Introduction to Learning

Introduction to Learning

- An agent is learning if it improves its performance after making observations about the world.

- Learning can range from the trivial, such as jotting down a shopping list, to the profound, as when Albert Einstein inferred a new theory of the universe.

- When the agent is a computer, we call it **machine learning**: a
  - computer observes some data,
  - builds a model based on the data,
  - uses the model as both a hypothesis about the world and a piece of software that can solve problems.

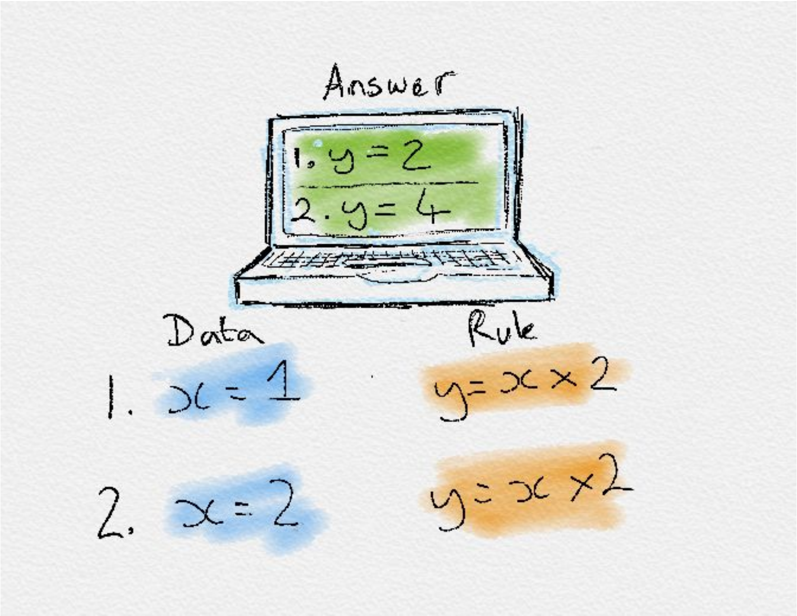
Why Machine Learning?

- Why would we want a machine to learn?
- Why not just program it the right way to begin with?

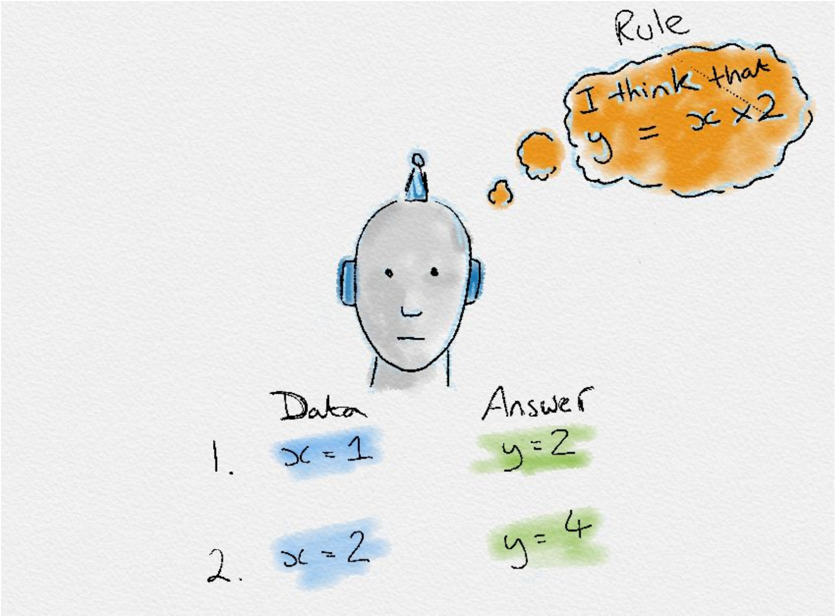
What is ML?

- Machine learning (ML) is the study of computer algorithms that improve automatically through experience.
- Machine-learning algorithms use statistics to find patterns in massive amounts of data.
- Traditionally, software engineering combined human created rules with data to create answers to a problem. Instead, machine learning uses data and answers to discover the rules behind a problem – **F. Chollet, Deep Learning with Python**

What is ML?



Traditional Programming



Machine Learning

Terminologies used in ML

- ML systems learn how to make inference from the input data samples to produce useful predictions on un-seen (test) data.
- Input data:
  - labelled examples: A labelled example includes feature(s) and the label. {features, label}: (x, y)For e.g.:

Features:	Label
Normal RBC, Normal HgB	Healthy
Low RBC, Low HgB	Anaemic

  - unlabelled examples: An unlabelled example contains features but not the label. {features, ?}: (x, ?)
    - For e.g.:

Features:

Housing type: 4BHK, Price: 40,000

Housing type: 4BHK, Price: 15,000

Housing type: 2BHK, Price: 25,000

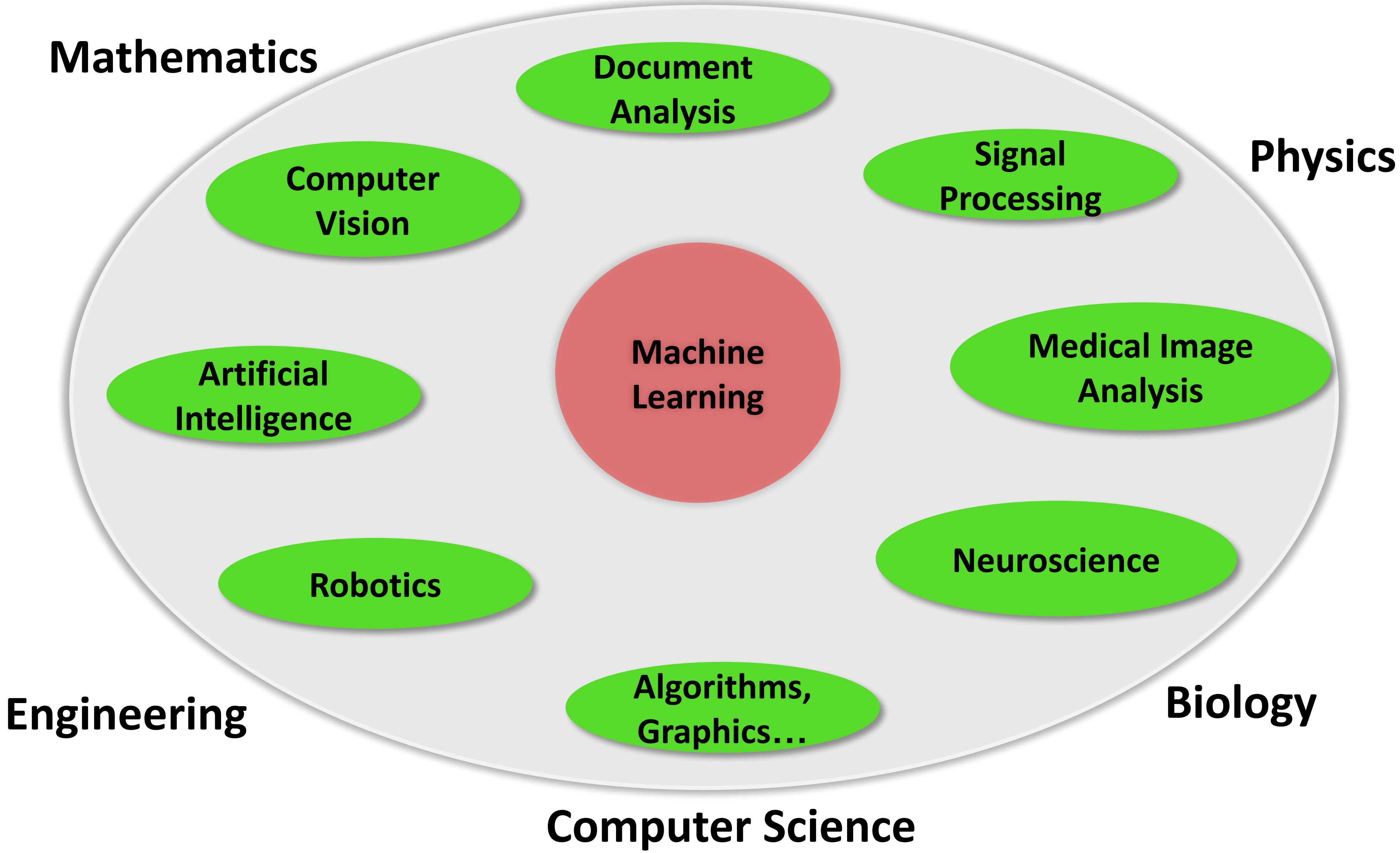
Terminologies used in ML

- Machine Learning Model:
  - A ML model defines the relationship between the features and label.
    - For e.g.: An anaemia diagnostic model might associate certain features strongly with “anaemic“ or “healthy”, and predict the labels based on the association rules it inferred.
  - Two Phases of ML model development
    - **Training** means creating or **learning** the model.
    - **Testing/Inference** means applying the trained model to unlabelled examples.

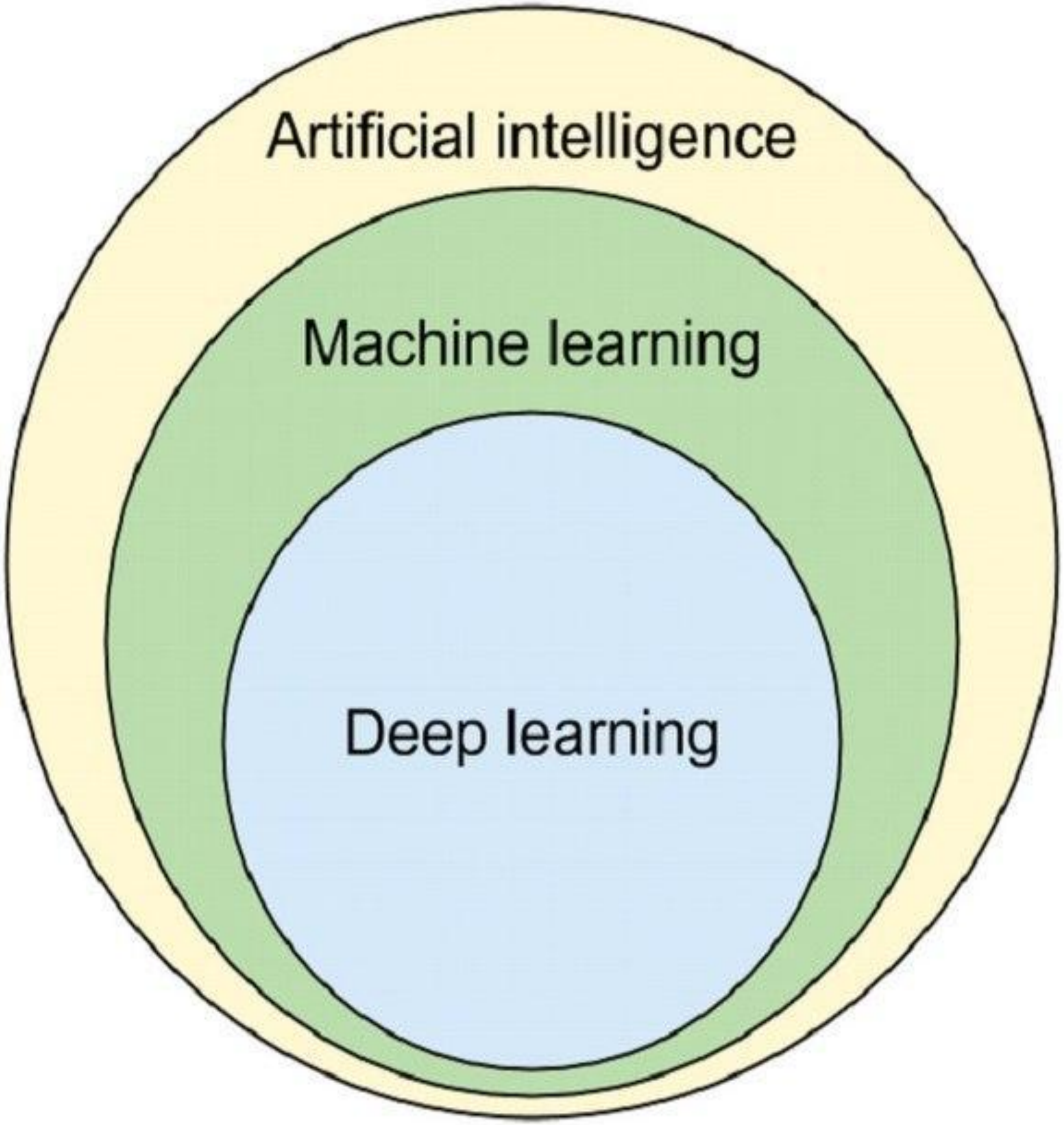
Applications

- Hand-written digit recognition
- Speech recognition
- Face detection
- Object classification
- Email spam detection
- Computational biology
- Autonomous cars
- Computer-aided diagnosis

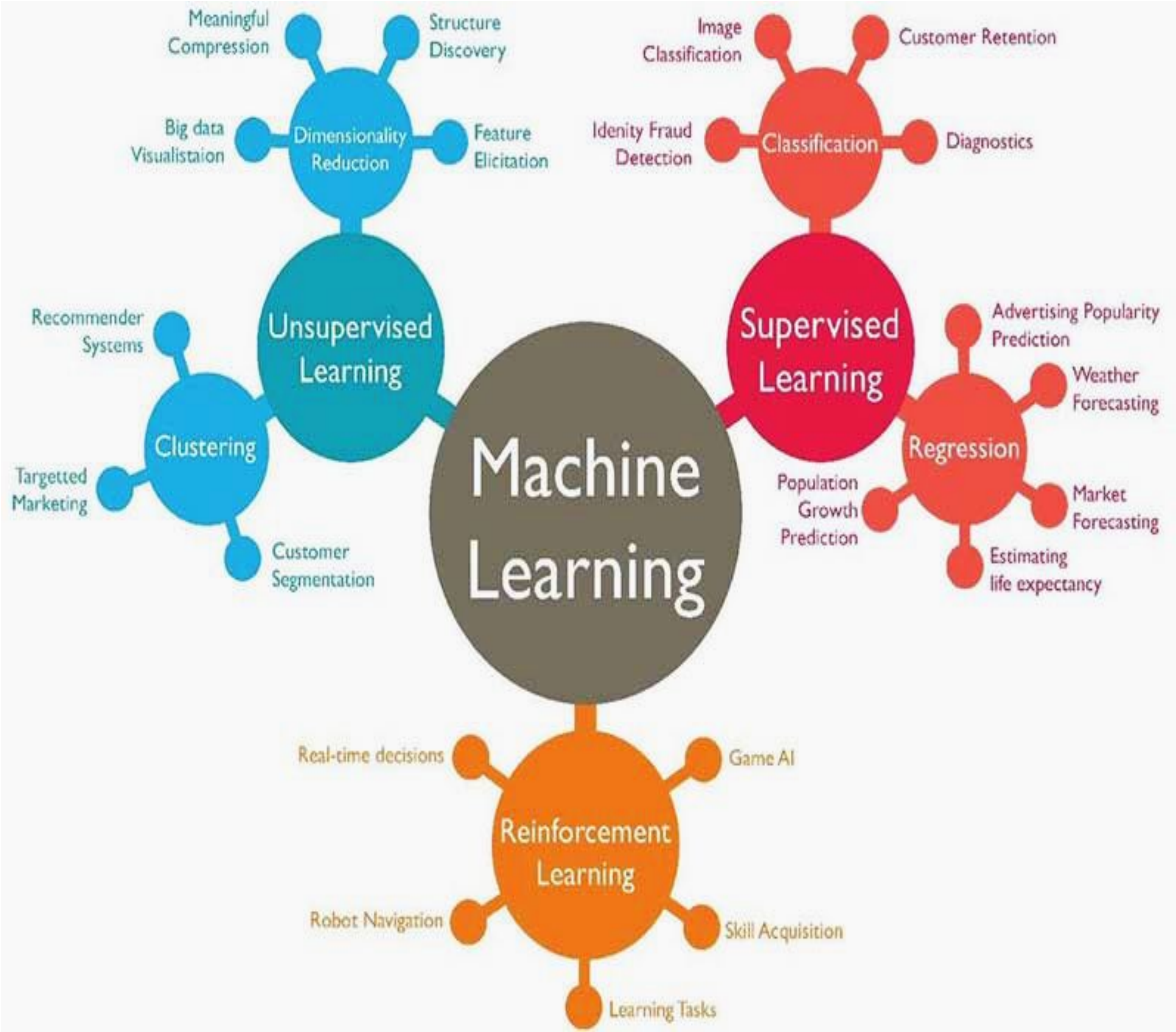
Relation with Other Fields



Relation with AI, ML and DL



Different Machine Learning Paradigms



Bayesian Learning

- ML works with data and hypotheses.
- Here, the data are evidence—that is, instantiations of some or all of the random variables describing the domain.
- The hypotheses are probabilistic theories of how the domain works, including logical theories as a special case.
- **Bayesian learning** simply calculates the probability of each hypothesis, given the data, and makes predictions on that basis.
    - i.e, the predictions are made by using all the hypotheses, weighted by their probabilities, rather than by using just a single “best” hypothesis.

Bayesian Learning

- **Marginal Probability**: The probability of an event irrespective of the outcomes of other random variables, e.g. P(A).

- **Joint Probability**: Probability of two (or more) simultaneous events, e.g. P(A and B) or P(A, B).

- **Conditional Probability**: Probability of one (or more) event given the occurrence of another event, e.g. P(A given B) or P(A | B)

- The joint probability can be calculated using the conditional probability; for example:
  $P(A, B) = P(A | B) * P(B)$

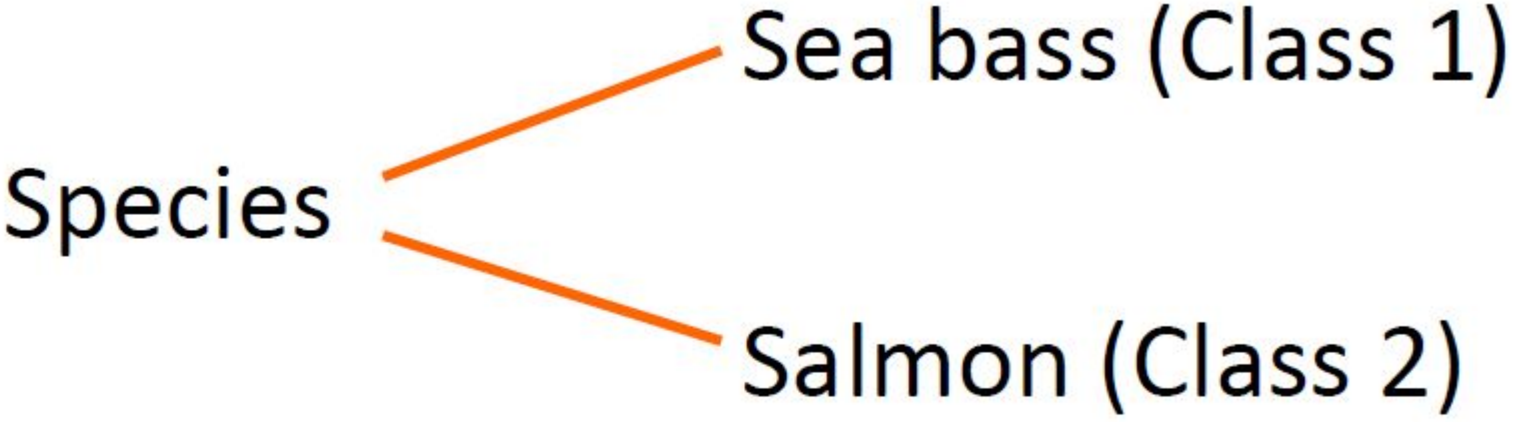
- This is called the product rule. Importantly, the joint probability is symmetrical, meaning that:
  $P(A, B) = P(B, A)$

- The conditional probability can be calculated using the joint probability; for example:
  $P(A | B) = P(A, B) / P(B)$

- The conditional probability is not symmetrical; for example:
  $P(A | B) != P(B | A)$

An Example

- “Sorting incoming Fish on a conveyor according to species using optical sensing”

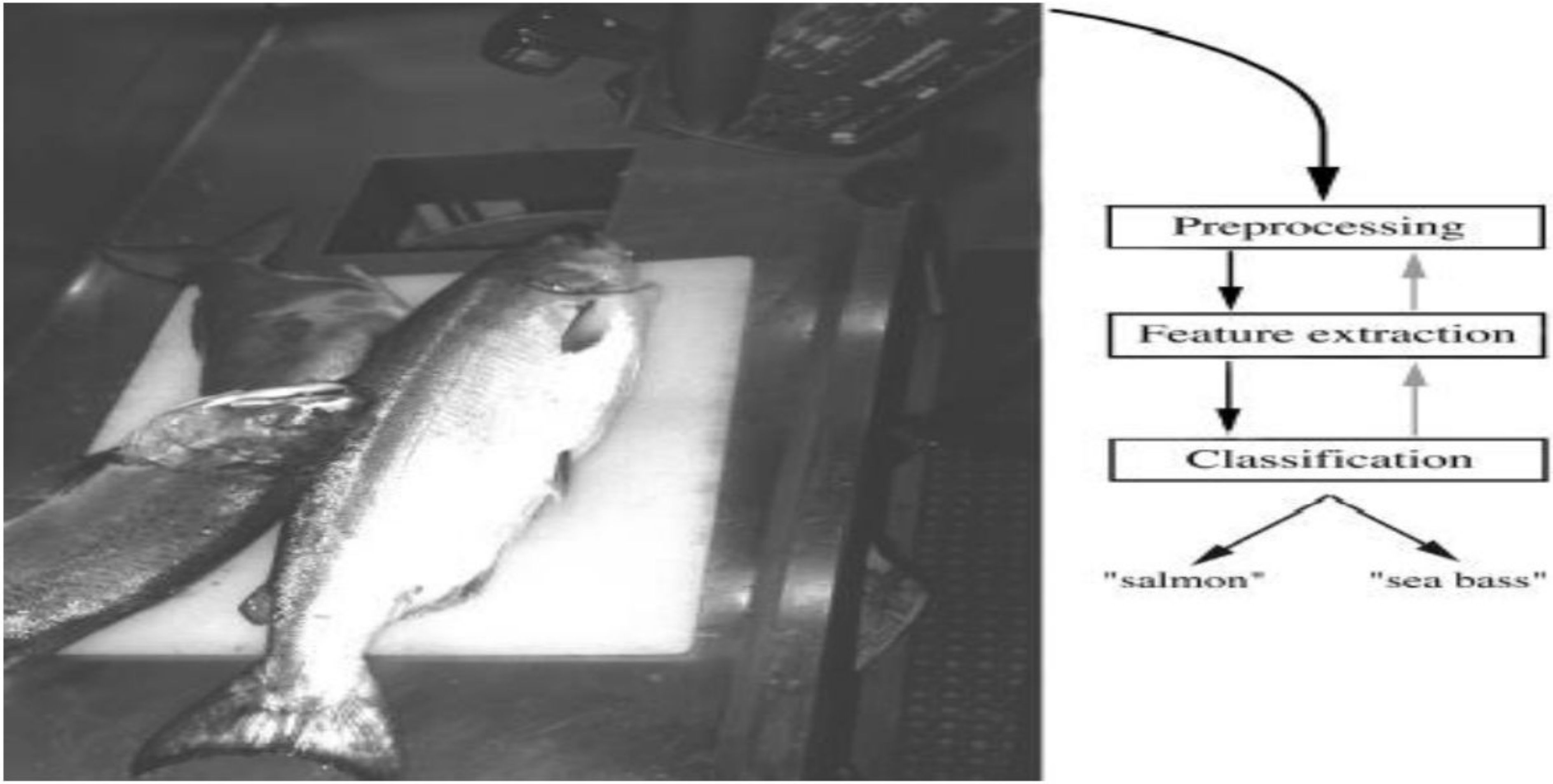


Species

Sea bass (Class 1)

Salmon (Class 2)

Bayesian Decision Theory Primary source of reference: *Pattern Classification* – Duda and Hart



Preprocessing

Feature extraction

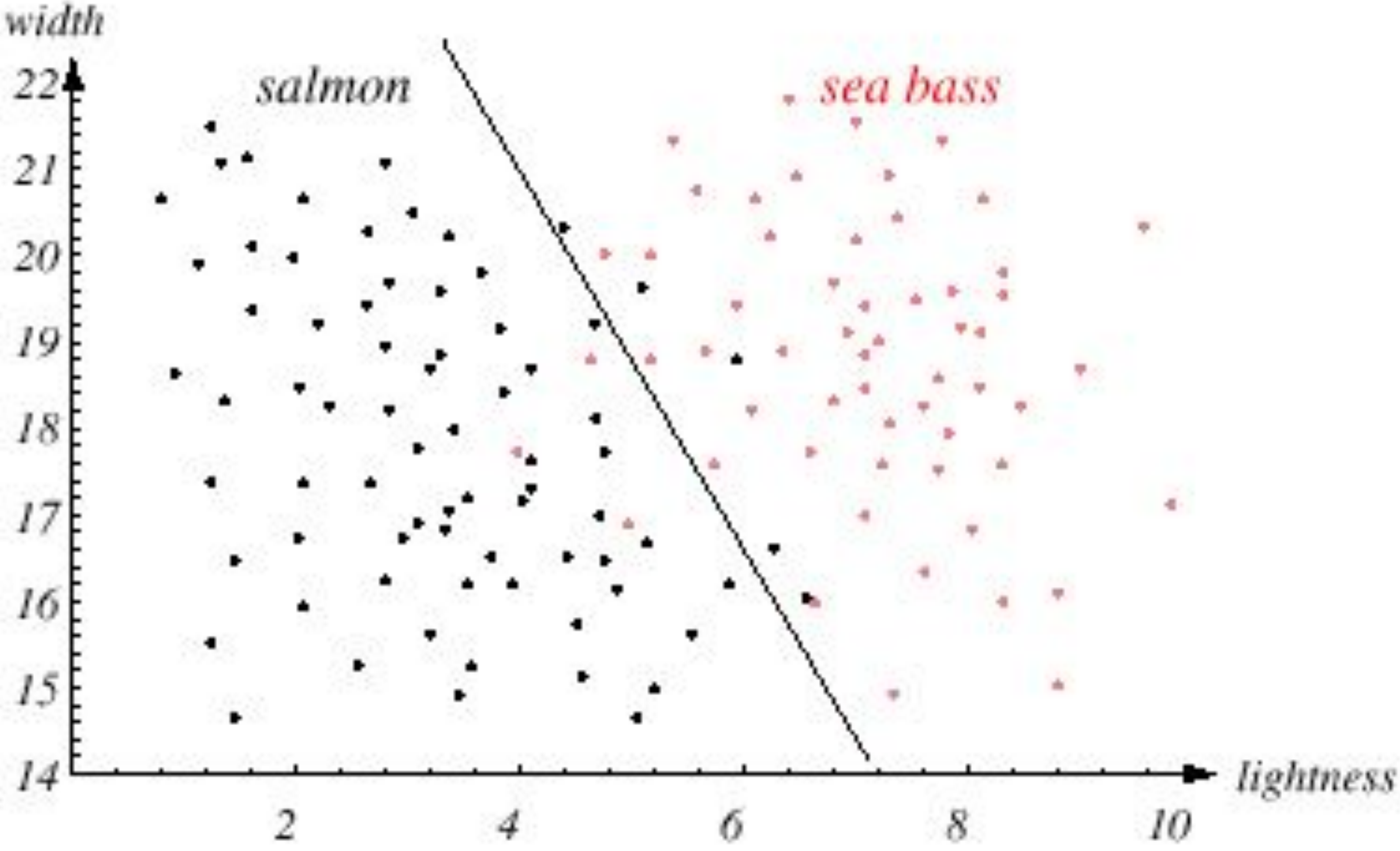
Classification

"salmon"

"sea bass"

Problem Analysis

- Set up a camera and take some sample images to extract features like
    - Length of the fish
    - Lightness (based on the gray level)
    - Width of the fish


width
22
21
20
19
18
17
16
15
14
salmon
sea bass
lightness
2
4
6
8
10

- The sea bass/salmon example (a two class problem)
- For example if we randomly catch 100 fishes and out of this if 75 are *sea bass* and 25 are *salmon*.
-
- Let the rule, in this case is: For any fish say its class is *sea bass*.
- What is the error rate of this rule?
- This information which is independent of feature values is called **apriori** knowledge.

Let the two classes are ω_1 and ω_2

- $P(\omega_1) + P(\omega_2) = 1$
- State of nature (class) is a random variable
- If $P(\omega_1) = P(\omega_2)$, we say it is of uniform priors
  - The catch of salmon and sea bass is equi-probable

- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$, otherwise decide ω_2
- *This is not a good classifier.*
- *We should take feature values into account !*
- *If x is the pattern we want to classify, then use the rule:*

If $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ *then assign class* ω_1
Else assign class ω_2

- $P(\omega_1 \mid x)$ *is called posteriori probability of class* ω_1 *given that the pattern is x.*

Bayes rule

- From data it might be possible for us to estimate $p(x \mid \omega_j)$, *where* $i = 1$ *or* 2. These are called class-conditional distributions.

- Also it is easy to find apriori probabilities $P(\omega_1)$ and $P(\omega_2)$. How this can be done?

- Bayes rule combines apriori probability with class conditional distributions to find posteriori probabilities.

Bayes rule

$$P(B|A) = \frac{P(A, B)}{P(A)} = \frac{P(A|B) * P(B)}{P(A)}$$

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) \,.\, P(\omega_j)}{p(x)}$$

- Where in case of two categories

$$p(x) = \sum_{j=1}^{j=2} p(x \mid \omega_j) P(\omega_j)$$

- Posterior = $\dfrac{\text{Likelihood . Prior}}{\text{Evidence}}$

Decision given the posterior probabilities

X is an observation for which:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ True state of nature = ω_1
if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ True state of nature = ω_2

Therefore:

whenever we observe a particular x, the probability of error is :

$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$
$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

- Minimizing the probability of error

- Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

  Therefore:

$$P(error \mid x) = min\ [P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

(error of Bayes decision)

Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\,white \mid \omega_1) = 0.2$, $P(\,white \mid \omega_2\,) = 0.6$, $P(\,dark \mid \omega_1) = 0.8$, $P(dark \mid \omega_2) = 0.4$ Find $P(\,error\,)$ of the Bayes Classifier.

$$P(white) = P(white|\omega_1)P(\omega_1) + P(white|\omega_2)P(\omega_2)$$

$$P(white) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\,white \mid \omega_1) = 0.2$, $P(\,white \mid \omega_2\,) = 0.6$, $P(\,dark \mid \omega_1) = 0.8$, $P(dark \mid \omega_2) = 0.4$ Find $P(\,error\,)$ of the Bayes Classifier.

$$P(white) = P(white|\omega_1)P(\omega_1) + P(white|\omega_2)P(\omega_2)$$

$$P(white) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

$$P(dark) = P(dark|\omega_1)P(\omega_1) + P(dark|\omega_2)P(\omega_2)$$

$$P(dark) = 0.8 * 0.75 + 0.4 * 0.25 = 0.7$$

Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\, white \mid \omega_1) = 0.2$, $P(\, white \mid \omega_2\,) = 0.6$, $P(\, dark \mid \omega_1) = 0.8$, $P(dark \mid \omega_2) = 0.4$ Find $P(\, error\,)$ of the Bayes Classifier.

$$P(white) = P(white|\omega_1)P(\omega_1) + P(white|\omega_2)P(\omega_2)$$

$$P(white) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

$$P(dark) = P(dark|\omega_1)P(\omega_1) + P(dark|\omega_2)P(\omega_2)$$

$$P(dark) = 0.8 * 0.75 + 0.4 * 0.25 = 0.7$$

$$P(\omega_1|white) = \frac{P(white|\omega_1)P(\omega_1)}{P(white)} = \frac{0.2 * 0.75}{0.3} = 0.5$$

$$P(\omega_2|white) = \frac{P(white|\omega_2)P(\omega_2)}{P(white)} = \frac{0.6 * 0.25}{0.3} = 0.5$$

Consider a one dimensional two class problem. The feature used is color of fish. Color can be either white or dark $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, $P(\, white \mid \omega_1) = 0.2$, $P(\, white \mid \omega_2\,) = 0.6$, $P(\, dark \mid \omega_1) = 0.8$, $P(dark \mid \omega_2\,) = 0.4$ Find $P(\,error\,)$ of the Bayes Classifier.

$$P(white) = P(white|\omega_1)P(\omega_1) + P(white|\omega_2)P(\omega_2)$$

$$P(white) = 0.2 * 0.75 + 0.6 * 0.25 = 0.3$$

$$P(dark) = P(dark|\omega_1)P(\omega_1) + P(dark|\omega_2)P(\omega_2)$$

$$P(dark) = 0.8 * 0.75 + 0.4 * 0.25 = 0.7$$

$$P(\omega_1|white) = \frac{P(white|\omega_1)P(\omega_1)}{P(white)} = \frac{0.2 * 0.75}{0.3} = 0.5$$

$$P(\omega_2|white) = \frac{P(white|\omega_2)P(\omega_2)}{P(white)} = \frac{0.6 * 0.25}{0.3} = 0.5$$

$$P(\omega_1|dark) = \frac{P(dark|\omega_1)P(\omega_1)}{P(dark)} = \frac{0.8 * 0.75}{0.7} = \frac{6}{7}$$

$$P(\omega_2|dark) = \frac{P(dark|\omega_2)P(\omega_2)}{P(dark)} = \frac{0.4 * 0.25}{0.7} = \frac{1}{7}$$

$$P(error) = P(error|white)P(white) + P(error|dark)P(dark)$$

$$P(error) = 0.5 * 0.3 + \frac{1}{7} * 0.7 = 0.25$$

- But, what is the error, if we use only apriori probabilities?

Since, $P(\omega_1) = 0.75$, $P(\omega_2) = 0.25$, every pattern is assigned to ω_1, So the error,

$$P(error) = P(\omega_2|white)P(white) + P(\omega_2|dark)P(dark)$$

Since, $P(\omega_1) = 0.75,\ P(\omega_2) = 0.25$, every pattern is assigned to ω_1, So the error,

$$P(error) = P(\omega_2|white)P(white) + P(\omega_2|dark)P(dark)$$

$$P(error) = \frac{P(white|\omega_2)P(\omega_2)}{P(white)}P(white) + \frac{P(dark|\omega_2)P(\omega_2)}{P(dark)}P(dark)$$

$$P(error) = \big(P(white|\omega_2) + P(dark|\omega_2)\big)P(\omega_2)$$

$$P(error) = P(\omega_2) = 0.25$$

- Same error? Where is the advantage?!

Consider $P(\omega_1) = 0.5,\ P(\omega_2) = 0.5$

$$P(white) = P(white|\omega_1)P(\omega_1) + P(white|\omega_2)P(\omega_2)$$

$$P(white) = 0.2 * 0.5 + 0.6 * 0.5 = 0.4$$

$$P(dark) = P(dark|\omega_1)P(\omega_1) + P(dark|\omega_2)P(\omega_2)$$

$$P(dark) = 0.8 * 0.5 + 0.4 * 0.5 = 0.6$$

$$P(\omega_1|white) = \frac{P(white|\omega_1)P(\omega_1)}{P(white)} = \frac{0.2 * 0.5}{0.4} = 0.25$$

$$P(\omega_2|white) = \frac{P(white|\omega_2)P(\omega_2)}{P(white)} = \frac{0.6 * 0.5}{0.4} = 0.75$$

$$P(\omega_1|dark) = \frac{P(dark|\omega_1)P(\omega_1)}{P(dark)} = \frac{0.8 * 0.5}{0.6} = \frac{2}{3}$$

$$P(\omega_2|dark) = \frac{P(dark|\omega_2)P(\omega_2)}{P(dark)} = \frac{0.4 * 0.5}{0.6} = \frac{1}{3}$$

$$P(error) = P(error|white)P(white) + P(error|dark)P(dark)$$

$$P(error) = 0.25 * 0.4 + \frac{1}{3} * 0.6 = 0.3$$

- But, P(error) based on apriori probabilities only is 0.5.
- Error based on the Bayes classifier is the lower bound.
  - Any classifier’s error is greater than or equal to this.

- Read Duda and Hart book.