Introduction

lecture 13

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

Note: ดึง (x_i, y_i) ใน pobability distribution X, Y ($P(X,Y)$) ถ้าอยากทำนาย X_{test} ต้องหาจาก $P(x,y)$ ซึ่งไม่สามารถเข้าถึงได้จึงต้องสร้าง modeling distribution มา $P(X,Y;\theta)$

$(x_i, y_i) \sim P(X,Y)$

assume it can access

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

Classifier $h(x_{test}) = \underset{y}{\operatorname{argmax}}\ P(Y=y \mid X=x_{test};\theta)$

Bayes

Need to learn $\boxed{P(Y|X)}$ **

① optimal classifier $h(x_t) = \underset{y}{\operatorname{argmax}}\ P(Y=y \mid X=x_t)$; pobs ที่ $X=x_t$ แล้ว Y เป็นของค่า y (เช่น samlon, mackael)

pobs ของ salmon เมื่อ $X=x_{test}$

pobs ของ mackel เมื่อ $X=x_{test}$

③ $P(Y|X)$

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

r.v ที่อยู่ใน $d=1$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad X = \begin{bmatrix} [X]_1 \\ [X]_2 \\ \vdots \\ [X]_d \end{bmatrix}$$

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

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

- เกิดจากการเรียก Bayes classifier

Bayes Classifier

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

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 $P(X=x)\ \forall x$ $\quad P(X)$

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

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

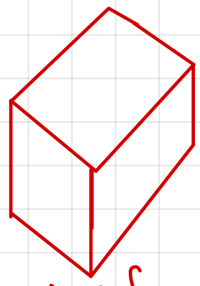
Coin tossing: n times

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



$i \times j$ faces

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

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

→ โยน สนใจที่หน้าออกเป็น y,x ใน Y,X

$I(E)$: indicator random varible

Assign ตัวแปรสุ่มที่มีค่า 1, 0

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

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

ใช้แทน ↓ corresponse

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

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

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

Assume X, Y follow Binominal distribution

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

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

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

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

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

- X เล็กที่หลายdimension => Vector (X ใหญ่ควรเป็น Vector)

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

$$\Downarrow$$

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

$$\vec{X} = \underbrace{\begin{bmatrix} [X]_1 \\ [X]_2 \\ \vdots \\ [X]_d \end{bmatrix}}_{\text{r.v.}} \qquad \vec{x} = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^d \end{bmatrix} \leftarrow \text{fixed}$$

r.v X = fixed x on each d

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

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

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

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: $P(Y=y \mid X=x) = \dfrac{P(X=x \mid Y=y)\,P(Y=y)}{P(X=x)}$

Bayes Classifier

$$= h(x) = \arg\max_y P(Y=y \mid X=x)$$

$$= \arg\max_y \frac{P(X=x \mid Y=y) \cdot P(Y=y)}{\cancel{P(X=x)}} \quad \text{ไม่เกี่ยวกับ } y$$

$$= \arg\max_y P(X=x \mid Y=y) \cdot P(Y=y)$$

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

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

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

$$= P([\vec{X}]_1 = \vec{x}^1 \land [\vec{X}]_2 = \vec{x}_2 \land \ldots \land [\vec{X}]_d = \vec{x}_d \mid Y=y)$$ 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 ที่จะได้ $[X]_i = x^i$ จะอิสระต่อกัน

Naive Bayes classifier

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

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 จำเพาะของ j y ที่ codinate α

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

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

$$\Rightarrow \frac{\sum_{i=1}^{n} I(Y_i = y) \cdot I(X_i^\alpha = j)}{\sum_{i=1}^{n} I(Y_i = y)}$$: Count เมื่อเจอ $Y_i = y$ 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)

Catagorical feature

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

Estimate $P([\vec{x}]_1 | Y)$

$\rightarrow P([\vec{x}]_1 = sunny \mid y = go\text{-}out)$

$\rightarrow P([\vec{x}]_1 = rainy \mid y = go\text{-}out)$

$\rightarrow P([\vec{x}]_1 = sunny \mid y = stay)$

$\rightarrow P([\vec{x}]_1 = rainy \mid y = stay)$

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

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." $\xrightarrow{\text{Back of word}}$ $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_d \end{bmatrix}$ ← # of word 1 appearing in the text

: its count of word in Sentense.

$\}$ cordinate

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

eg. Span filter

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

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

Estimate $P([\vec{x}]_\alpha = j \mid Y=y)$; $\alpha : 5, j = 10, y = \text{spam}$ Given W_5 = Princess (เกิด 10ครั้ง)

$P([\vec{x}]_5 = 10 \mid Y = \text{SPAM})$; given spam ที่ตัวแปรส่วน cordinate 5 = 10

Modeling

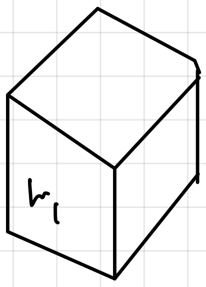
$$\boxed{P([\vec{x}]_\alpha = j \mid Y=y) = \left[\binom{m}{j} \cdot P(w_\alpha \mid Y=y)\right]^j}$$

eg. $P[[\vec{x}]_2 = 3 \mid y = \text{spam}] = \left[\binom{5}{3} \times P(w_2 \mid Y = \text{spam})\right]^3$

$\downarrow$ word ที่ 2

- $P(w_2 \mid Y=y)$ 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)

$$\vec{x} = \begin{pmatrix} 1 & x^1 \\ 4 & x^2 \\ \vdots & \vdots \\ & x^d \end{pmatrix} \begin{matrix} - W_1 \\ - W_2 \\ \\ - W_d \end{matrix}$$

Assume throw and get R1:

$$\underbrace{W_2, W_2, W_1, W_2, W_2, W_3, W_2}_{m \text{ times } = 7}$$

R2:

$$\underbrace{W_1, W_2, W_2, W_3, W_2, W_2, W_2}_{n \text{ times } = 7}$$

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

- We model $P(W_\alpha \mid Y = \text{spam}) = [\theta_{spam}]_\alpha$

$P(W_\alpha \mid Y = \text{ham}) = [\theta_{ham}]_\alpha$

- Estimate

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

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

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

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

∴ โจทย์ หาครั้งเห็น money 2ครั้ง

แต่ไม่รู้อยู่ตรงไหนบ้าง เลยใช้ $\binom{m}{2}$ ด้วย

Summary

Bayes Rule → $h(x) = \underset{y}{\arg\max}\, P(X = x \mid Y = y) \cdot P(y)$

Naive Bayes Classifier → $= \underset{y}{\arg\max} \prod_{\alpha=1}^{d} P([x]_\alpha = x^\alpha \mid Y = y) \cdot P(y)$

Spam filter (text classification)

"An ant is an animal" ; 5 words

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

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

Ingeneral

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

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

$\to$ จน. ครั้งแต่ละ word ปรากฏอยู่ใน y

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

$$= \underbrace{P([X]_1 = x^1 \mid Y = y)}_{} \times \ldots \times P([X]_\alpha = x^\alpha \mid Y = y)$$

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

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

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

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

$0!$

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

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

Binary Classification

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

take $\log_e$ into both sides

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

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

3 notable class

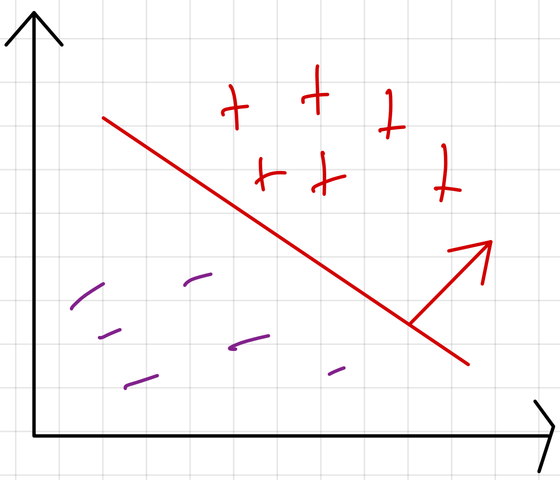
① Categorical features ($[X]_\alpha$ เป็นประเภท(แทน word) ไม่ใช่จำนวนครั้ง
$[X]_\alpha \in \{0, 1, \dots, k-1\}$ ของการปรากฏ) Catagorical Naive Bayes Classifier

② multinomial feature ($[X]_\alpha$ แทนความถี่ของการปรากฏ ของแต่ละ word)
$[X]_\alpha \in \{0, 1, \dots, m\}$ $\quad m = \sum_{\alpha=1}^{d} x^\alpha$ เป็นความถี่ (จน.ครั้งของการปรากฏ)
↳ multinomial Naive Bayes Classifier

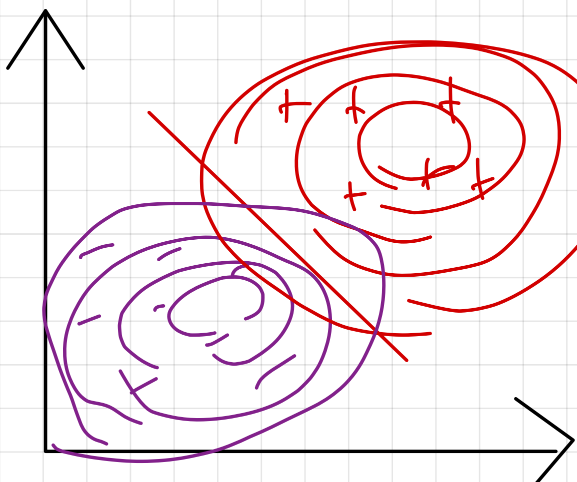
③ Continous feature
$[X]_\alpha \in \mathbb{R}$ ($[X]_\alpha$ เป็นค่าต่อเนื่อง)
↳ Guassian Naive Bayes Classifier

Summary of Naive Bayes

- Naive Bayes = Bayes classifier + Bayes rule + Naive Bayes assumption
- The assumption says "all feature values are Indepedent."
    - พิจารณา W_α เมื่อให้ label $y = spam$ $P(W_\alpha \mid y = spam)$
- 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

- สร้าง distribution ของ probs. ให้ sample

- hyperplane separate ตาม data
- separate ตาม distribution

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

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

ทำนาย y โดยให้ 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 : $P(Y|x) = \dfrac{P(x|Y) \times P(y)}{P(x)}$

Prove multinomial Naive Bayes is a linear Classifier

Proof - Assume $y \in \{-1, +1\}$ $\quad$ $T \rightarrow T = T$ $\quad$ $T \rightarrow F = F$

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

ด้านหากล่า

$$\text{iff } \boxed{P(x|y{=}{+1})} \times P(y{=}1) \geq \boxed{P(x|y{=}{-1})} \times P(y{=}{-1})$$

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

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

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

$$\text{iff } \sum_{\alpha=1}^{d} \left(\log_e P(\underline{x_\alpha|y{=}{+1}}) - \log_e P(x_\alpha|y{=}{-1})\right) + \left(\log_e P(y{=}{+1}) - \log_e P(y{=}{-1})\right) \geq 0$$

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

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

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

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

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

Case 3 : Continuous feature : Guassian Naive Bayes Classifier

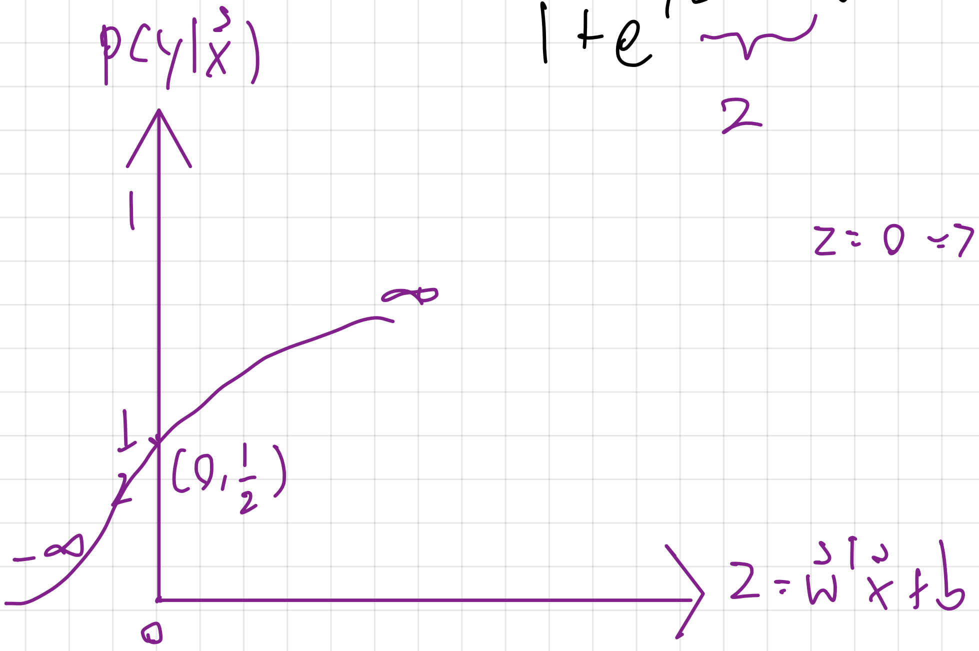
- $[x]_\alpha = \mathbb{R}$
- Model $P([x]_\alpha = j \mid Y = y)$

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

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

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

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



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

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 : $P(Y|X) = \dfrac{P(X|Y) \times P(Y)}{P(X)}$

Recall before Prove

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

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

eg. Perceptron is a discrimative algorithm

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

↳ pobability

เมื่อ $y = +1$

∴ มั่นใจได้ว่าโอกาสคือ 1 ถ้าไม่ใช่ก็เป็น 0

eg. Naive Bayes is a generative algorithm

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

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

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

$\therefore$ ยัด b ไปใน w $\Rightarrow w^T x \geq 0$

Note: h() จะเป็น linear ถ้า $h(x)=+1$ 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

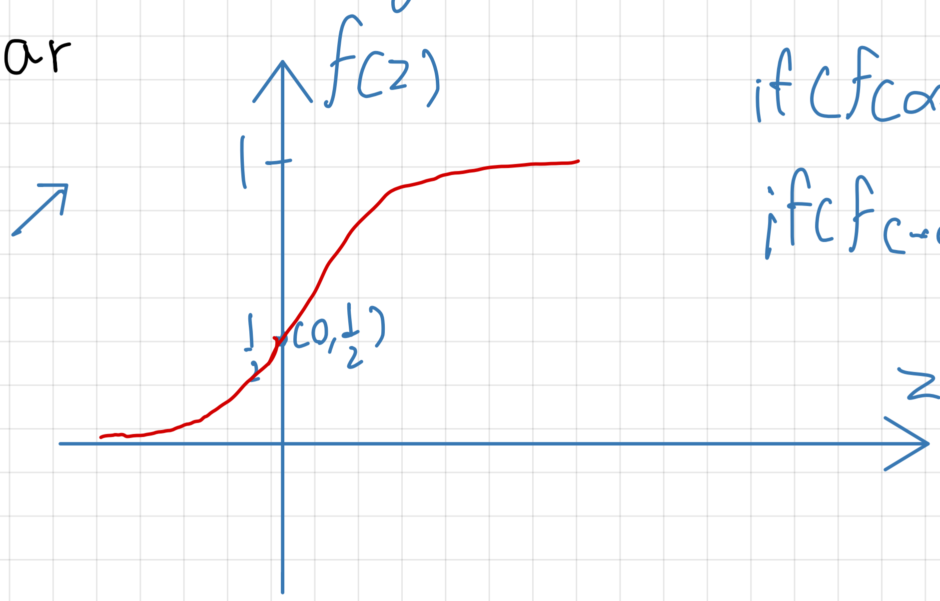
$$P(y|x) = \boxed{\frac{1}{1+e^{-y(\vec{w}^T \vec{x})}}} \; ; \; y \in \{-1,+1\}$$

sigmond function ♡

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

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

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



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

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

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

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

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

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

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

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

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 $P(y|x) =$