

Classification: Probabilistic Generative Model

Classification



- Credit Scoring
 - Input: income, savings, profession, age, past financial history
 - Output: **accept or refuse**
- Medical Diagnosis
 - Input: current symptoms, age, gender, past medical history
 - Output: **which kind** of diseases
- Handwritten character recognition
- Face recognition
 - Input: image of a face, output: person

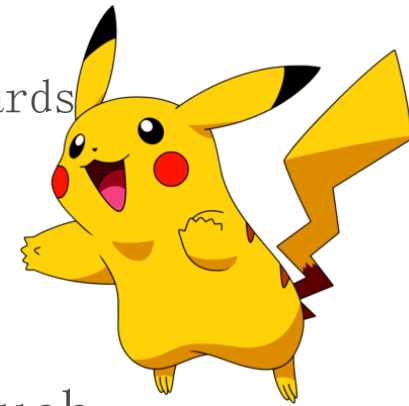
Input:  output: **金**

Example Application



$$f(\text{Pikachu}) = \text{Electric}$$
$$f(\text{Squirtle}) = \text{Water}$$
$$f(\text{Bulbasaur}) = \text{Grass}$$

pokemon games
(*NOT* pokemon cards
or Pokemon Go)



Example Application

- **HP:** hit points, or health, defines how much damage a pokemon can withstand before fainting **35**
- **Attack:** the base modifier for normal attacks (eg. Scratch **55** power)
- **Defense:** the base damage resistance against normal attacks **40**
- **SP Atk:** special attack, **50** base modifier for special attacks (e.g. fire blast, bubble beam) **50**
- **SP Def:** the base damage resistance against special attacks **90**

- **Speed** Can we predict the “type” of pokemon based on the information?

Example Application

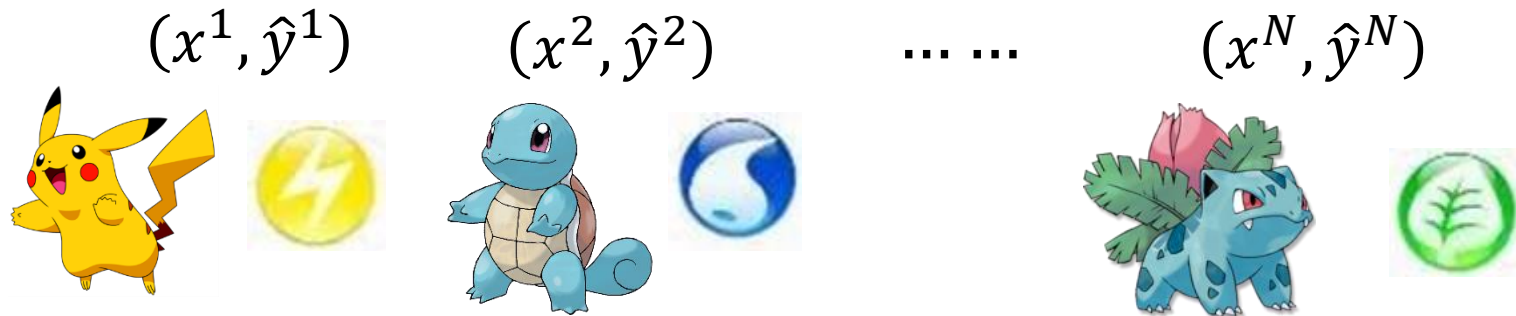
×	防禦方的屬性																		
	一般	格鬥	飛行	毒	地面	岩石	蟲	幽靈	鋼	火	水	草	電	超能力	冰	龍	惡	妖精	
攻擊方的屬性	一般	1×	1×	1×	1×	1/2×	1×	0×	1/2×	1×	1×	1×	1×	1×	1×	1×	1×	1×	
	格鬥	2×	1×	1/2×	1/2×	1×	2×	1/2×	0×	2×	1×	1×	1×	1×	1/2×	2×	1×	2×	1/2×
	飛行	1×	2×	1×	1×	1×	1/2×	2×	1×	1/2×	1×	1×	2×	1/2×	1×	1×	1×	1×	1×
	毒	1×	1×	1×	1/2×	1/2×	1/2×	1×	1/2×	0×	1×	1×	2×	1×	1×	1×	1×	1×	2×
	地面	1×	1×	0×	2×	1×	2×	1/2×	1×	2×	2×	1×	1/2×	2×	1×	1×	1×	1×	1×
	岩石	1×	1/2×	2×	1×	1/2×	1×	2×	1×	1/2×	2×	1×	1×	1×	1×	2×	1×	1×	1×
	蟲	1×	1/2×	1/2×	1/2×	1×	1×	1×	1/2×	1/2×	1/2×	1×	2×	1×	2×	1×	1×	2×	1/2×
	幽靈	0×	1×	1×	1×	1×	1×	1×	2×	1×	1×	1×	1×	1×	2×	1×	1×	1/2×	1×
	鋼	1×	1×	1×	1×	1×	2×	1×	1×	1/2×	1/2×	1/2×	1×	1/2×	1×	2×	1×	1×	2×
	火	1×	1×	1×	1×	1×	1/2×	2×	1×	2×	1/2×	1/2×	2×	1×	1×	2×	1/2×	1×	1×
	水	1×	1×	1×	1×	2×	2×	1×	1×	1×	2×	1/2×	1/2×	1×	1×	1×	1/2×	1×	1×
	草	1×	1×	1/2×	1/2×	2×	2×	1/2×	1×	1/2×	1/2×	2×	1/2×	1×	1×	1×	1/2×	1×	1×
	電	1×	1×	2×	1×	0×	1×	1×	1×	1×	1×	2×	1/2×	1/2×	1×	1×	1/2×	1×	1×
	超能力	1×	2×	1×	2×	1×	1×	1×	1×	1/2×	1×	1×	1×	1×	1/2×	1×	1×	0×	1×
	冰	1×	1×	2×	1×	2×	1×	1×	1×	1/2×	1/2×	1/2×	2×	1×	1×	1/2×	2×	1×	1×
	龍	1×	1×	1×	1×	1×	1×	1×	1×	1/2×	1×	1×	1×	1×	1×	1×	2×	1×	0×
	惡	1×	1/2×	1×	1×	1×	1×	1×	2×	1×	1×	1×	1×	1×	2×	1×	1×	1/2×	1/2×
妖精	1×	2×	1×	1/2×	1×	1×	1×	1×	1/2×	1/2×	1×	1×	1×	1×	1×	2×	2×	1×	

這些倍數適用於XY及之後的遊戲。

這些倍數適用於XY及之後的遊戲。

How to do Classification

- Training data for Classification

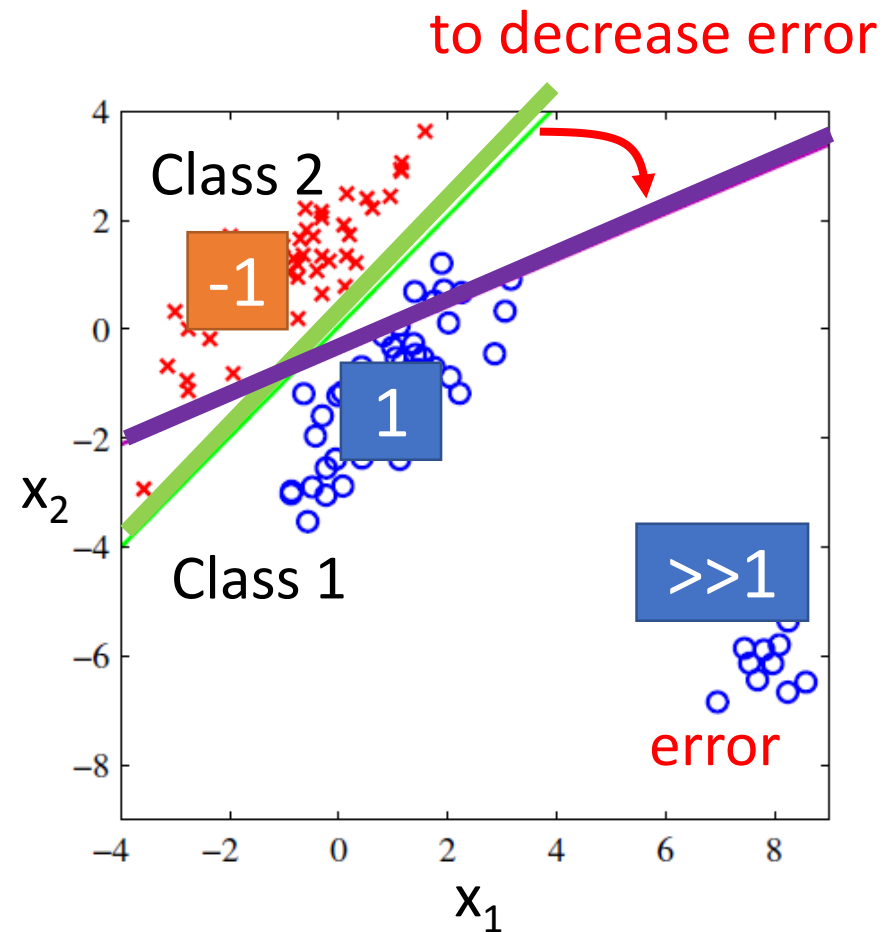
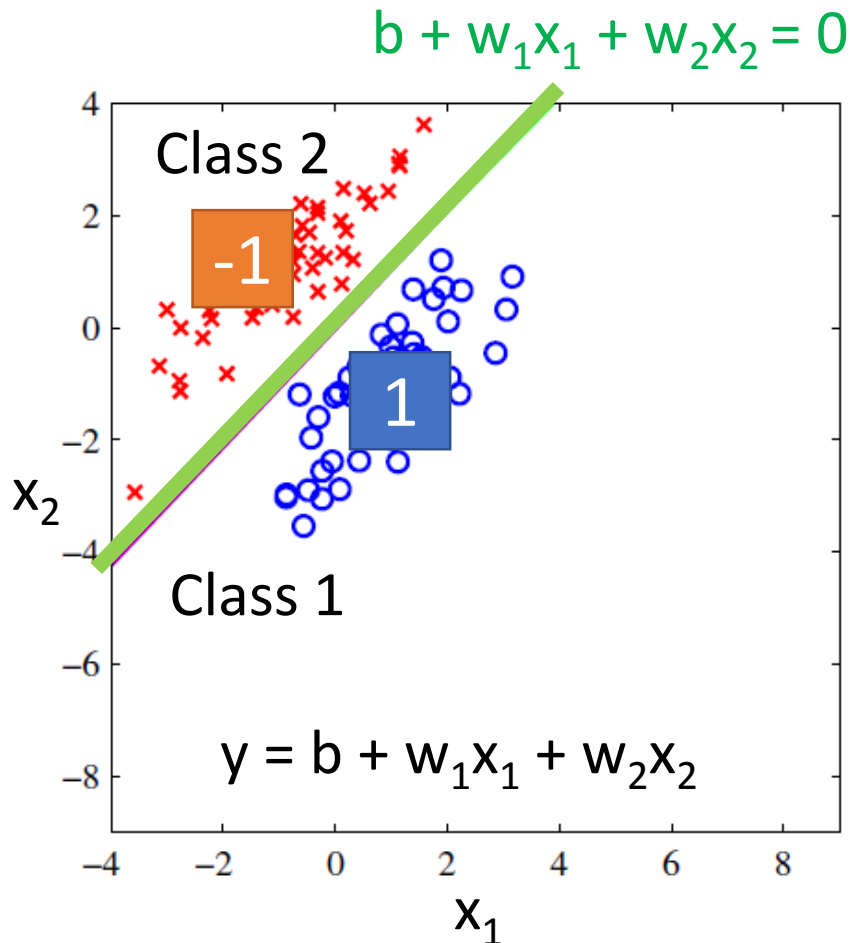


Classification as Regression?

Binary classification as example

Training: Class 1 means the target is 1; Class 2 means the target is -1

Testing: closer to 1 \rightarrow class 1; closer to -1 \rightarrow class 2

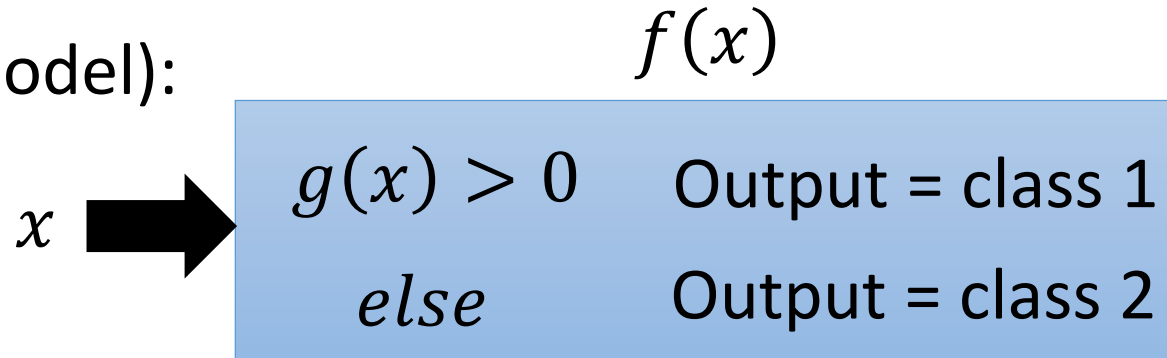


Penalize to the examples that are “too correct” ... (Bishop, P186)

- Multiple class: Class 1 means the target is 1; Class 2 means the target is 2; Class 3 means the target is 3 problematic

Ideal Alternatives

- Function (Model):



- Loss function:

$$L(f) = \sum_n \delta(f(x^n) \neq \hat{y}^n)$$

只关心对错，不关心错多少远的问题

The number of times f get incorrect results on training data.

- Find the best function:

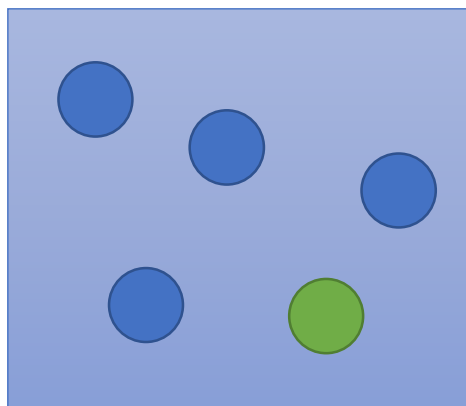
- Example: Perceptron, SVM

Not Today

Two Boxes

Box 1

$$P(B_1) = 2/3$$

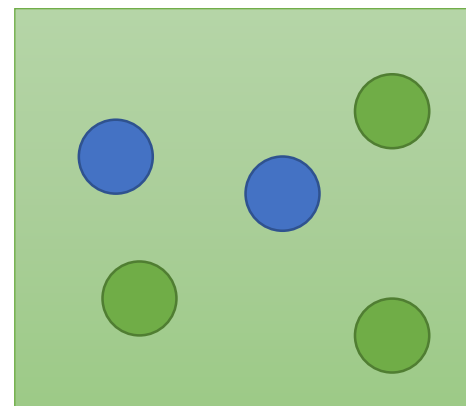


$$P(\text{Blue} | B_1) = 4/5$$

$$P(\text{Green} | B_1) = 1/5$$

Box 2

$$P(B_2) = 1/3$$



$$P(\text{Blue} | B_2) = 2/5$$

$$P(\text{Green} | B_2) = 3/5$$

 from one of the boxes

Where does it come from?

随便从box1或box2中取一球，
从盒子1中取出一个篮球的概率是
 $P(B_1) \cdot P(\text{Blue} | B_1)$ 区分与 已经知道取的是
个蓝球，再求蓝球来自于box1的机率

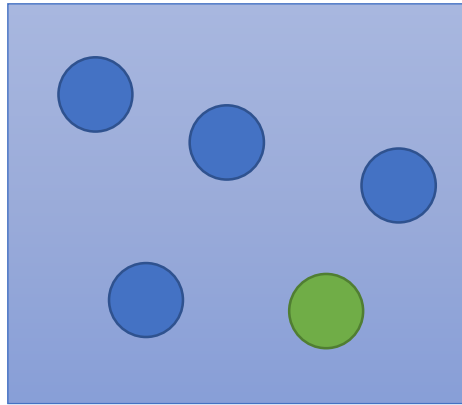
$$P(B_1 | \text{Blue}) = \frac{P(\text{Blue} | B_1)P(B_1)}{P(\text{Blue} | B_1)P(B_1) + P(\text{Blue} | B_2)P(B_2)}$$

Two Classes

Estimating the Probabilities
From training data

Class 1

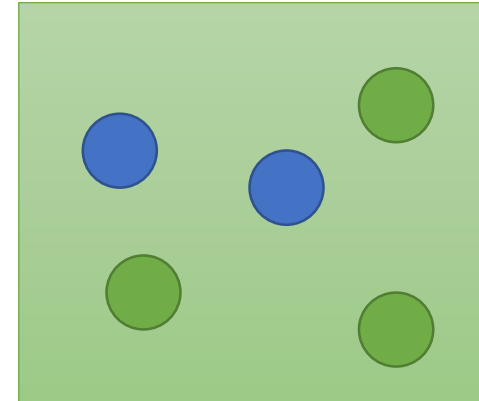
$P(C_1)$



$P(x|C_1)$

Class 2

$P(C_2)$



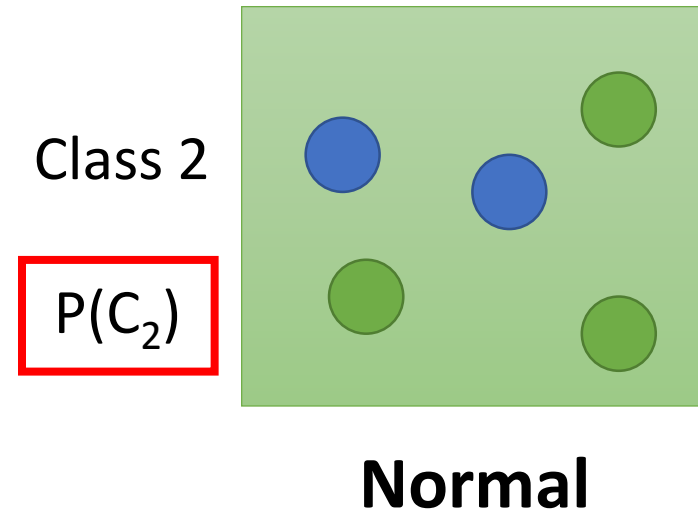
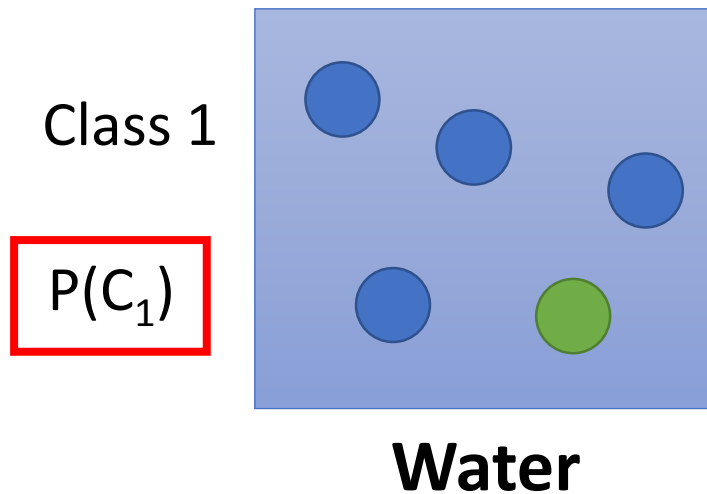
$P(x|C_2)$

Given an x , which class does it belong to

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

Generative Model $P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$

Prior



Water and Normal type with ID < 400 for training,
rest for testing

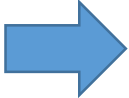
Training: 79 Water, 61 Normal

$$P(C_1) = 79 / (79 + 61) = 0.56$$

$$P(C_2) = 61 / (79 + 61) = 0.44$$

Probability from Class

$$P(x|C_1) = ? \quad P(\text{  | \text{Water}) = ?$$

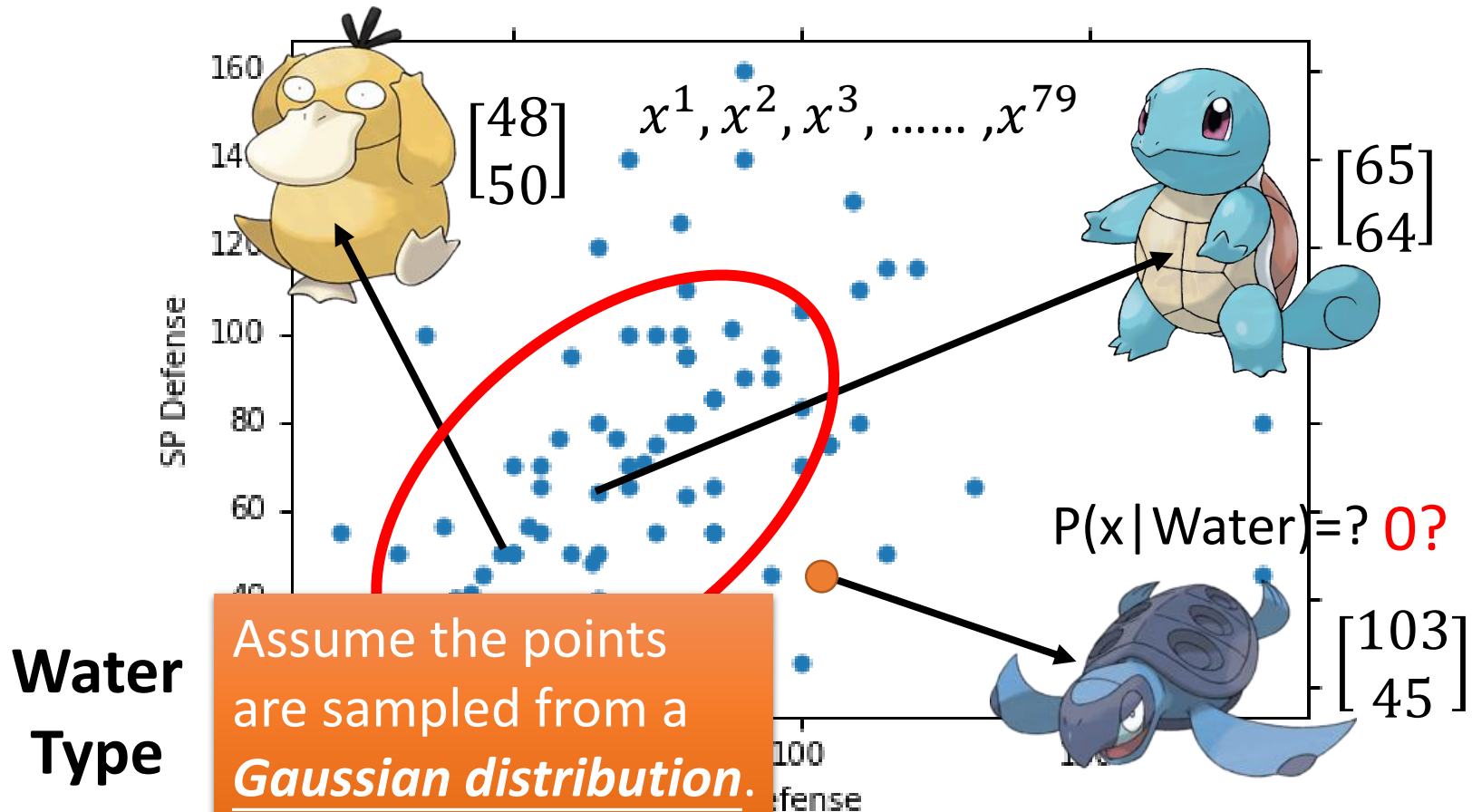
Each Pokémon is represented as a vector by its attribute.  feature

**Water
Type**



Probability from Class - Feature

- Considering **Defense** and **SP Defense**

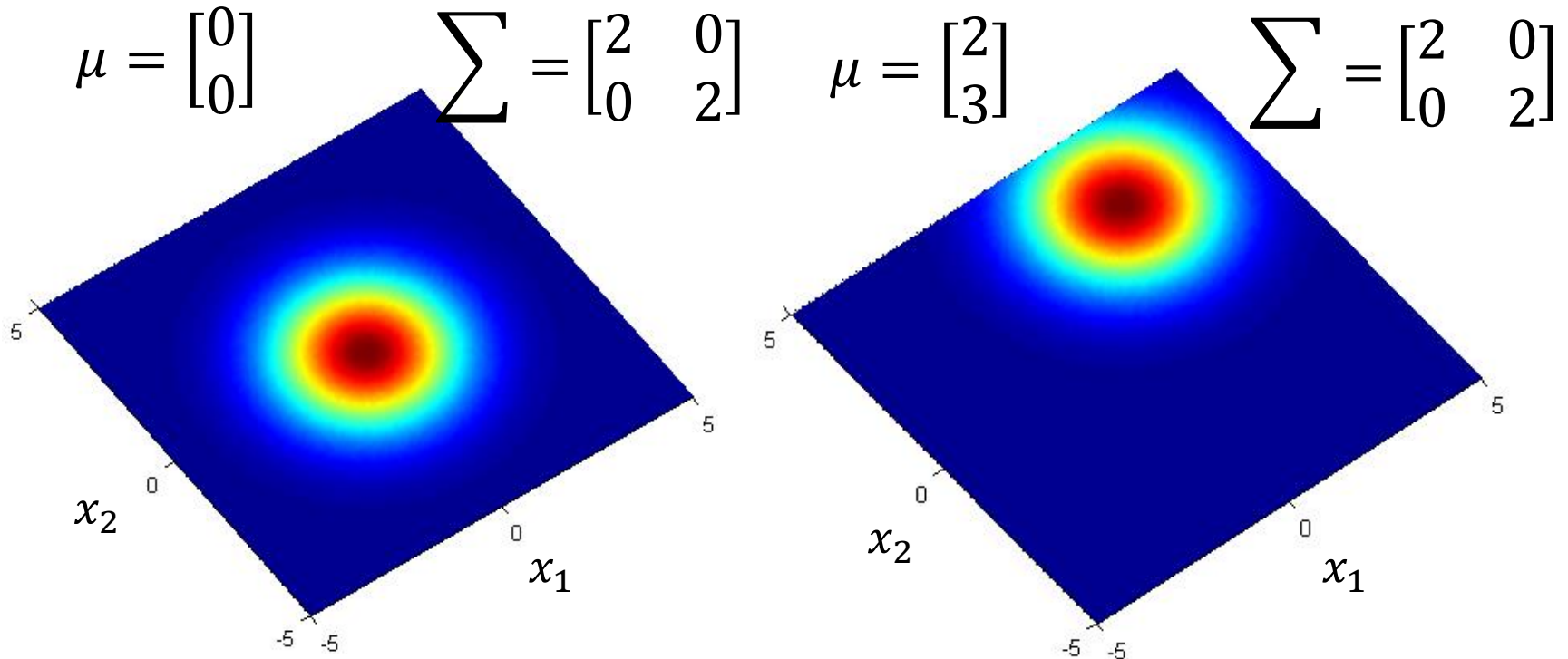


Gaussian Distribution

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

Input: vector x , output: probability of sampling x

The shape of the function determines by **mean μ** and **covariance matrix Σ**

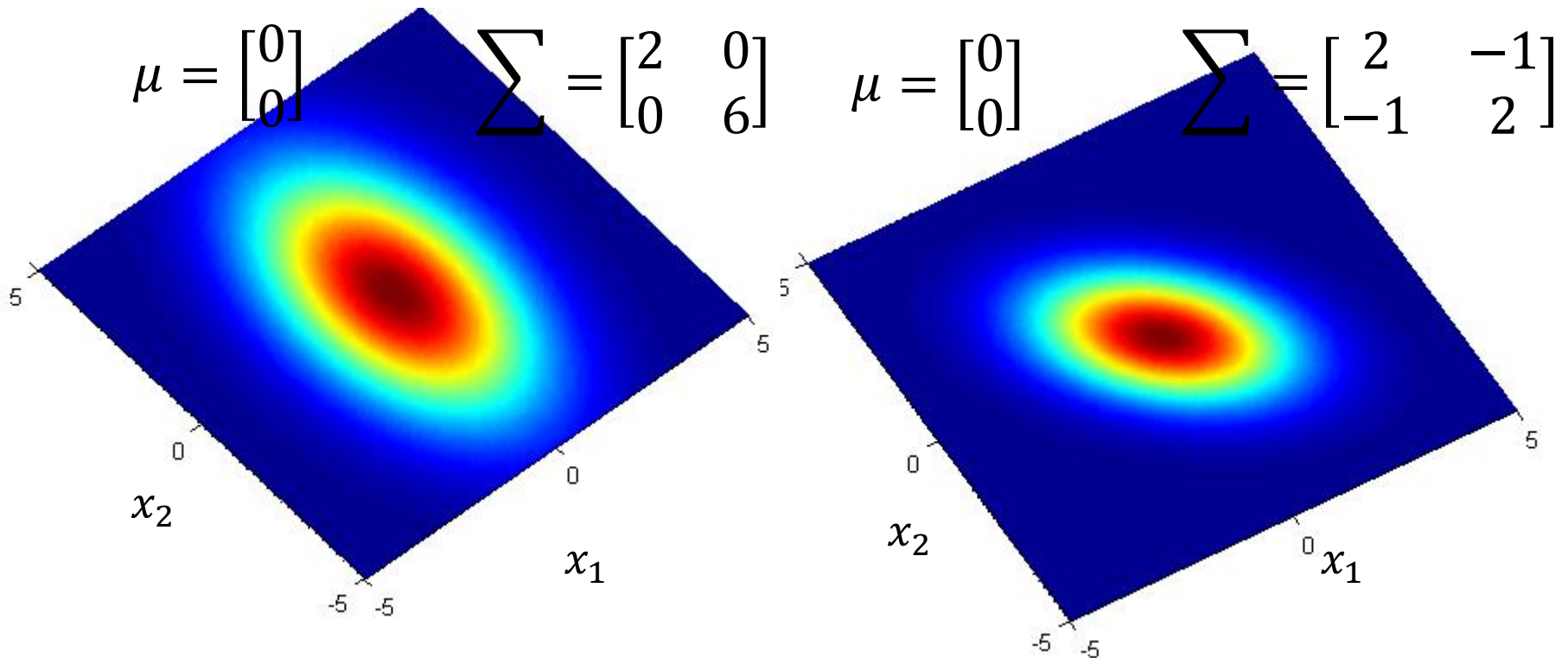


Gaussian Distribution

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

Input: vector x , output: probability of sampling x

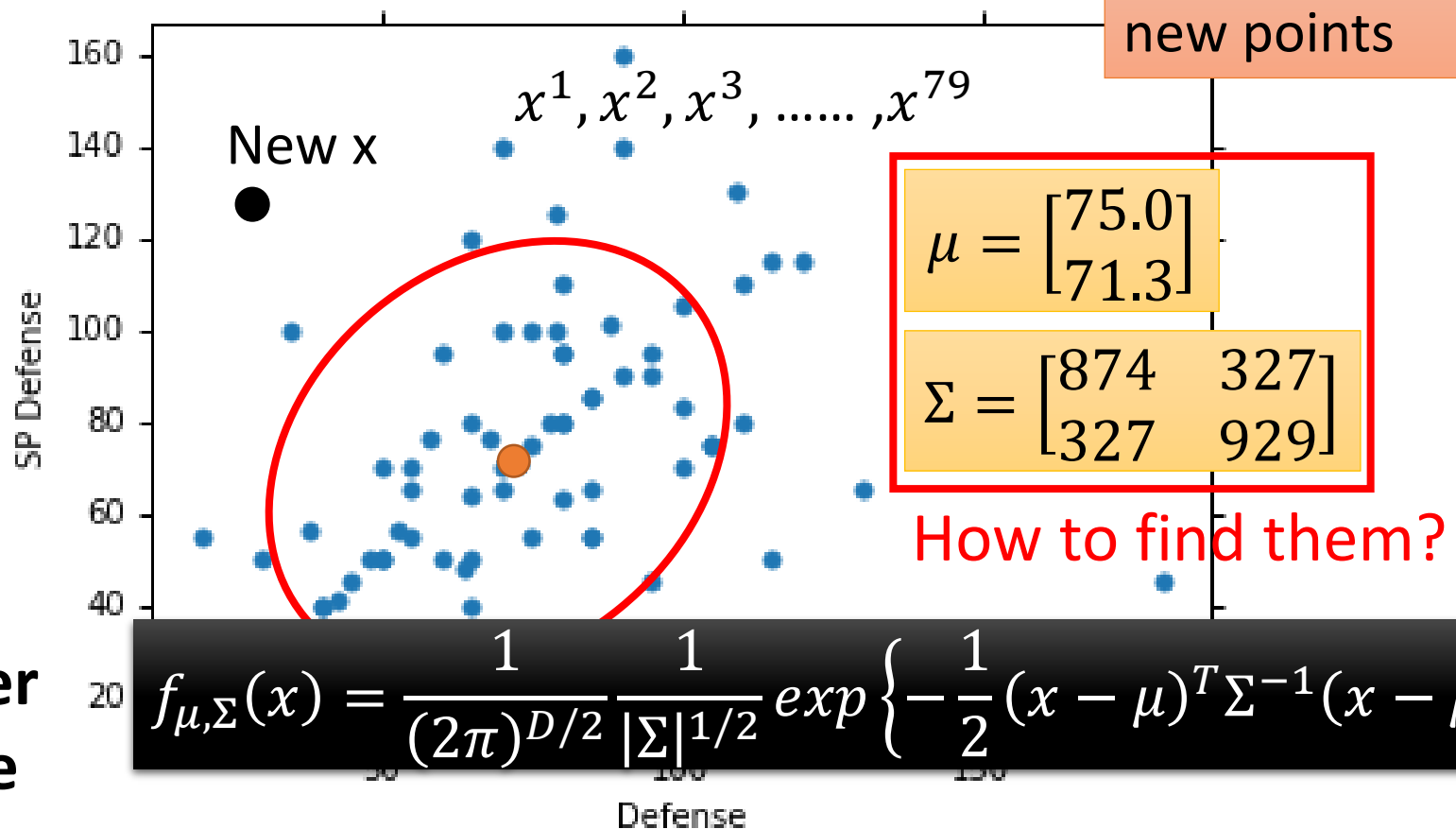
The shape of the function determines by **mean μ** and **covariance matrix Σ**



Probability from Class

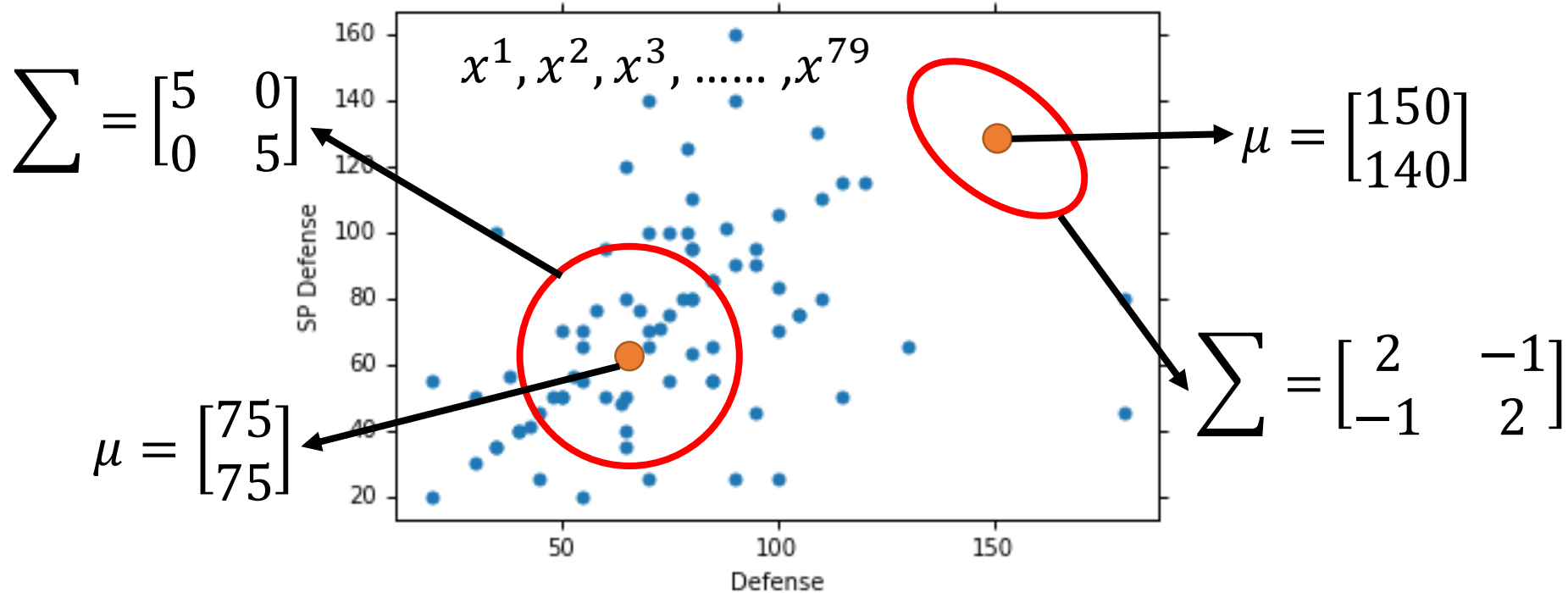
Assume the points are sampled from a Gaussian distribution

Find the Gaussian distribution behind them → Probability for new points



Water
Type

Maximum Likelihood $f_{\mu, \Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$



The Gaussian with any mean μ and covariance matrix Σ can generate these points. ➡ Different Likelihood

Likelihood of a Gaussian with mean μ and covariance matrix Σ
 = the probability of the Gaussian samples $x^1, x^2, x^3, \dots, x^{79}$

$$L(\mu, \Sigma) = f_{\mu, \Sigma}(x^1) f_{\mu, \Sigma}(x^2) f_{\mu, \Sigma}(x^3) \dots f_{\mu, \Sigma}(x^{79})$$

Maximum Likelihood

We have the “Water” type Pokémons: $x^1, x^2, x^3, \dots, x^{79}$

We assume $x^1, x^2, x^3, \dots, x^{79}$ generate from the Gaussian (μ^*, Σ^*) with the **maximum likelihood**

$$L(\mu, \Sigma) = f_{\mu, \Sigma}(x^1) f_{\mu, \Sigma}(x^2) f_{\mu, \Sigma}(x^3) \dots f_{\mu, \Sigma}(x^{79})$$

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

$$\mu^*, \Sigma^* = \arg \max_{\mu, \Sigma} L(\mu, \Sigma) \quad \Sigma^* = \frac{1}{N} X^T X \quad \begin{array}{l} X \text{是所有样本的集合} \\ \text{每一行代表一个样本} \end{array}$$

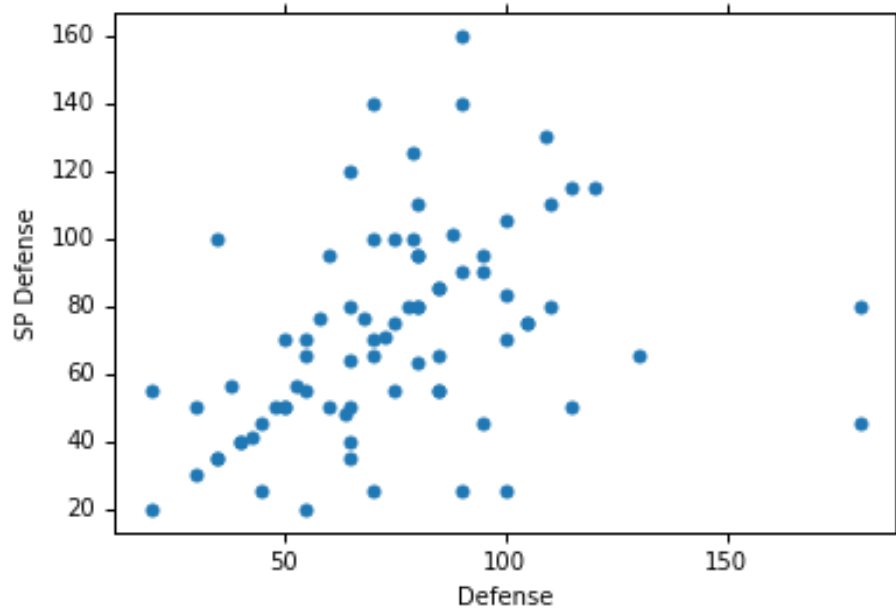
$$\mu^* = \frac{1}{79} \sum_{n=1}^{79} x^n$$

average

$$\Sigma^* = \frac{1}{79} \sum_{n=1}^{79} (x^n - \mu^*) (x^n - \mu^*)^T$$

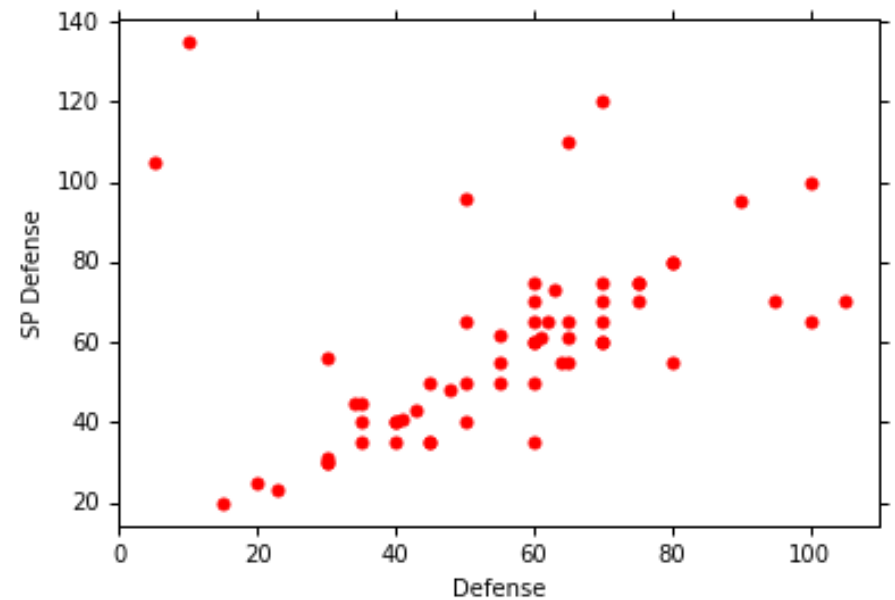
Maximum Likelihood

Class 1: Water



$$\mu^1 = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

Class 2: Normal



$$\mu^2 = \begin{bmatrix} 55.6 \\ 59.8 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 847 & 422 \\ 422 & 685 \end{bmatrix}$$

Now we can do classification 😊

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

$P(C_1) = 79 / (79 + 61) = 0.56$

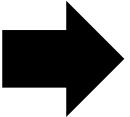
$$\mu^1 = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

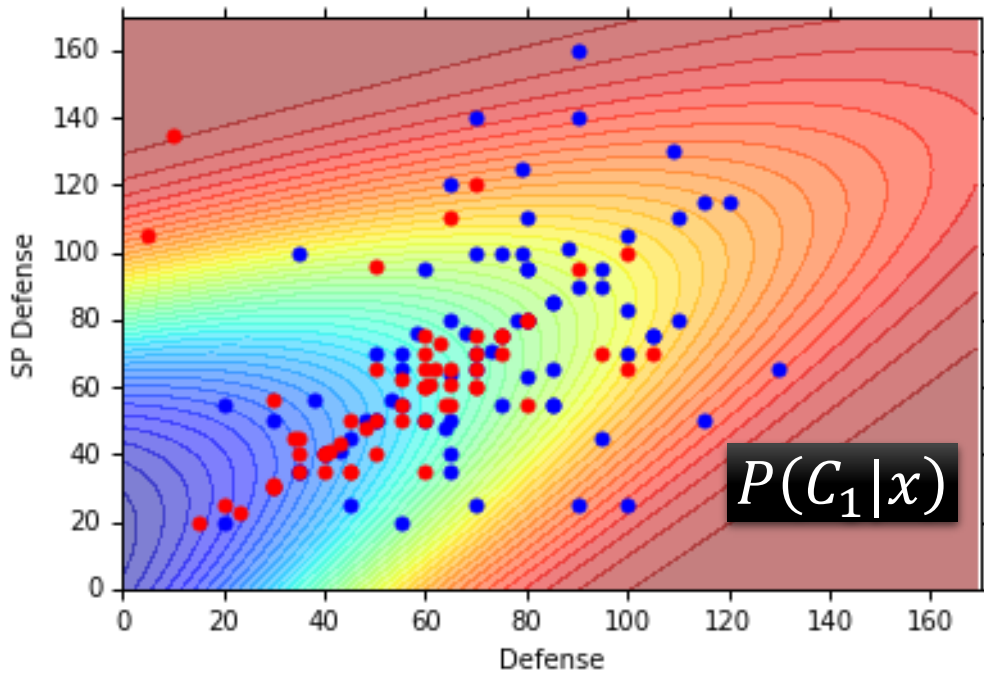
$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

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

$P(C_2) = 61 / (79 + 61) = 0.44$

$$\mu^2 = \begin{bmatrix} 55.6 \\ 59.8 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 847 & 422 \\ 422 & 685 \end{bmatrix}$$

If $P(C_1|x) > 0.5$  x belongs to class 1 (Water)



Blue points: C_1 (Water), Red points: C_2 (Normal)

How's the results?

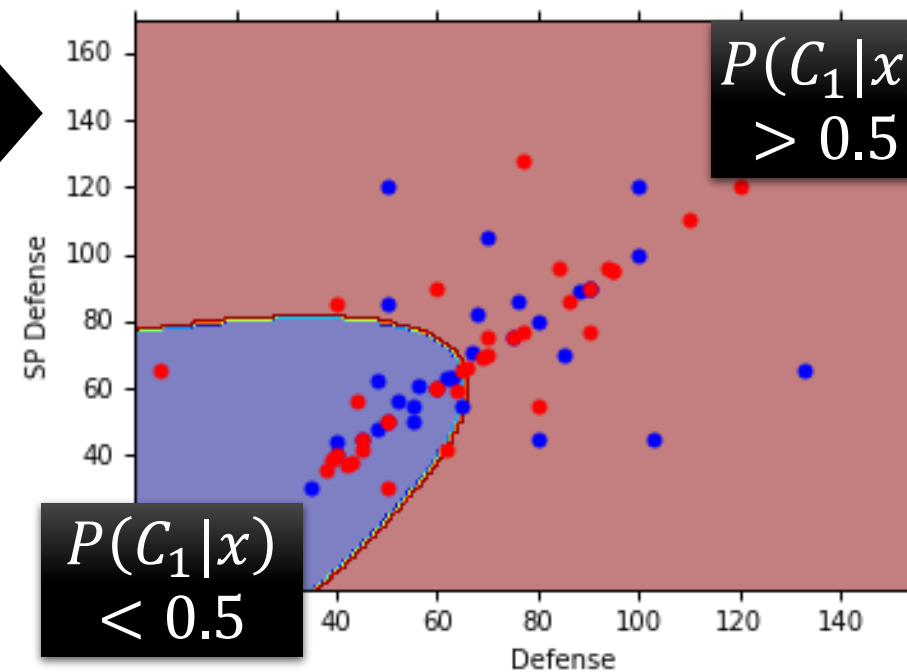
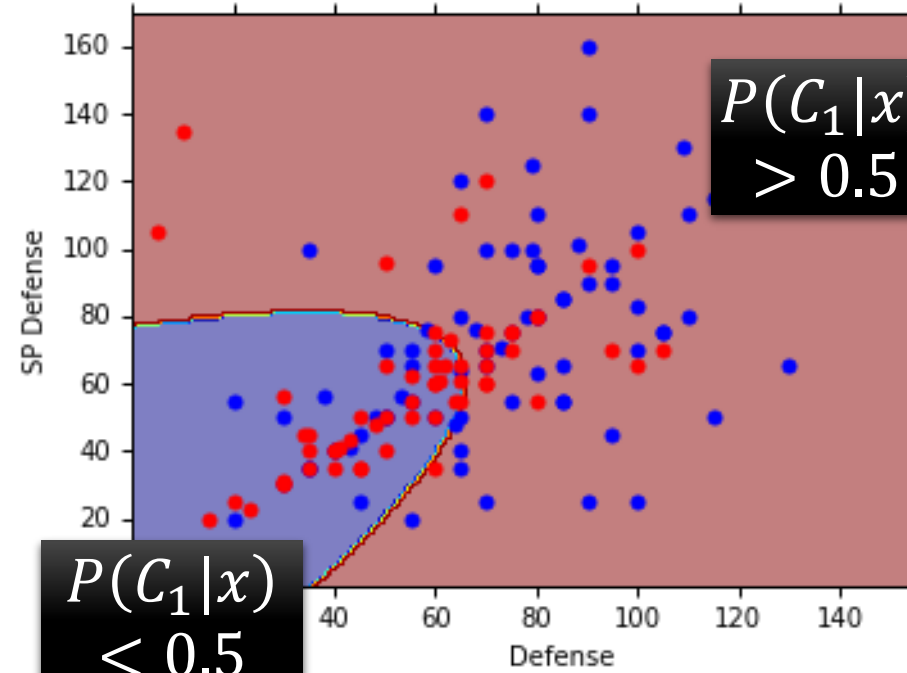
Testing data: 47% accuracy ☹️

All: hp, att, sp att,
de, sp de, speed (6 features)

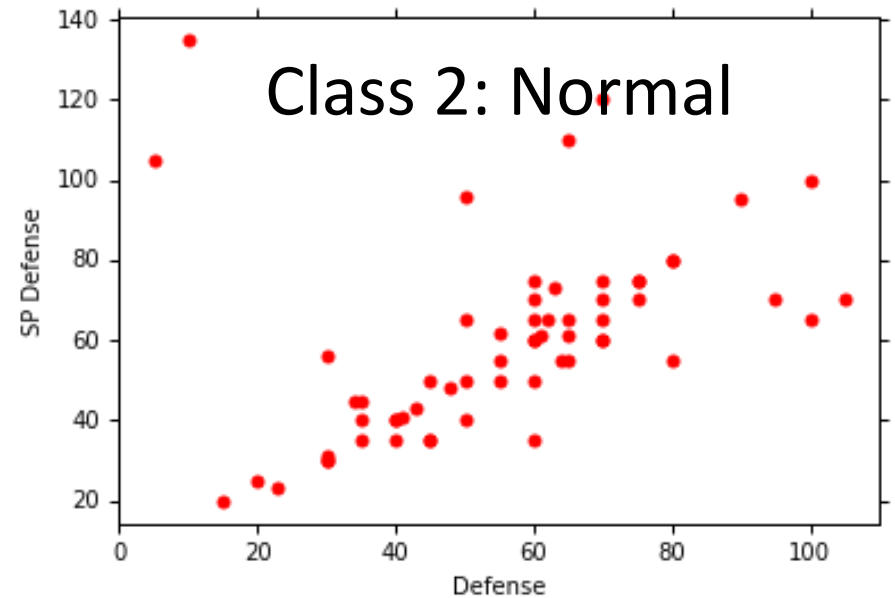
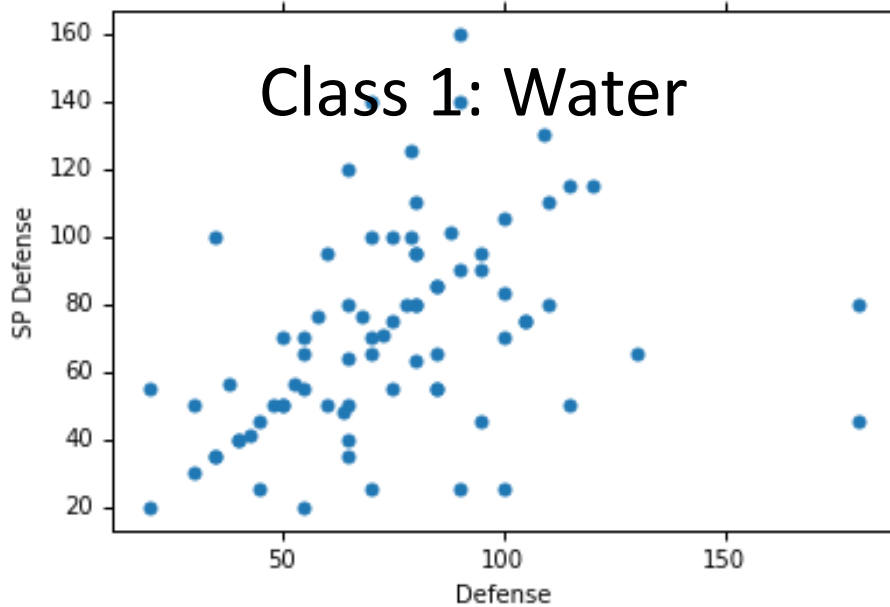
μ^1, μ^2 : 6-dim vector

Σ^1, Σ^2 : 6 x 6 matrices

64% accuracy ...



Modifying Model



$$\mu^1 = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix} \quad \Sigma^1 = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

$$\mu^2 = \begin{bmatrix} 55.6 \\ 59.8 \end{bmatrix} \quad \Sigma^2 = \begin{bmatrix} 847 & 422 \\ 422 & 685 \end{bmatrix}$$

The same Σ

Less parameters

Modifying Model

Ref: Bishop,
chapter 4.2.2

- Maximum likelihood

“Water” type Pokémons:

$x^1, x^2, x^3, \dots, x^{79}$

μ^1

“Normal” type Pokémons:

$x^{80}, x^{81}, x^{82}, \dots, x^{140}$

μ^2

Σ

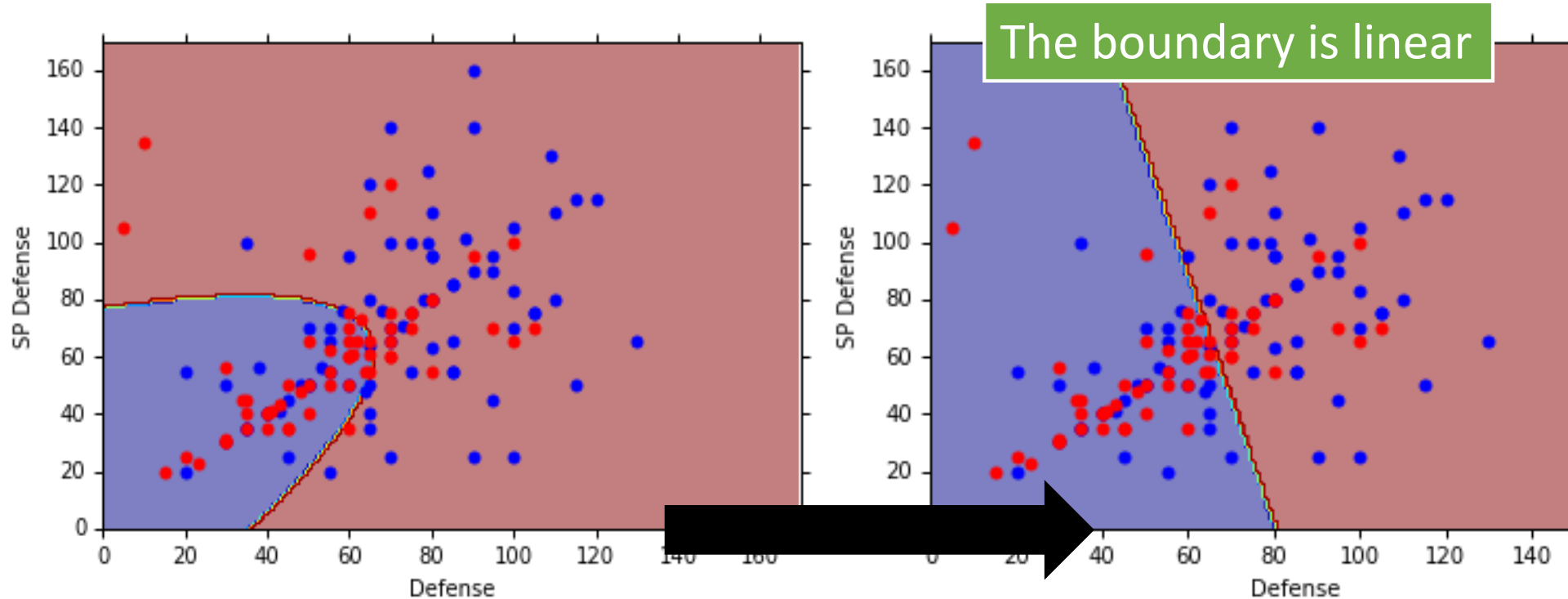
Find μ^1, μ^2, Σ maximizing the likelihood $L(\mu^1, \mu^2, \Sigma)$

$$L(\mu^1, \mu^2, \Sigma) = f_{\mu^1, \Sigma}(x^1) f_{\mu^1, \Sigma}(x^2) \cdots f_{\mu^1, \Sigma}(x^{79}) \\ \times f_{\mu^2, \Sigma}(x^{80}) f_{\mu^2, \Sigma}(x^{81}) \cdots f_{\mu^2, \Sigma}(x^{140})$$

μ^1 and μ^2 is the same

$$\Sigma = \frac{79}{140} \Sigma^1 + \frac{61}{140} \Sigma^2$$

Modifying Model



The same covariance matrix

All: hp, att, sp att, de, sp de, speed

54% accuracy → 73% accuracy

Three Steps

- Function Set (Model):

x 

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

If $P(C_1|x) > 0.5$, output: class 1
Otherwise, output: class 2

- Goodness of a function:
 - The mean μ and covariance Σ that maximizing the likelihood (the probability of generating data)
- Find the best function: easy

Probability Distribution

- You can always use the distribution you like 😊

$$P(x|C_1) = P(x_1|C_1) P(x_2|C_1) \cdots P(x_k|C_1) \cdots$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \\ x_K \end{bmatrix}$$

1-D Gaussian

For binary features, you may assume they are from Bernoulli distributions.

If you assume all the dimensions are independent, then you are using *Naive Bayes Classifier*.

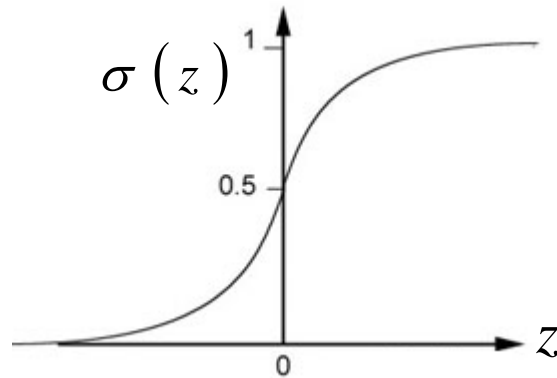
Posterior Probability

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{1}{1 + \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}} = \frac{1}{1 + \exp(-z)} = \sigma(z)$$

Sigmoid function

$$z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$



Warning of Math

Posterior Probability

$$P(C_1|x) = \sigma(z) \quad \text{sigmoid} \quad z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

$$z = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{P(C_1)}{P(C_2)} \rightarrow \frac{\frac{N_1}{N_1 + N_2}}{\frac{N_2}{N_1 + N_2}} = \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) \right\}$$

$$P(x|C_2) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^2|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \right\}$$

$$z = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{P(C_1)}{P(C_2)} = \frac{N_1}{N_2}$$

$$P(x|C_1) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) \right\}$$

$$P(x|C_2) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^2|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \right\}$$

$$\ln \frac{\cancel{\frac{1}{(2\pi)^{D/2}}} \frac{1}{|\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) \right\}}{\cancel{\frac{1}{(2\pi)^{D/2}}} \frac{1}{|\Sigma^2|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2) \right\}}$$

$$= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} \exp \left\{ -\frac{1}{2} [(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) - (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2)] \right\}$$

$$= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2} [(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1) - (x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2)]$$

$$z = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{P(C_1)}{P(C_2)} = \frac{N_1}{N_2}$$

$$= \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2} [\underbrace{(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1)}_{\text{red}} - \underbrace{(x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2)}_{\text{red}}]$$

$$(x - \mu^1)^T (\Sigma^1)^{-1} (x - \mu^1)$$

$$= x^T (\Sigma^1)^{-1} x - \underbrace{x^T (\Sigma^1)^{-1} \mu^1 - (\mu^1)^T (\Sigma^1)^{-1} x}_{\text{blue}} + (\mu^1)^T (\Sigma^1)^{-1} \mu^1$$

$$= x^T (\Sigma^1)^{-1} x - \underbrace{2(\mu^1)^T (\Sigma^1)^{-1} x}_{\text{blue}} + (\mu^1)^T (\Sigma^1)^{-1} \mu^1$$

$$(x - \mu^2)^T (\Sigma^2)^{-1} (x - \mu^2)$$

$$= x^T (\Sigma^2)^{-1} x - 2(\mu^2)^T (\Sigma^2)^{-1} x + (\mu^2)^T (\Sigma^2)^{-1} \mu^2$$

$$z = \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2} x^T (\Sigma^1)^{-1} x + (\mu^1)^T (\Sigma^1)^{-1} x - \frac{1}{2} (\mu^1)^T (\Sigma^1)^{-1} \mu^1 \\ + \frac{1}{2} x^T (\Sigma^2)^{-1} x - (\mu^2)^T (\Sigma^2)^{-1} x + \frac{1}{2} (\mu^2)^T (\Sigma^2)^{-1} \mu^2 + \ln \frac{N_1}{N_2}$$

End of Warning

$$P(C_1|x) = \sigma(z)$$

$$z = \cancel{\ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}}} - \cancel{\frac{1}{2} x^T (\Sigma^1)^{-1} x} + (\mu^1)^T (\Sigma^1)^{-1} x - \frac{1}{2} (\mu^1)^T (\Sigma^1)^{-1} \mu^1 \\ + \cancel{\frac{1}{2} x^T (\Sigma^2)^{-1} x} - (\mu^2)^T (\Sigma^2)^{-1} x + \frac{1}{2} (\mu^2)^T (\Sigma^2)^{-1} \mu^2 + \ln \frac{N_1}{N_2}$$

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$z = \underbrace{(\mu^1 - \mu^2)^T \Sigma^{-1} x}_{\mathbf{w}^T} - \underbrace{\frac{1}{2} (\mu^1)^T \Sigma^{-1} \mu^1 + \frac{1}{2} (\mu^2)^T \Sigma^{-1} \mu^2}_{b} + \ln \frac{N_1}{N_2}$$

$$P(C_1|x) = \sigma(\mathbf{w} \cdot x + b) \quad \text{How about directly find } \mathbf{w} \text{ and } b?$$

In generative model, we estimate $N_1, N_2, \mu^1, \mu^2, \Sigma$

Then we have \mathbf{w} and b

Reference

- Bishop: Chapter 4.1 – 4.2
- Data: <https://www.kaggle.com/abcsds/pokemon>
- Useful posts:
 - <https://www.kaggle.com/nishantbhadauria/d/abcsds/pokemon/pokemon-speed-attack-hp-defense-analysis-by-type>
 - <https://www.kaggle.com/nikos90/d/abcsds/pokemon/mastering-pokebars/discussion>
 - <https://www.kaggle.com/ndrewgele/d/abcsds/pokemon/visualizing-pok-mon-stats-with-seaborn/discussion>

Acknowledgment

- 感謝 江貫榮 同學發現課程網頁上的日期錯誤
- 感謝 范廷瀚 同學提供寶可夢的 domain knowledge
- 感謝 Victor Chen 發現投影片上的打字錯誤