Probabilistic Reasoning and Bayes' Theorem
Fundamentals of AI (AE1FAI)

Slides created by Dr. Rong Qu， Prof. Hyung-Yi Lee
Modified by Dr. Huan Jin

Outline

- Probability Theory
  - Overview
  - Disjoint, Independence
- Bayesian Theorem
  - From Joint Distribution
  - Using Independence / Factoring
  - From Sources of Evidence
- Naive Bayes Learner

Basic Concepts

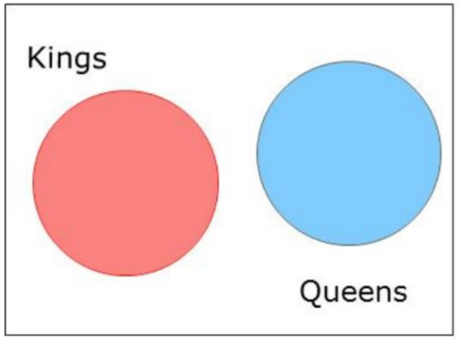
- A probability model is a mathematical representation, defined by its sample space S, events A within the sample space, and probabilities P(A), which is a numerical value describing its long-tun relative frequency.
- Example: Give a probability model for chance process of tossing of a coin.

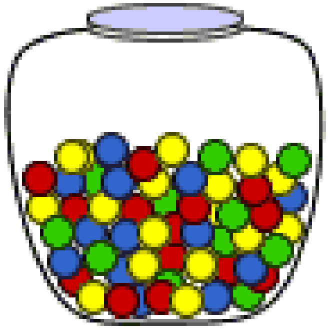
  Sample space, S = { Head, Tail }

  Each of these outcomes has probability ½
- First two basic rules of probability are:-
  - **Rule 1: Any probability P(A) is number between 0 and 1 (0 ≤ P(A) ≤ 1)**
  - **Rule 2: The probability of the sample space S is always equal to 1 (P(S) = 1)**

Basic Concepts

- If two events have no outcome in common, they are called disjoint.
  - **Rule 3: If two events A and B are disjoint, the probability of either event is the sum of probabilities of two events P(A or B) = P(A) + P(B)**
- The chance of any (one or more) of two more events occurring is called the union of events. The probability of union of disjoint events is the sum of their individual probability.
  - E.g., The probability to get yellow and red: ¼ + ¼ = ½





Basic Concepts

- If there are 3 red and 2 blue marbles, probability of drawing red is 3/5 = 0.6 whereas probability of drawing blue is 2/5 = 0.4. Given sample space is { red, blue }, and the total probability is always equal to 1. The event of drawing blue marble is same as the event red marble was not drawn (complement of event).
  - **Rule 4: $P(A^c) = 1 – P(A)$**
- Consider the coin flipping event, flip the coin two times, what the probability of getting “head”s. The outcome of the 1st event (1st flip) has no effect on the probability of the 2nd event (2nd flip), then the two events are independent.
  - **Rule 5: If two events are independent, then the probability of both events happening is the product of probabilities of each event: P(A and B) = P(A)P(B)**

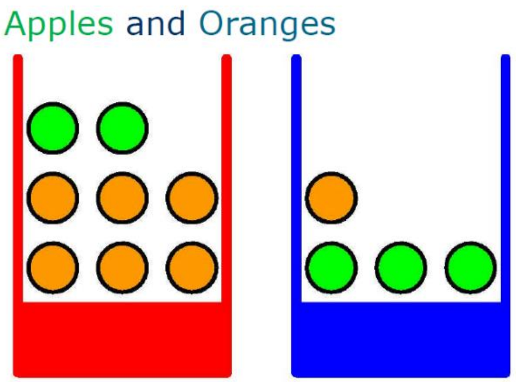
Basic Concepts

- The chance of two events occurring is called the intersection of events. For independent events, the probability of the intersection of two or more events is the product of the probabilities
  - E.g., observing coin flip to get 2 heads is ½ * ½ = ¼
  - E.g., observing coin flip to get 4 heads is ½ * ½ * ½ * ½ = 1/16

Probability Theory

- One of the boxes is randomly picked and item of fruit selected from that box. The fruit is replaced back in the same box after observation.
- Let B be the random variable denoting the identity of the box and F the random variable indicating the identity of the fruit drawn.
- If P(B=redbox) = 4/10 and P(B=bluebox) = 6/10.

1) What is the overall probability that this selection procedure will pick an apple?
2) Given that we have chosen an orange, what is the probability that the box chosen is the blue one?



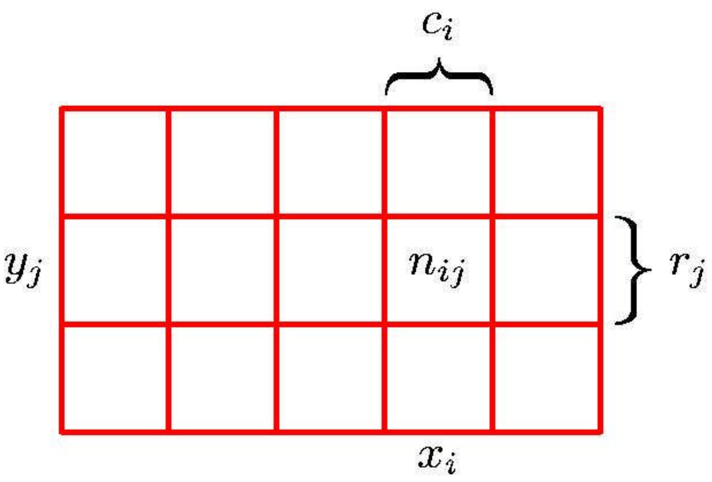
Probability theory

Suppose X can take any of the values x_i where i = 1,…., M and Y can take the values y_j where j = 1,…., L. After a total of N trials in which both variables X and Y are sampled.

If the number of trials in which X = x_i and Y = y_j is n_{ij} then:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- Joint Probability



Let the number of trials in which X takes the value x_i , irrespective of the value that Y takes, be denote by c_i and similarly the number of trials in which Y takes the value y_j be denoted by r_j .

The probability that X takes the value x_i irrespective of the value of Y is:

$$p(X = x_i) = \frac{c_i}{N}.$$

- Marginal Probability (prior)

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

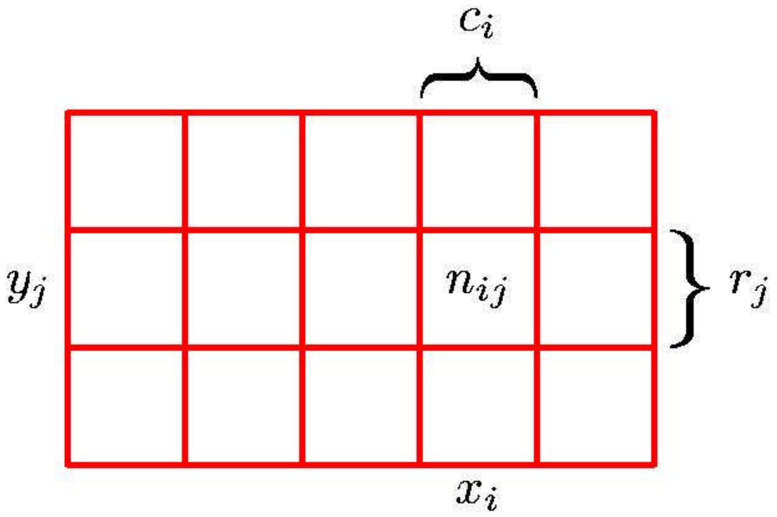
Pattern Recognition and Machine Learning by Christopher Bishop

Probability theory

If we consider only those instances for which X = x_i then the fraction of such instances for which Y = y_j is:

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

- Conditional Probability (posterior)



The Rules of Probability:

Product Rule:

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

$$p(X, Y) = p(Y|X)p(X)$$

Sum Rule:

$$p(X) = \sum_Y p(X, Y)$$

Bayes' Theorem

- Product Rule:

$$p(X, Y) = p(Y|X)p(X)$$

- Together with the symmetry property p(X, Y) = p(Y, X), the relationship between conditional probabilities can be derived, known as Bayes' theorem.

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

posterior ∝ likelihood × prior

- Using the sum rule, the denominator can be expressed in terms of the quantities appearing in the numerator.

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

exercise

- Given:
  - P(r) = 4/10; P(b) = 6/10
  - P(a|r) = 2/8; R(o/r) = 6/8
  - P(a|b) = 3/4; P(o/b) = 1/4

P(a|b) obeys the same rules as probabilities, i.e., P(a | b) + P(NOT(a) | b) = 1

e.g. P(a|r) + P(o|r) = 1
and P(a|b) + P(o|b) = 1

- Joint probability
  - P(a, b) = P(a|b) * P(b) = 3/4 * 6/10 = 9/20
  - P(a, r) = P(a|r) * P(r) = 2/8 * 4/10 = 2/20
  - P(o, b) = P(o|b) * P(b) = 1/4 * 6/10 = 3/20
  - P(o, r) = P(o|r) * P(r) = 6/8 * 4/10 = 6/20

  P(a, b) + P(a, r) + P(o, b) + P(o, r) = 1

Joint Probability Table

	Box = b	Box = r
Fruit = a	9/20	2/20
Fruit = o	3/20	6/20

- Prior probability $p(X) = \sum_{Y} p(X, Y)$
  - P(a) = P(a, b) + P(a, r)
    = 11/20

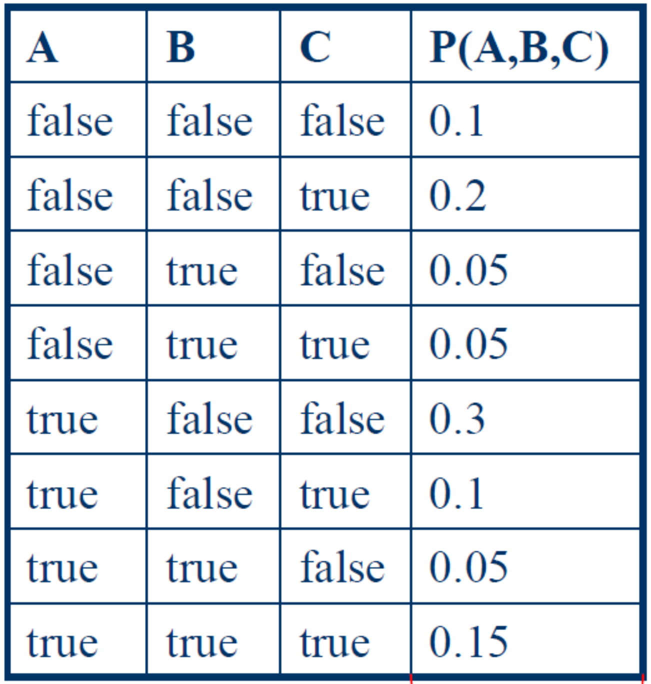
- Conditional probability
  - P(b|o) = P(o|b) * P(b)/P(o)
    = (1/4 * 6/10)/(9/20) = 1/3
  - P(r|o) = 1 – 1/3 = 2/3

Random Variables

- A random variable is the basic element of probability, representing an event with some degree of uncertainty as to the event's outcome
- Let's start with simplest type of random variables, Boolean, which take only **true** or **false**, indicating event occurring or not.
- Examples (Let A be a Boolean random variable):
  - A = getting heads on a coin flip
  - A = it will rain tomorrow
  - A = Mary has influenza
  - A = The US president in 2024 will be female

The Joint Probability distribution

- Joint probabilities can be between any number of variables
- For each combination of variables, we can show how probable that combination is
- The probabilities of these combinations need to sum to 1

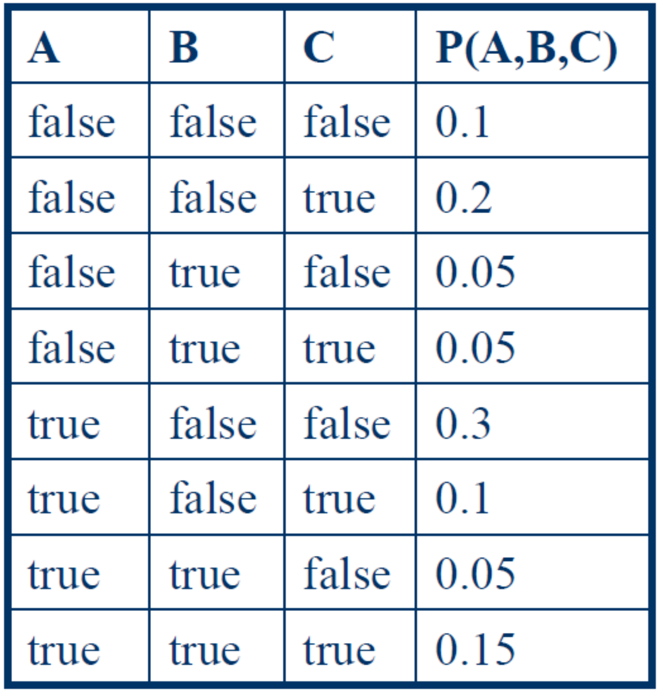


A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Sums to 1

The Joint Probability distribution

- Once there is a joint probability distribution, it is possible to calculate any probability that involved A, B and C.
  - May have to use marginalization and Bayes' rule



A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

- Example:
  - P(A=true) = sum of P(A,B,C) in rows with A = true
  - P(A=true,B=true|C=true) = P(A=true,B=true,C=true)/P(C=true)

Bayes' Rule

- Bayes' rule is derived from the product rule:
  - P(Y | X) = P(X | Y) P(Y) / P(X) where

  P(Y | X) is the probability that hypothesis Y is true given evidence X

  P(X | Y) is the probability observe X given hypothesis Y

  P(Y) is prior probability that hypothesis Y is true without any evidence.

- In general form:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y) + P(X|\neg Y)P(\neg Y)}$$

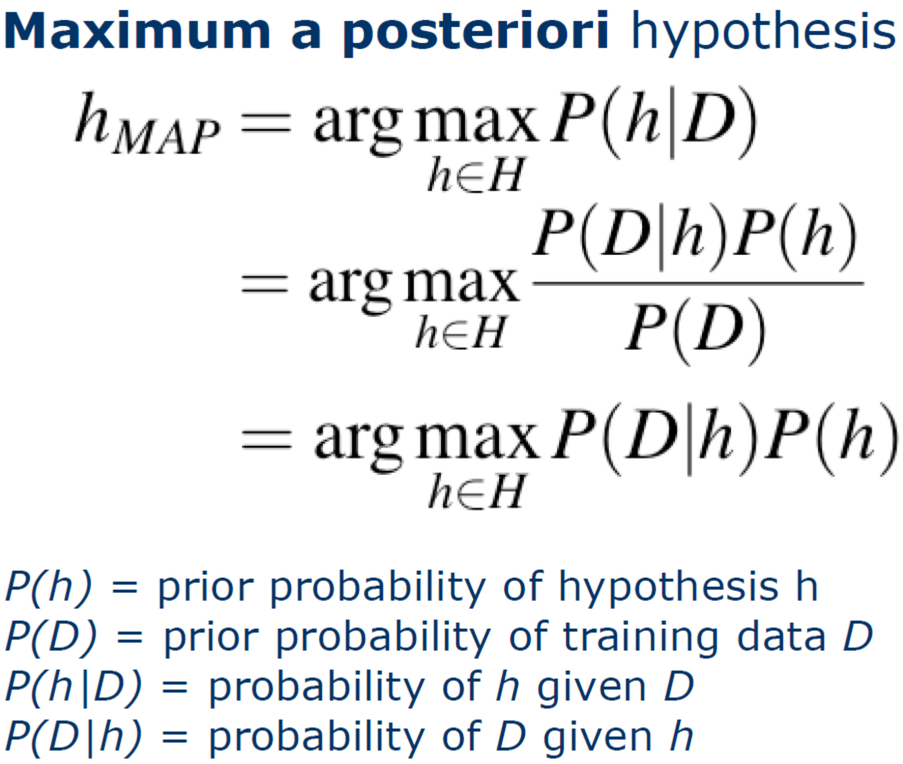
Example-1

- A child has rash
- A doctor knows
  - 10% of sick children have flu, and 3% of children has rash
  - 5% of children with flu develop a rash
  - Has the child got a flu?
- Diagnostic hypothesis
  - P(Y) – probability of children has flu = 0.01
  - P(X) – probability of children has rash = 0.03
  - P(X|Y) – probability observe evidence of rash given children has flue = 0.05
  - P(Y|X) – probability that children has flu given evidence of rash

    = 0.05 * 0.1 / 0.03 = 0.17

Choosing Hypothesis

- Generally want the most probable hypothesis given the training data.



Maximum a posteriori hypothesis h_{MAP}:

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

$$= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

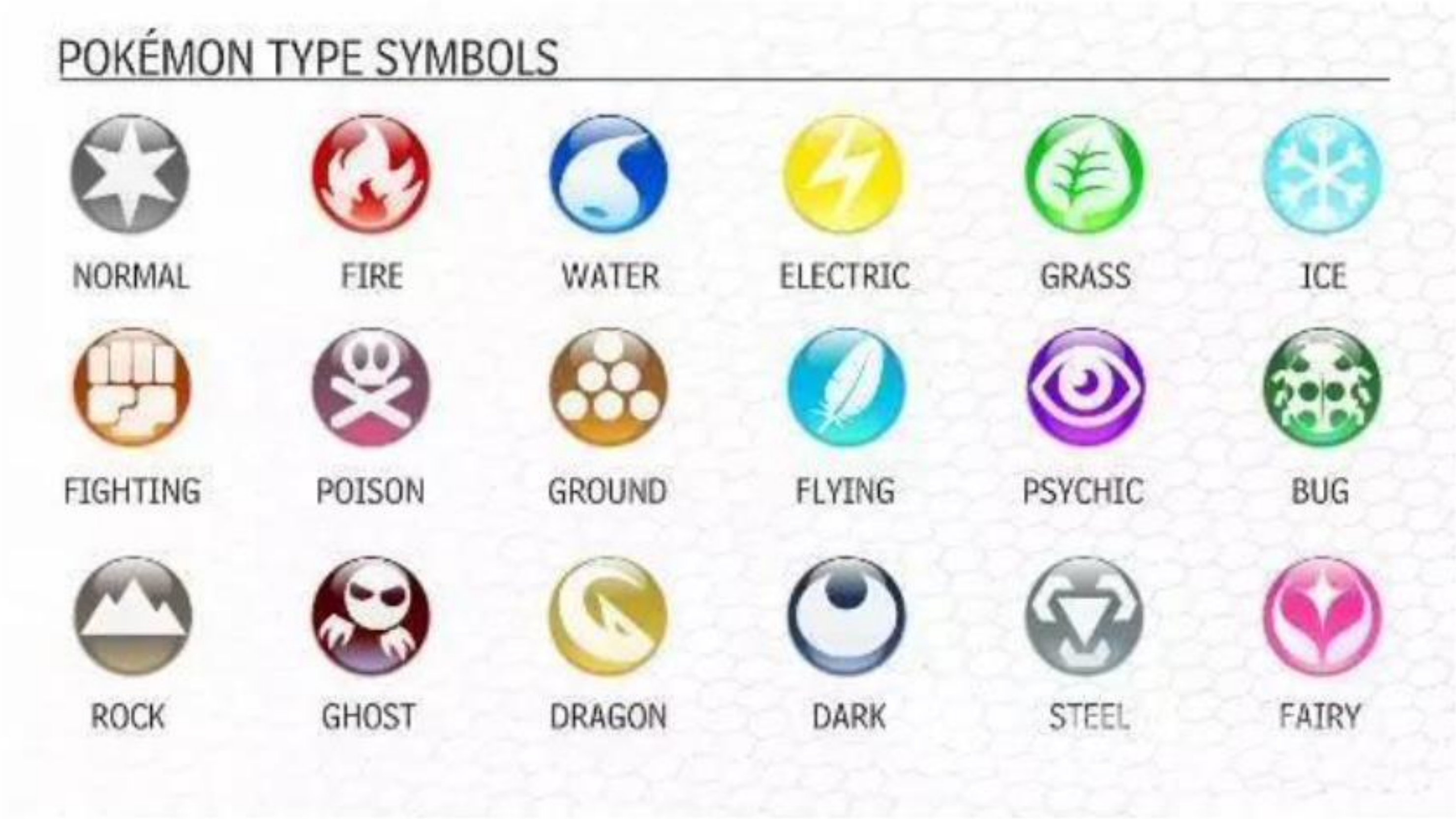
$$= \arg\max_{h \in H} P(D|h)P(h)$$

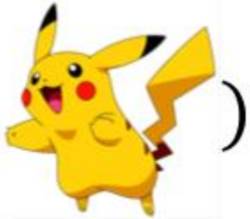
P(h) = prior probability of hypothesis h
P(D) = prior probability of training data *D*
P(h|D) = probability of *h* given *D*
P(D|h) = probability of *D* given *h*

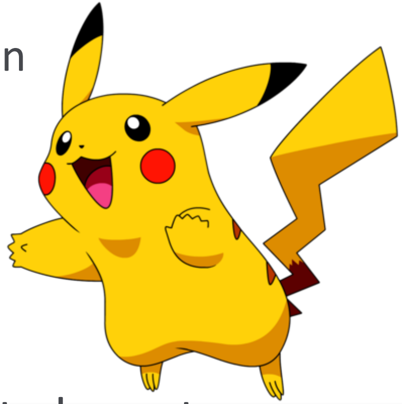
Outline

- Probability Theory
  - Overview
  - Disjoint, Independence
  - Sum and Product Rules
- Bayesian Theorem
  - From Joint Distribution
  - Using Independence/Factoring
  - Simple and General Form
- NAIVE BAYES LEARNER

Example Application



$f($  $) =$ $f($ $) =$ $f($ $) =$ pokemon games (*NOT* pokemon cards or Pokemon Go)



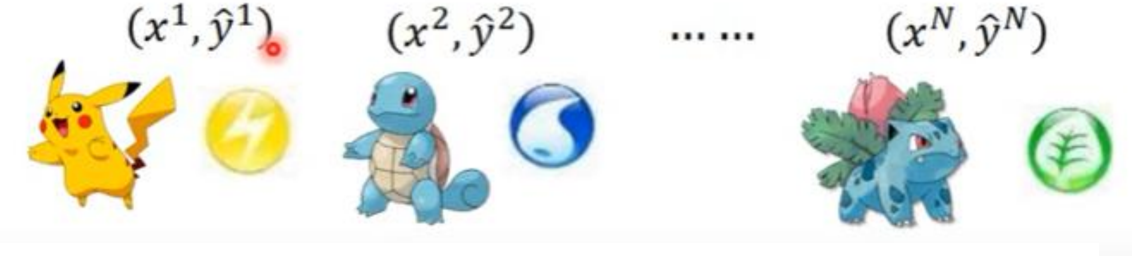
Example Application

- **Total**: sum of all stats that come after this, a general guide to how strong a pokemon is 320
- **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, Punch) 55
- **Defense**: the base damage resistance against normal attacks 40
- **SP Atk**: special attack, the base modifier for special attacks (e.g. fire blast, bubble beam) 50
- **SP Def**: the base damage resistance against special attacks 50
- **Speed**: determines which pokemon attacks first each round 90

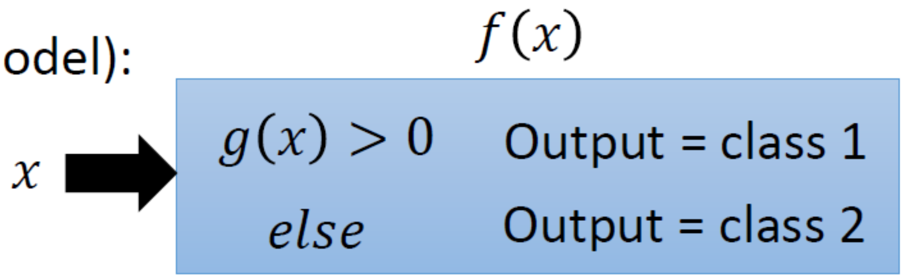
Can we predict the "type" of pokemon based on the information?

How to do Classification

- Training data for classification



- Function (Model):



$f(x)$

$x \rightarrow$

$g(x) > 0$	Output = class 1
$else$	Output = class 2

- Loss function:

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

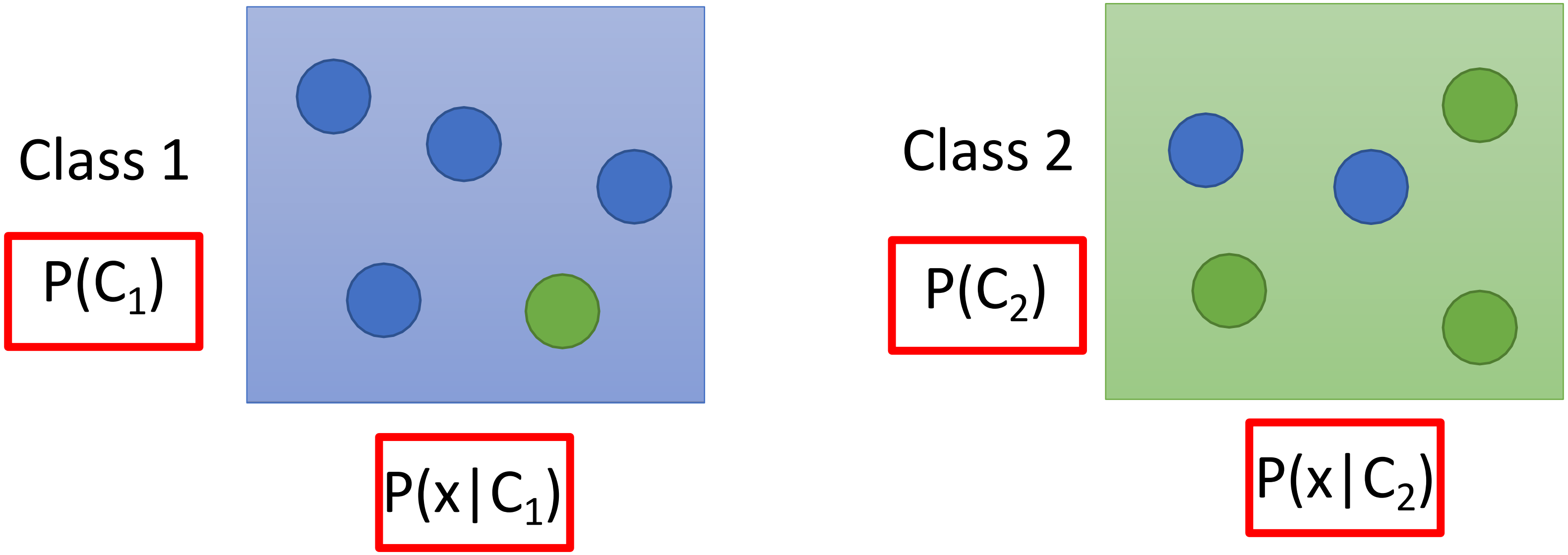
The number of times f get incorrect results on training data.

- Find the best function:
  - Example: Perceptron, SVM

Not Today

Two Classes

Estimating the Probabilities
From training data



Class 1

P(C$_1$)

P(x|C$_1$)

Class 2

P(C$_2$)

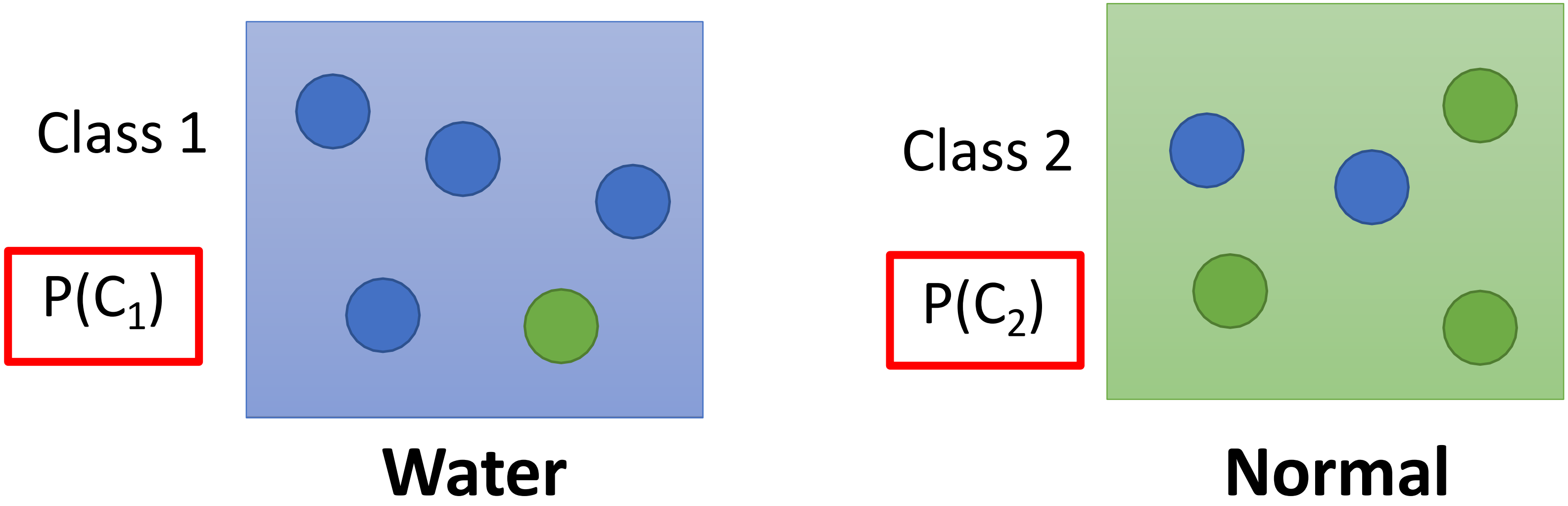
P(x|C$_2$)

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



Class 1

$P(C_1)$

Water

Class 2

$P(C_2)$

Normal

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

$P(x|C_1) = ?$ P($|$Water) = ?

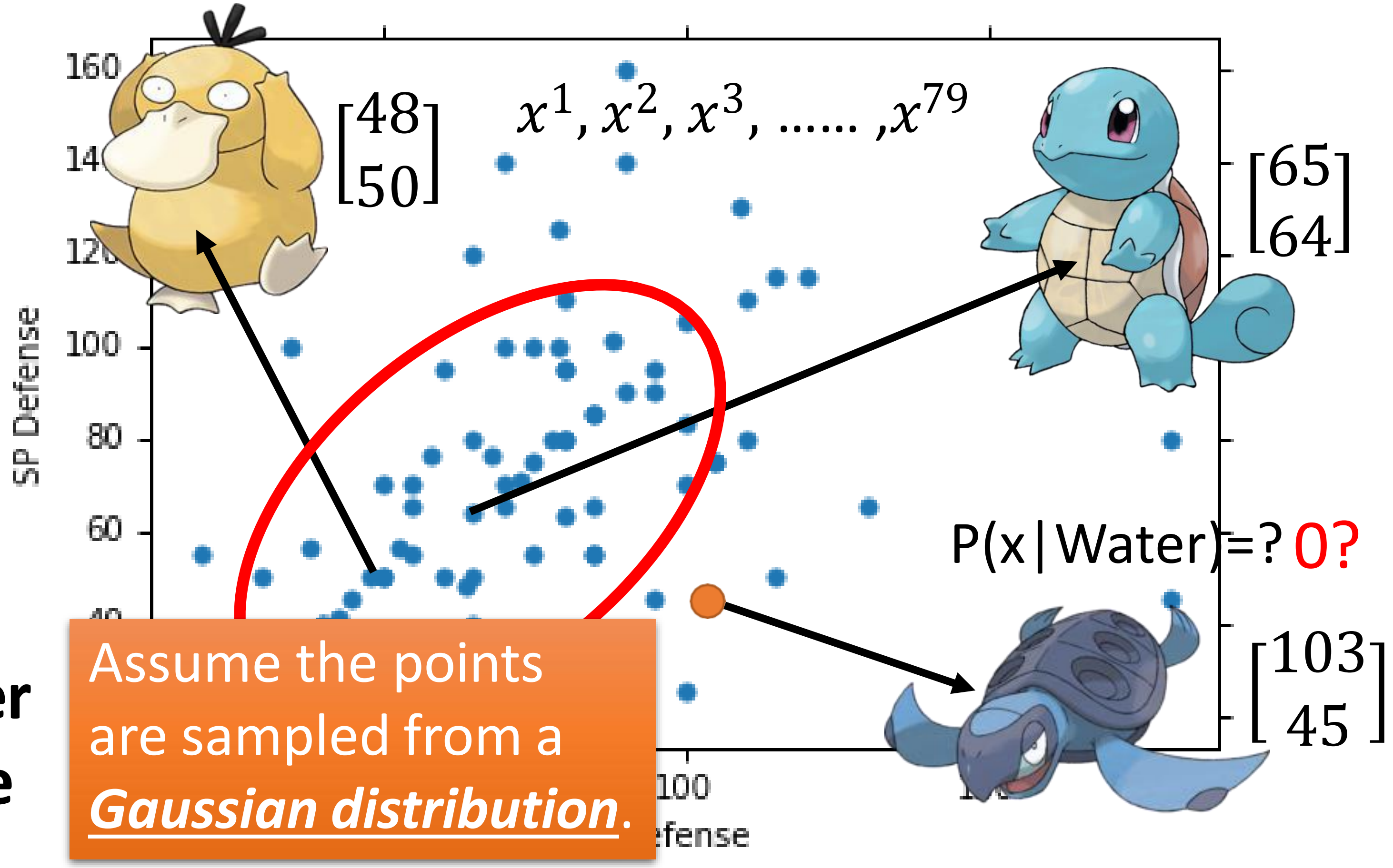
Each Pokémon is represented as a vector by its attribute. ⇨ feature

Water Type

79 in total

Probability from Class - Feature

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



Assume the points are sampled from a ***Gaussian distribution***.

Water Type

https://blog.slinuxer.com/tag/pca

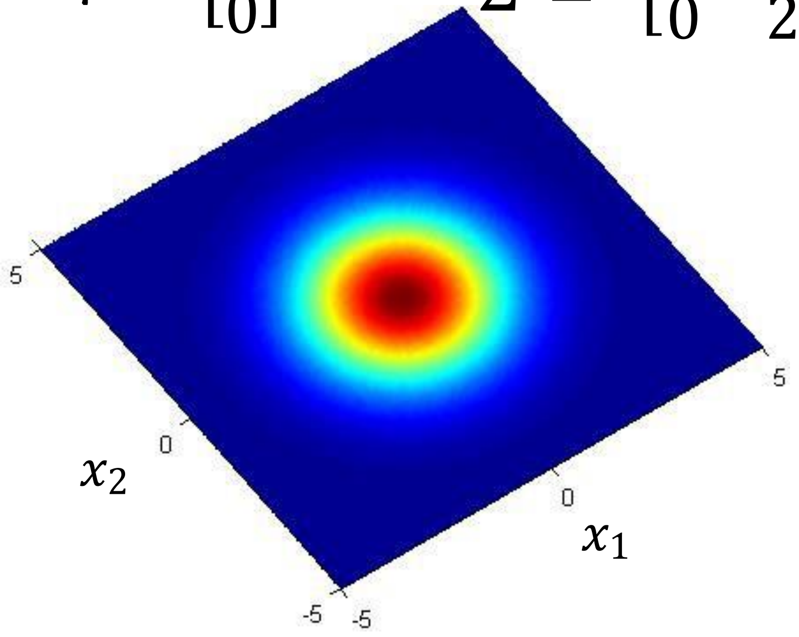
Gaussian Distribution

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

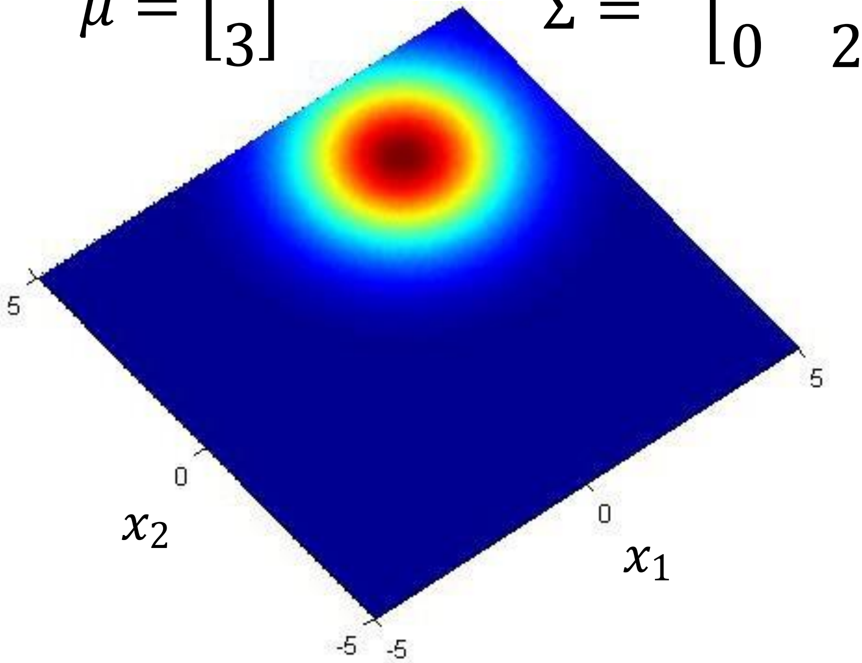
Input: vector x, output: probability of sampling x

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

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$



$\mu = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$



https://blog.slinuxer.com/tag/pca

Gaussian Distribution

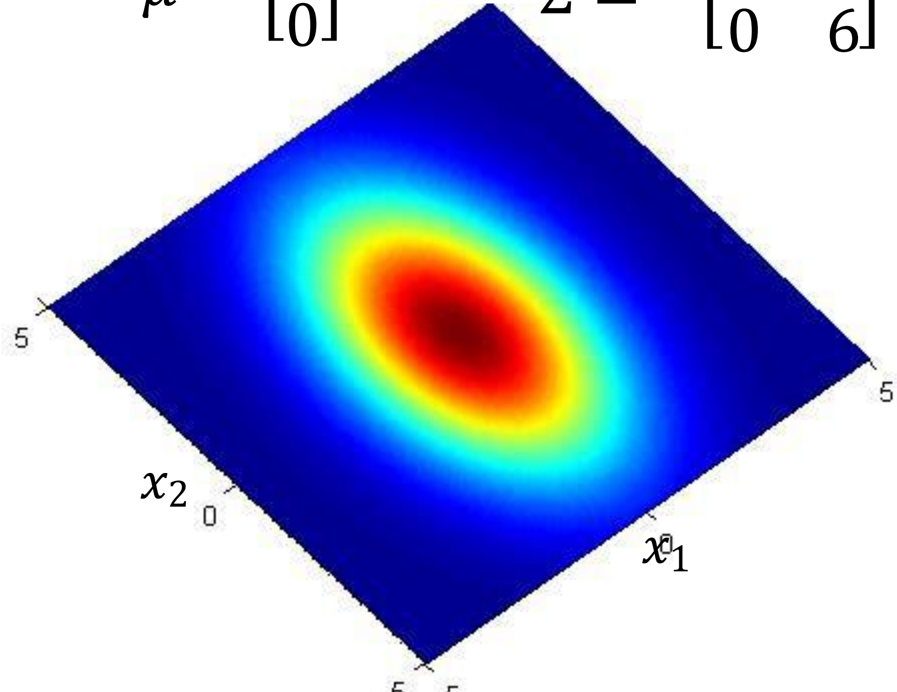
$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

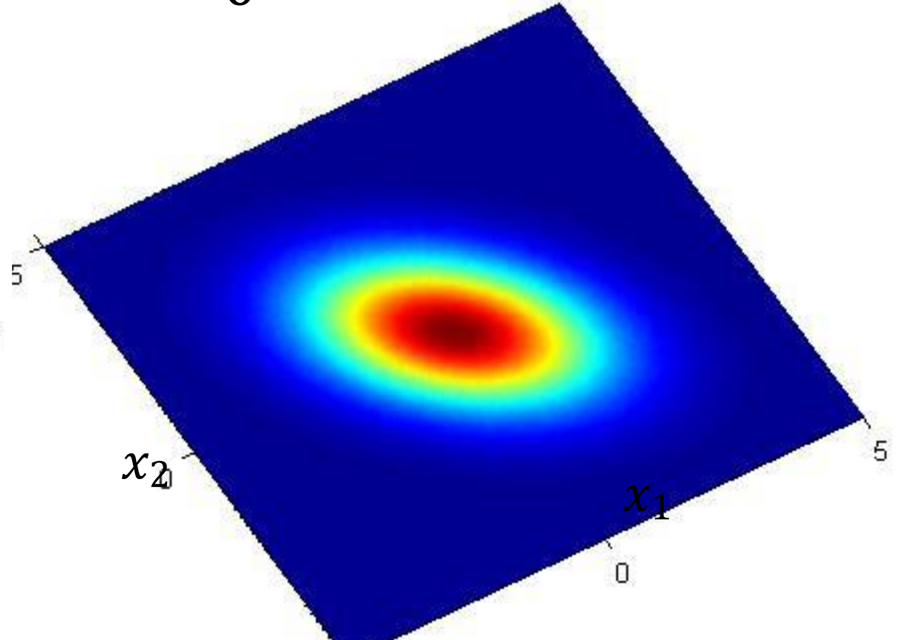
Input: vector x, output: probability of sampling x

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

$$\mu = \begin{bmatrix}0\\0\end{bmatrix} \qquad \Sigma = \begin{bmatrix}2 & 0\\0 & 6\end{bmatrix}$$

$$\mu = \begin{bmatrix}0\\0\end{bmatrix} \qquad \Sigma = \begin{bmatrix}2 & -1\\-1 & 2\end{bmatrix}$$

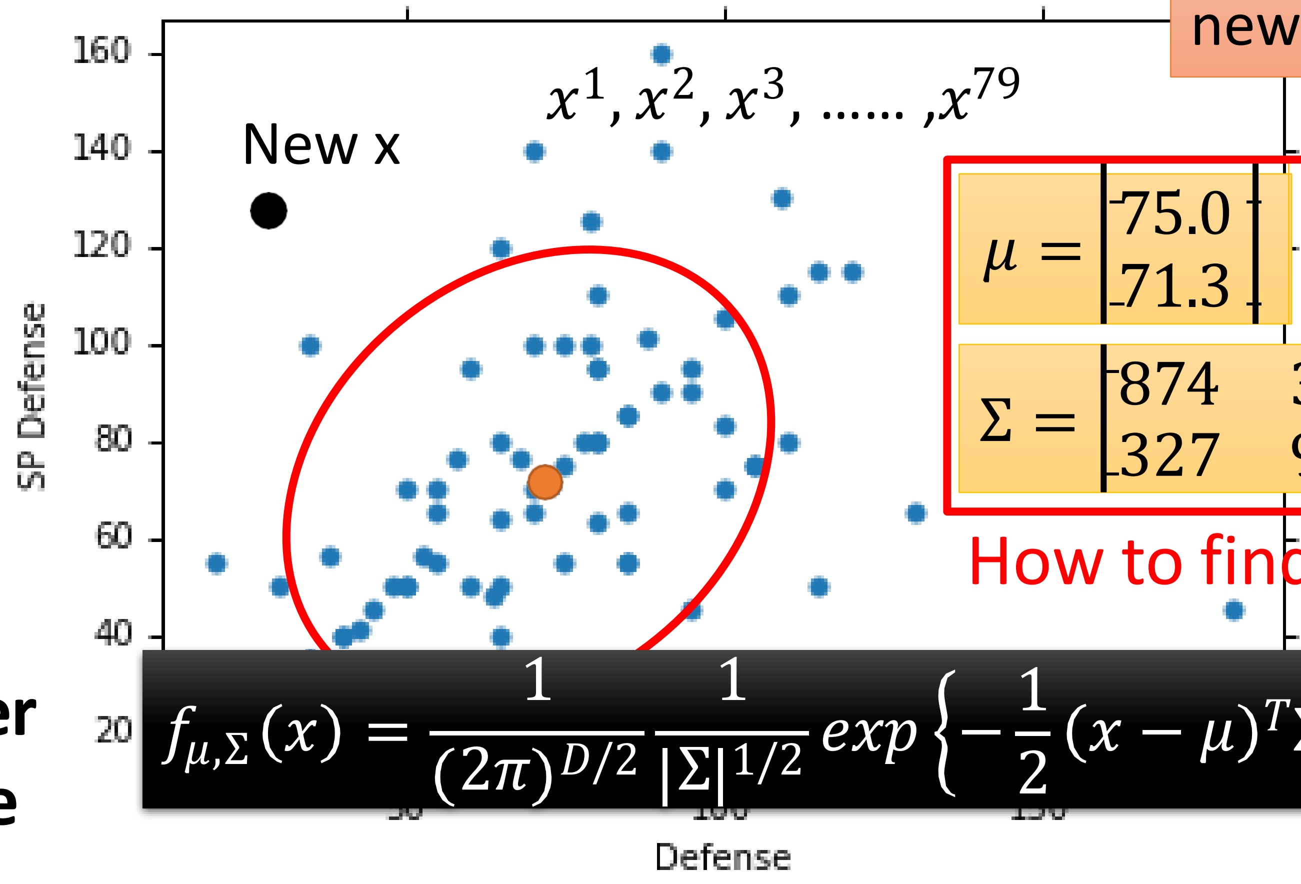




Probability from Class

Assume the points are sampled from a Gaussian distribution

Find the Gaussian distribution behind them → Probability for new points



$$\mu = \begin{bmatrix} 75.0 \\ 71.3 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 874 & 327 \\ 327 & 929 \end{bmatrix}$$

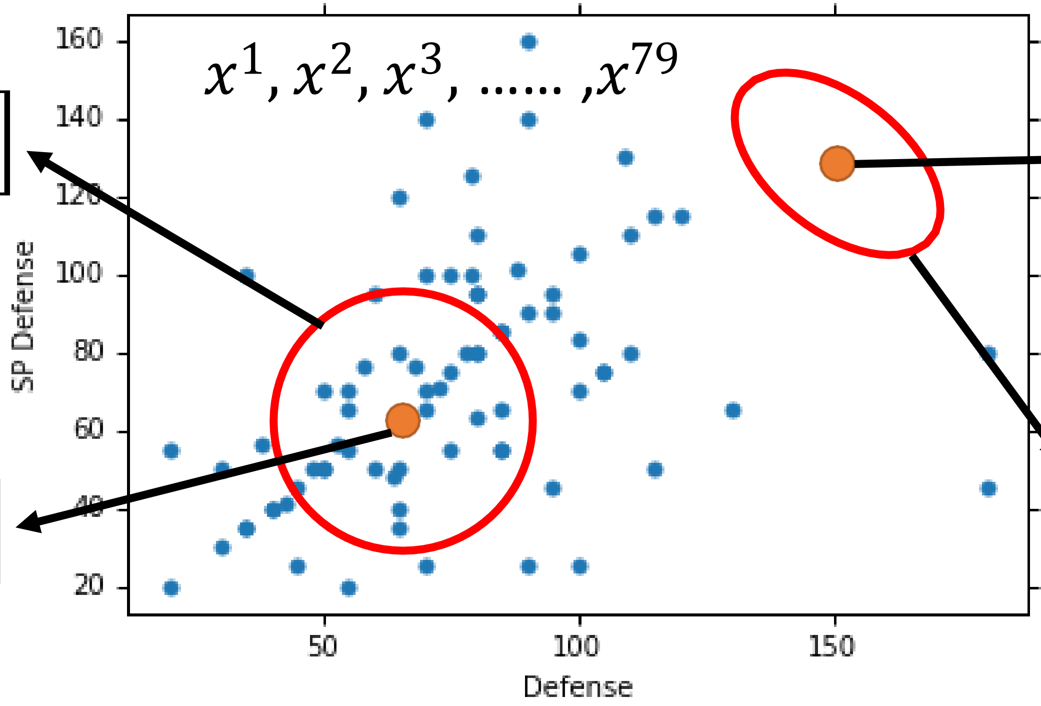
How to find them?

Water Type

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

Maximum Likelihood

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



⍰ $= \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$

$\mu = \begin{bmatrix} 75 \\ 75 \end{bmatrix}$

$\mu = \begin{bmatrix} 150 \\ 140 \end{bmatrix}$

⍰ $= \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$

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

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

Maximum Likelihood

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

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

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

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

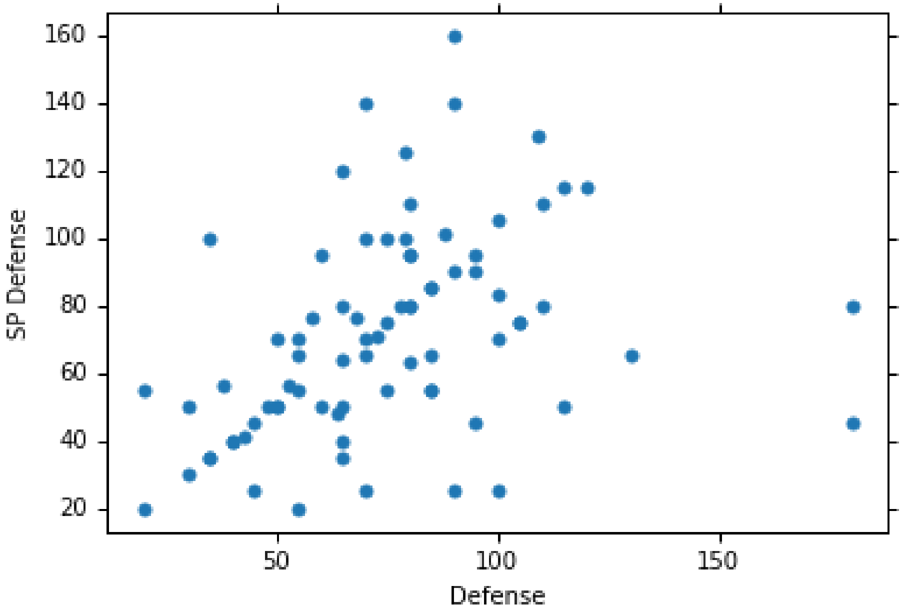
$$\mu^*, \Sigma^* = arg \max_{\mu,\Sigma} L(\mu, \Sigma)$$

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

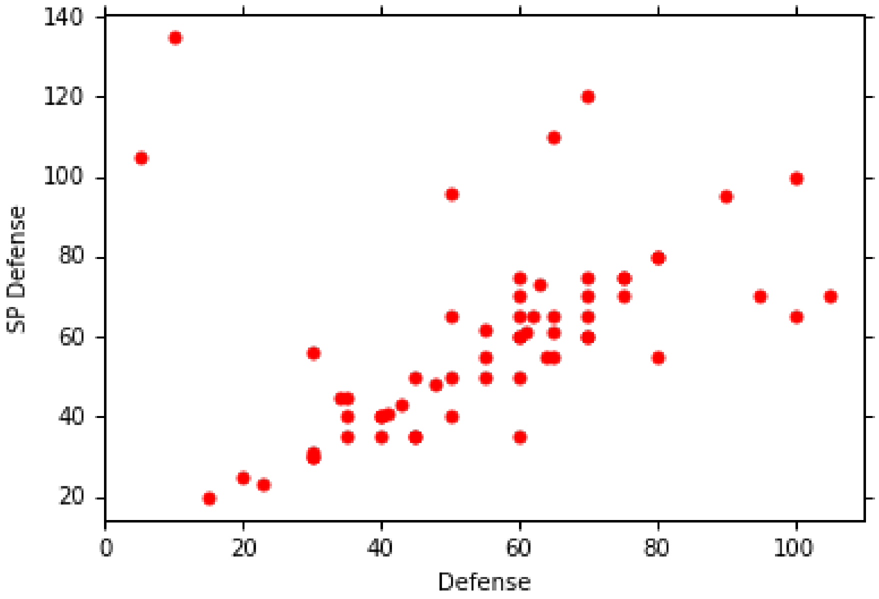
average

Maximum Likelihood

Class 1: Water



Class 2: Normal



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

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

Now we can do classification ☺

$$f_{\mu^1,\Sigma^1}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^1|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_1)^T(\Sigma_1)^{-1}(x-\mu^1)\right\}$$

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

P(C1)
= 79 / (79 + 61) =0.56

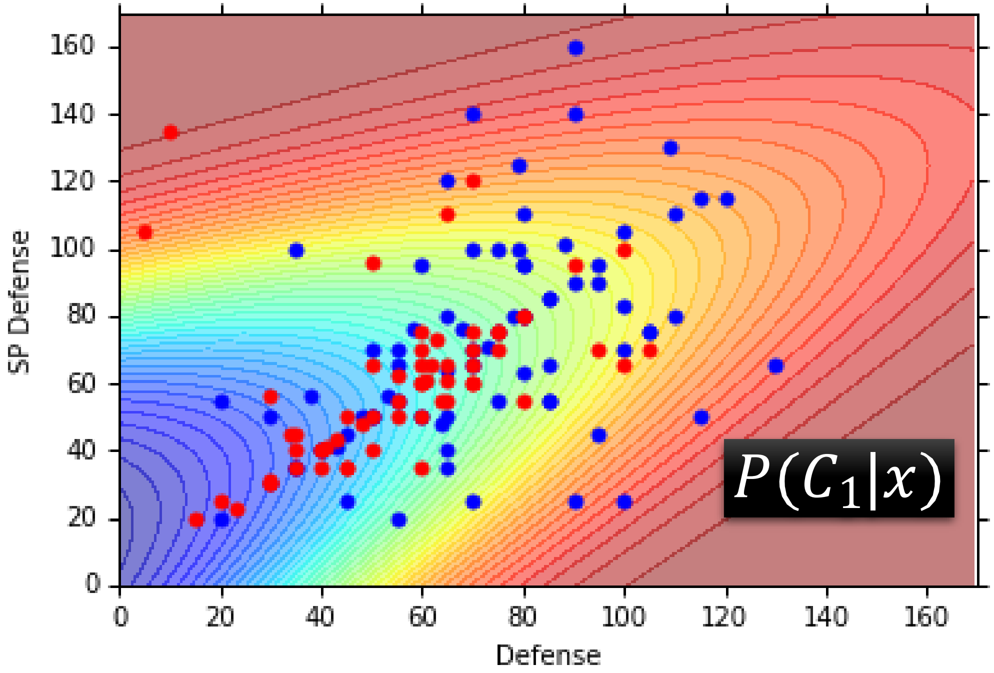
$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

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

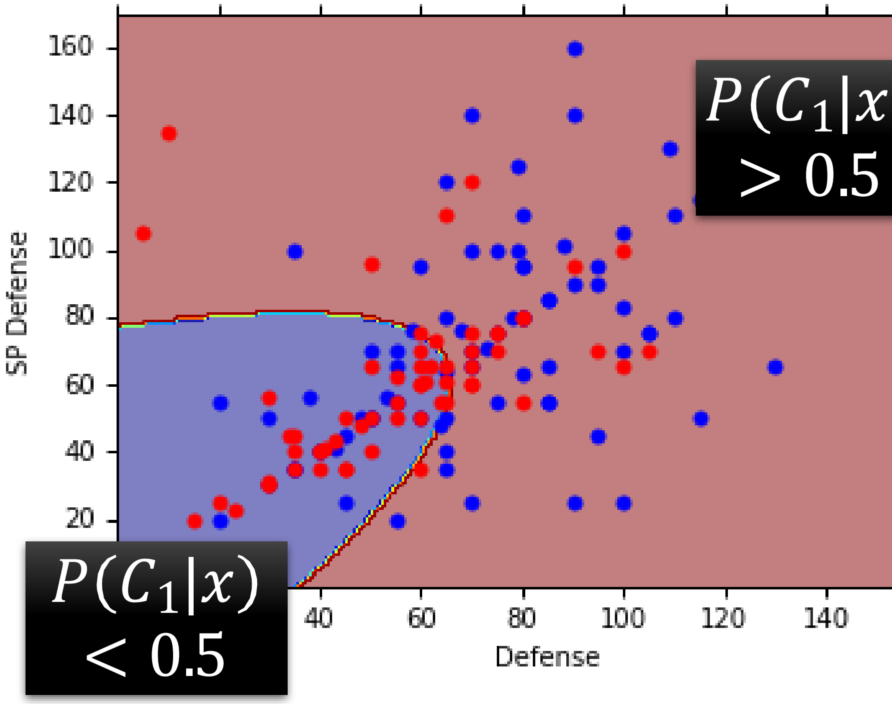
P(C2)
= 61 / (79 + 61)
=0.44

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

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?

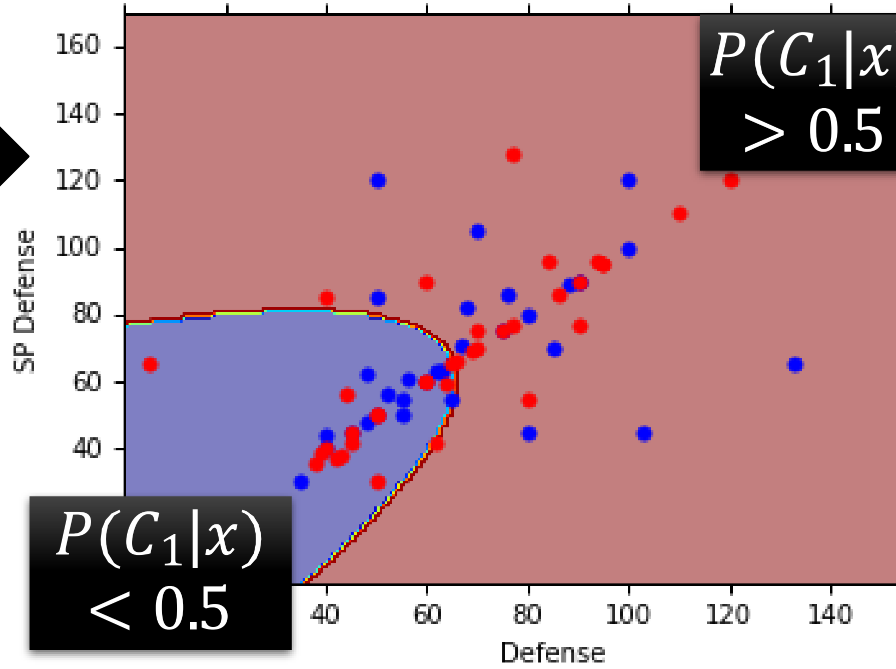
Testing data: 47% accuracy

All: total, hp, att, sp att, de, sp de, speed (7 features)

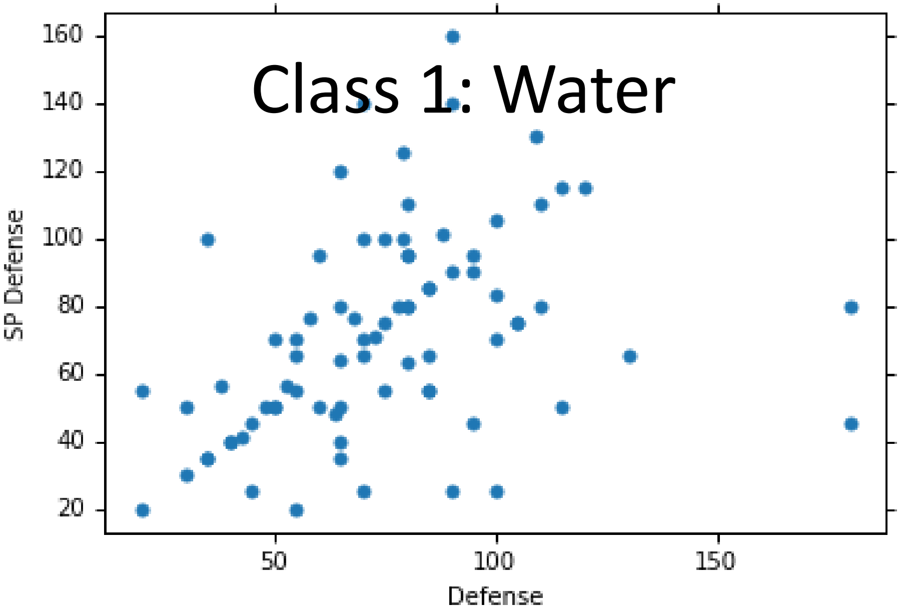
μ^1, μ^2: 7-dim vector

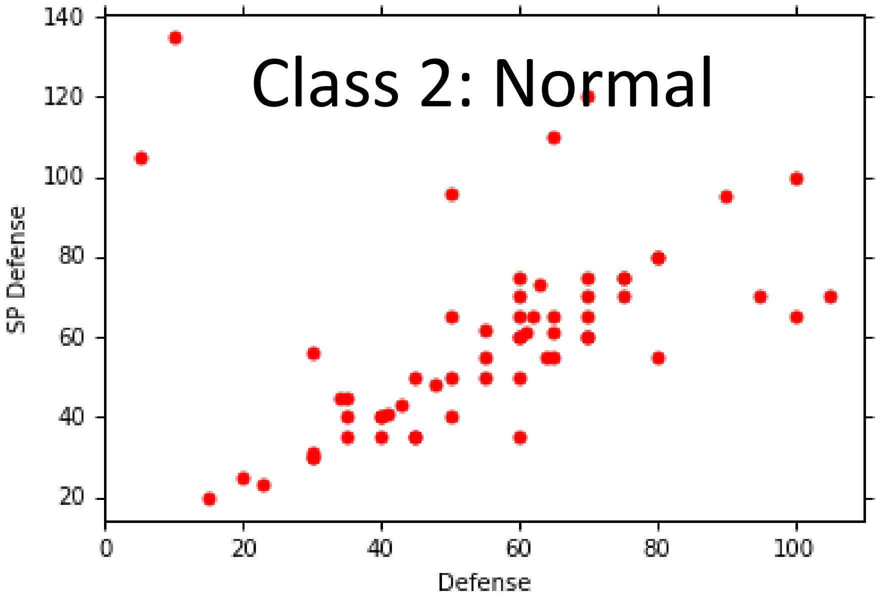
Σ^1, Σ^2: 7 x 7 matrices

54% accuracy … ☹



Modifying Model


Class 1: Water


Class 2: Normal

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

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

The same Σ

Less parameters

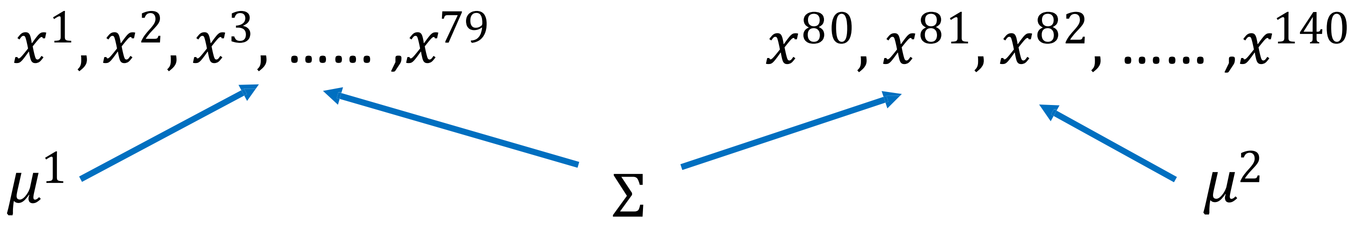
Modifying Model

Ref: Bishop, chapter 4.2.2

- Maximum likelihood

“Water” type Pokémons:

“Normal” type Pokémons:



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

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

μ^1

Σ

μ^2

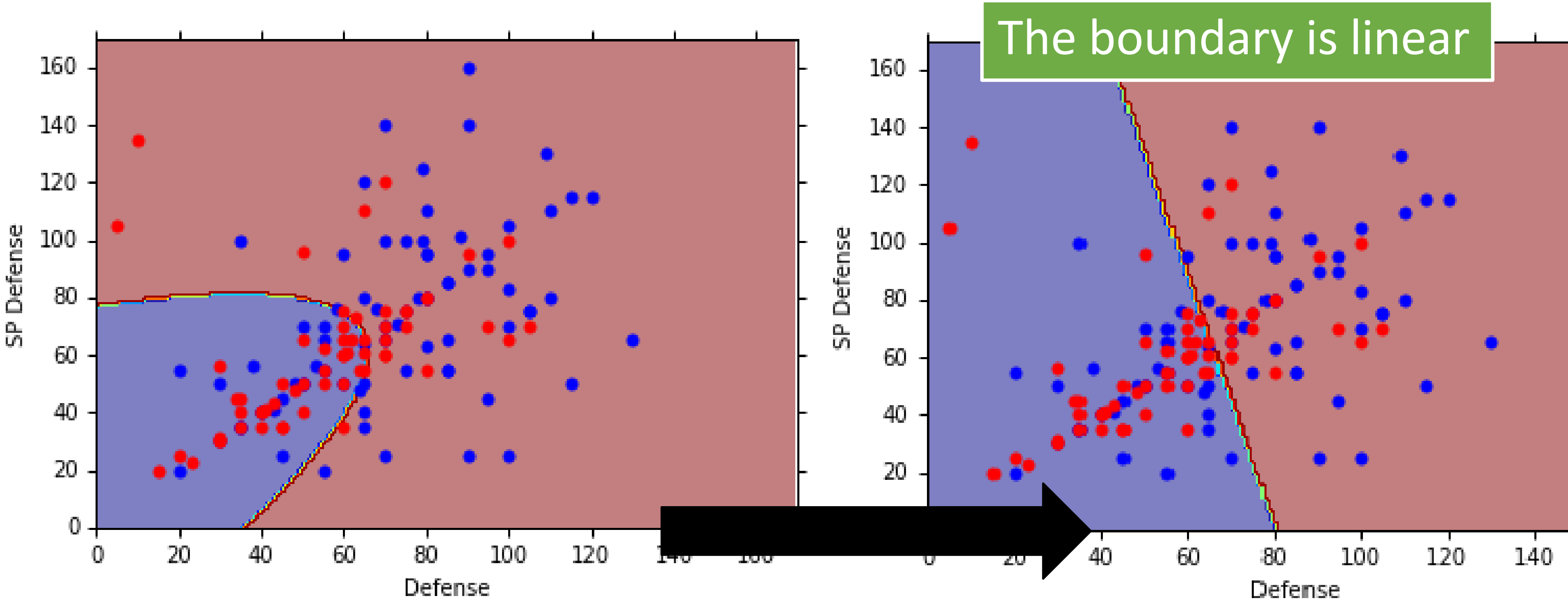
Find μ^1, μ^2, Σ maximizing the likelihood $L(\mu^1, \mu^2, \Sigma)$

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

μ^1 and μ^2 is the same

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

Modifying Model



The boundary is linear

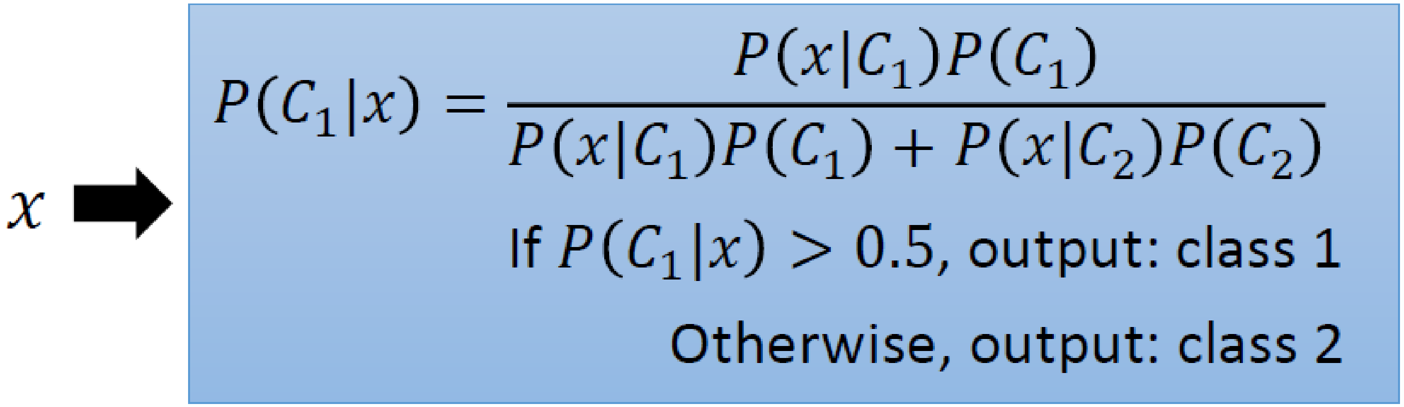
The same covariance matrix

All: total, 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

Naive Bayes Discrete

- We have a dataset with discrete feature values. We want to classify an item with a set of features (F). Essentially what we want to do is to predict the class of an item given the features.
- For a specific class, **Class=**c_1, we will find the conditional probability given the item features:
- $P(Class{=}c_1|F) = \frac{P(F|Class{=}c_1)*P(Class{=}c_1)}{P(F)}$
- The features are a vector with many elements.
- $P(Class{=}c_1|F) = \frac{P(\ F1\ |Class{=}c_1)*P(\ F2\ |Class{=}c_1)*\ldots*P(\ Fn\ |Class{=}c_1)*P(Class{=}c_1)}{P(\ F1\ )*P(\ F2\ )*\ldots*P(\ Fn\ )}$

Same for different classes.
Constant! Ignore!

Naive Bayes Discrete

- The probability of Class in the dataset.
  - $P(Class = c_1)$= # of times Class c_1 occurs in the dataset=$\frac{|D_{c_1}|}{|\text{D}|}$
    - $|D_{c_1}|$: # of examples in Class c_1
    - |D|: Total # of examples in dataset
- The conditional probability of each feature occurring in an item classified in Class.
  - Discrete: how many times a feature value occurs in items of each class.

Naive Bayes Discrete

- The conditional probability

$$P(\,F1 = x1\,|Class = c_1) = \frac{|D_{c,F_1=x_1}|}{|D_{c_1}|}$$

- $D_{c=c_1,F_1=x_1}$: number of examples where $F_1 = x_1$ in class c_1

Reference

- Textbook 20.1-20.2
- Bishop: Chapter4.1-4.2

Acknowledgment

- Thanks to Professor Hyung-Yi Lee form National Taiwan University for these lecture slides.