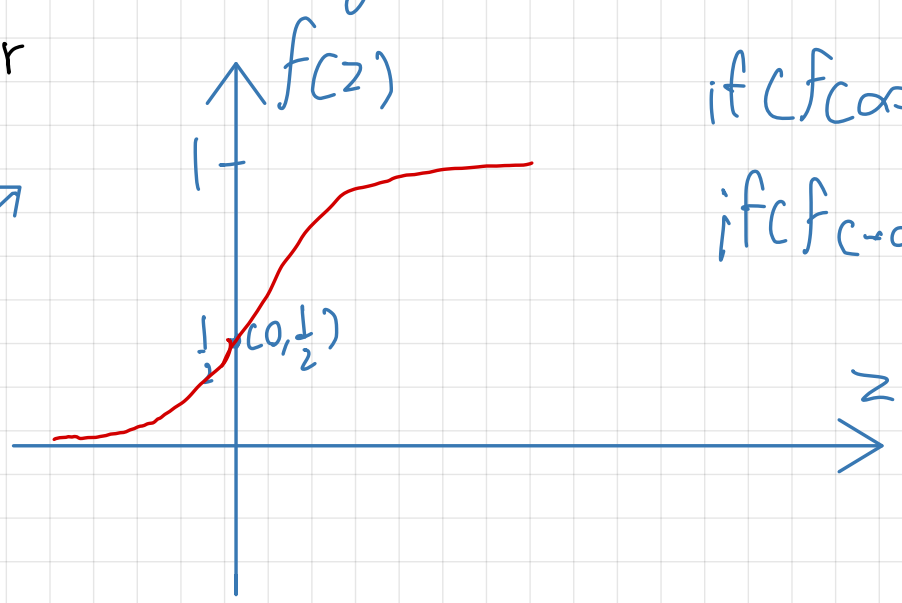
$$P(y|x) = \frac{1}{1 + e^{-y(\vec{w}^T \vec{x})}} \quad ; y \in \{-1, +1\}$$

Define $z = w^T x$; z is sclar

$$f(z, y=+1) = \frac{1}{1 + e^{-z}}$$

$$f(z, y=-1) = \frac{1}{1 + e^{z}}$$

sigmond function ♡



if $(f(\infty) = \frac{1}{1 + z^{-\infty}} = 1)$

if $(f(-\infty) = 0)$

$0 \leq f(z) \leq 1$ same with probability

Recall #1 missclassification occurs when $y(w^T x) < 0$ $P(y|x) < \frac{1}{2}$

correct classification $y(w^T x) = yz \geq 0$ $P(y|x) \geq \frac{1}{2}$

Recall #2 In correct classification (คือ $yz \geq 0$)

$w^T x$ measures the distance from x to the hyperplane. and

x is very far from the hyperplane, then $w^T x$ will be large quantity (คือ $y=+1$)

observations:

- If x lies on the right side of the hyperplane and very far from the hyperplane, then $P(y|x) = 1$
- If x lies on the wrong side of the hyperplane and x is very far from the hyperplane, the $P(y|x) =$