Pattern Recognition and Machine Learning

Dr Suresh Sundaram

sureshsundaram@iitg.ernet.in

Lets get started

- Person identification systems -> Biometrics, Aadhar,

Human Perception

- How did we learn the alphabet of the English language?

Trained ourselves to recognize alphabets, so that given a new alphabet, we use our memory / intelligence in recognizing it.

Machine Perception

- How about providing such capabilities to machines to recognize alphabets ?

- The field of pattern recognition exactly does that.

Idea

- Build a machine that can recognize patterns:
  - Speech recognition
  - Fingerprint identification
  - OCR (Optical Character Recognition)
  - DNA sequence identification

A basic PR framework

- Training samples
- Testing samples
- An algorithm for recognizing an unknown test sample

- Samples are labeled (supervised learning)

Typical supervised PR problem

- Alphabets – 26 in number (upper case)

- # of alphabets/ classes to recognize – 26.
- Collect samples of each of the 26 alphabets and train using an algorithm.
- Once trained, test system using unknown test sample/ alphabeth.

Basics

pattern $\longrightarrow \boxed{\text{feature extractor}} \xrightarrow{X} \boxed{\text{classifier}} \longrightarrow$ class label

- Feature extractor makes some measurements on the input pattern.
- X is called *Feature Vector*. Often, $X \in \Re^n$.
- Classifier maps each feature vector to a class label.
- Features to be used are problem-specific.

So what's a pattern ?

A pattern is an entity, vaguely defined, that could be given a name, e.g.,

- fingerprint image,
- handwritten word,
- human face,
- speech signal,
- DNA sequence
- alphabeth

Handwriting Recognition

From Nov 10, 1999

Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To

Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

Machine print document

From Nov 10, 1999

Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To

Dr. Rob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm

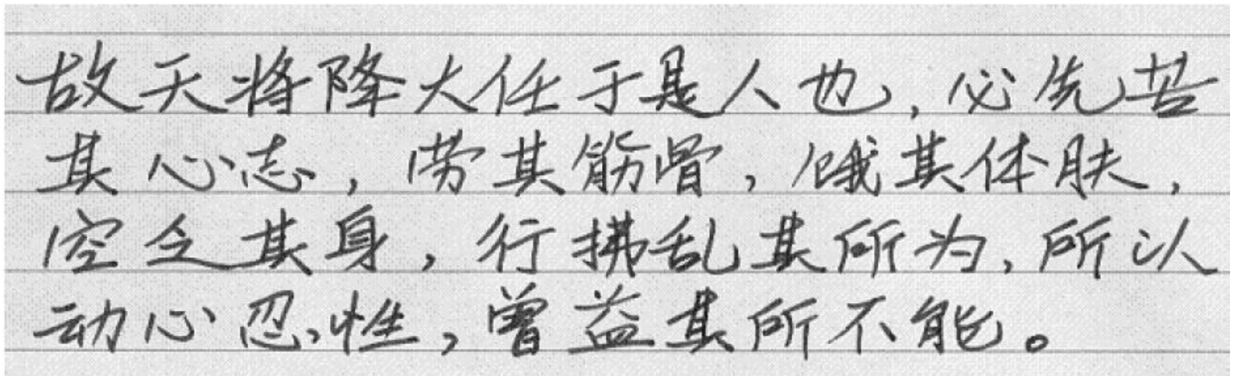
However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

Input handwritten document

Handwriting recognition

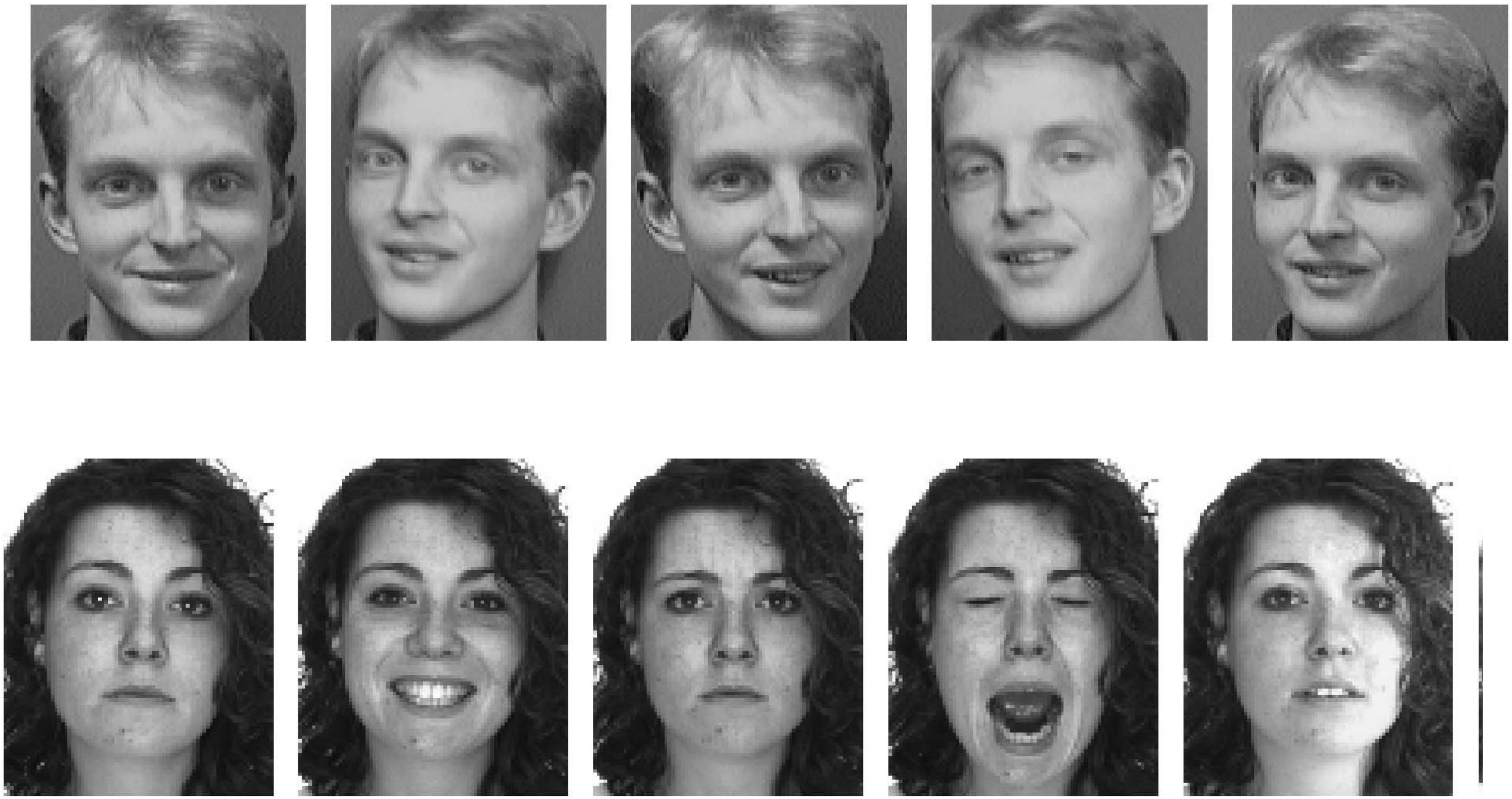

故天将降大任于是人也，必先苦
其心志，劳其筋骨，饿其体肤，
空乏其身，行拂乱其所为，所以
动心忍性，曾益其所不能。

(a) Handwriting

故天将降大任于是人也，必先苦
其心志，劳其筋骨，饿其体肤，
空乏其身，行拂乱其所为，所以
动心忍性，曾益其所不能。

(b) Corresponding Machine Print

Face recognition



Fingerprint recognition

Other Applications

- Object classification
- Signature verification (genuine vs forgery)
- Iris recognition
- Writer adaptation
- Speaker recognition
- Bioinformatics (gene classification)
- Communication System Design
- Medical Image processing

Pattern Recognition Algorithms

- Bag of algorithms that can used to provide some intelligence to a machine.

- These algorithms have a solid probabilistic framework.

- Algorithms work on certain characteristics defining a class -refered as ‘features’.

What is a feature?

- Features across classes need to be discriminative for better classification peformance.

Pattern l

Pattern i

- Presence of a dot in ‘i’ can distinguish these ‘i’ from ‘l’ and is a feature.
- Features values can be discrete or continuous in nature (floating value).
- In practice, a single feature may not suffice for discrimination.

Pattern Recognition Algorithms

- Bag of algorithms that can used to provide some intelligence to a machine.
- These algorithms have a solid probabilistic framework.
- Algorithms work on certain characteristics defining a class -refered as ‘features’.

What is a feature?

- Features across classes need to be discriminative for better classification peformance.

Pattern l

Pattern i

- Presence of a dot in ‘i’ can distinguish these ‘i’ from ‘l’ and is a feature.
- Features values can be discrete or continuous in nature (floating value).
- In practice, a single feature may not suffice for discrimination.

Feature selection

In practice, a single feature may not suffice for discrimination.

- A possible solution is to look out for many features and select a set (possibly with feature selection algorithms). The goal is to improve the recognition performance of unseen test data.

- The different features selected can be represented with a vector called as 'feature vector'.

Dimension of a feature vector

- Suppose we select d features, we can represent them with a d-dimensional feature vector.

- Pixels of an image of size M XN can be represented with a MN*1 dimensional feature vector.

Feature selection

- Domain Knowledge helps in extracting features

- Feature discriminability measures are available like Fisher scores to measure the effectiveness of features.

List of features used in literature

- Pixels in an image
- Edge based features in an image
- Transformed coefficients

DFT (Shape description)
DCT (Compression)
Wavelets (Palm print recognition)
KLT /PCA (Face recognition)
Gabor (Texture classification, script identification)
MFCCs (Speech systems)

Features

- Feature to be discriminative
- Specific to applications…… no universal feature for all pattern recognition problems …. Ugly Duckling Theorem

- To be robust to translation, rotation, occlusion, scaling

Features

- Continuous, real valued
- Discrete
- Binary
- Mixed

Features

- Features depend on the problem. Measure ‘relevant’ quantities.
- Some techniques available to extract ‘more relevant’ quantities from the initial measurements. (e.g., PCA)
- After feature extraction each pattern is a vector
- Classifier is a function to map such vectors into class labels.
- Many general techniques of classifier design are available.
- Need to test and validate the final system.

Curse of dimensionality

- If limited data is available, too many features may degrade the performance ….. We need as large number of training samples for better generalization….to beat the `curse of dimensionality'!

- Need arises to come up with techniques such as PCA to pick the `relevant features'.

Basic Pattern Recognition

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

Species

Sea bass

Salmon

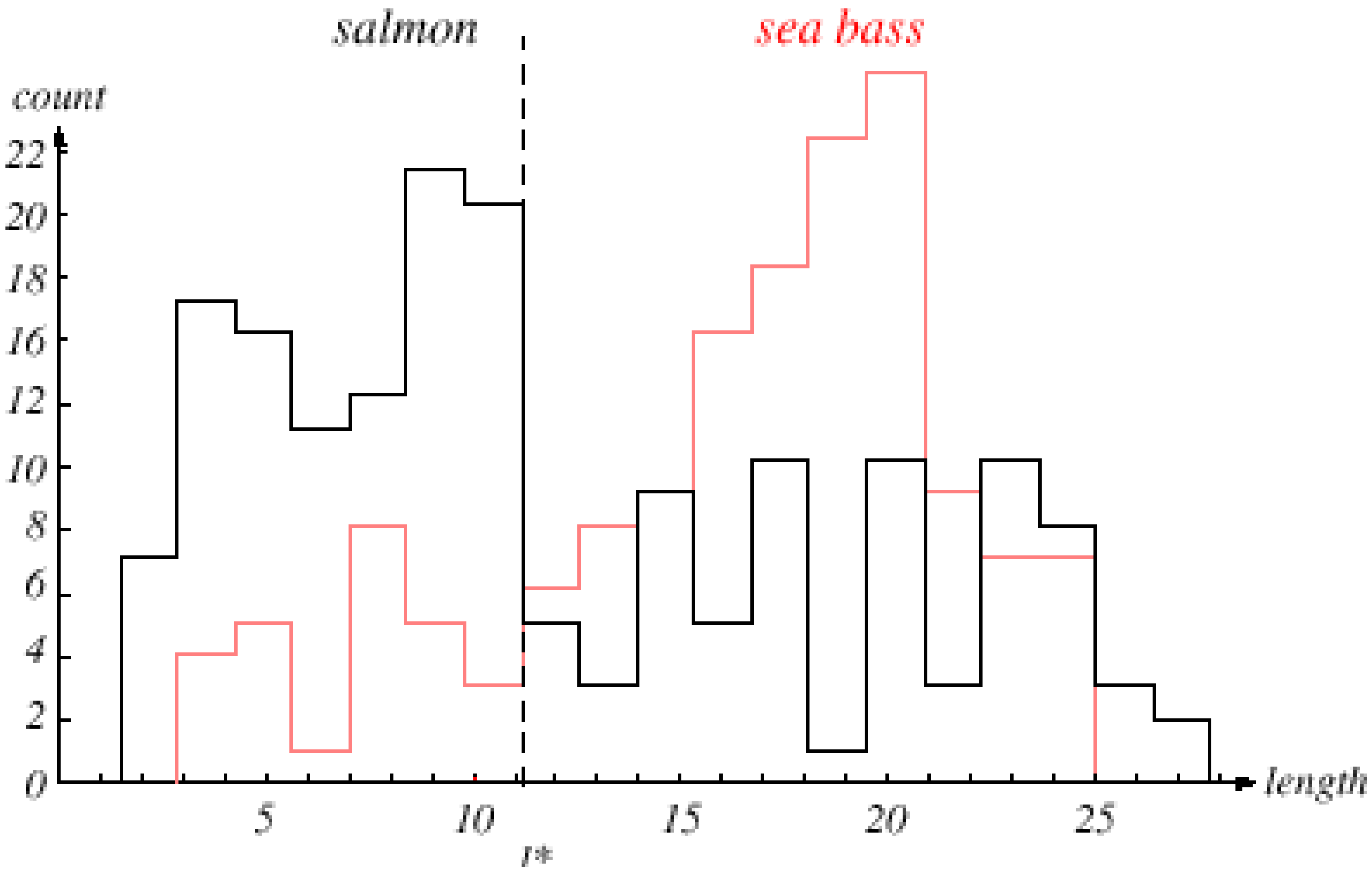

Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

- Problem Analysis
  - Set up a camera and take some sample images to extract features
    - Length
    - Lightness
    - Width
    - Number and shape of fins
    - Position of the mouth, etc…

    - This is the set of all suggested features to explore for use in our classifier!

- Preprocessing
  - Use a segmentation operation to isolate fishes from one another and from the background
- Information from a single fish is sent to a feature extractor whose purpose is to reduce the data by measuring certain features
- The features are passed to a classifier

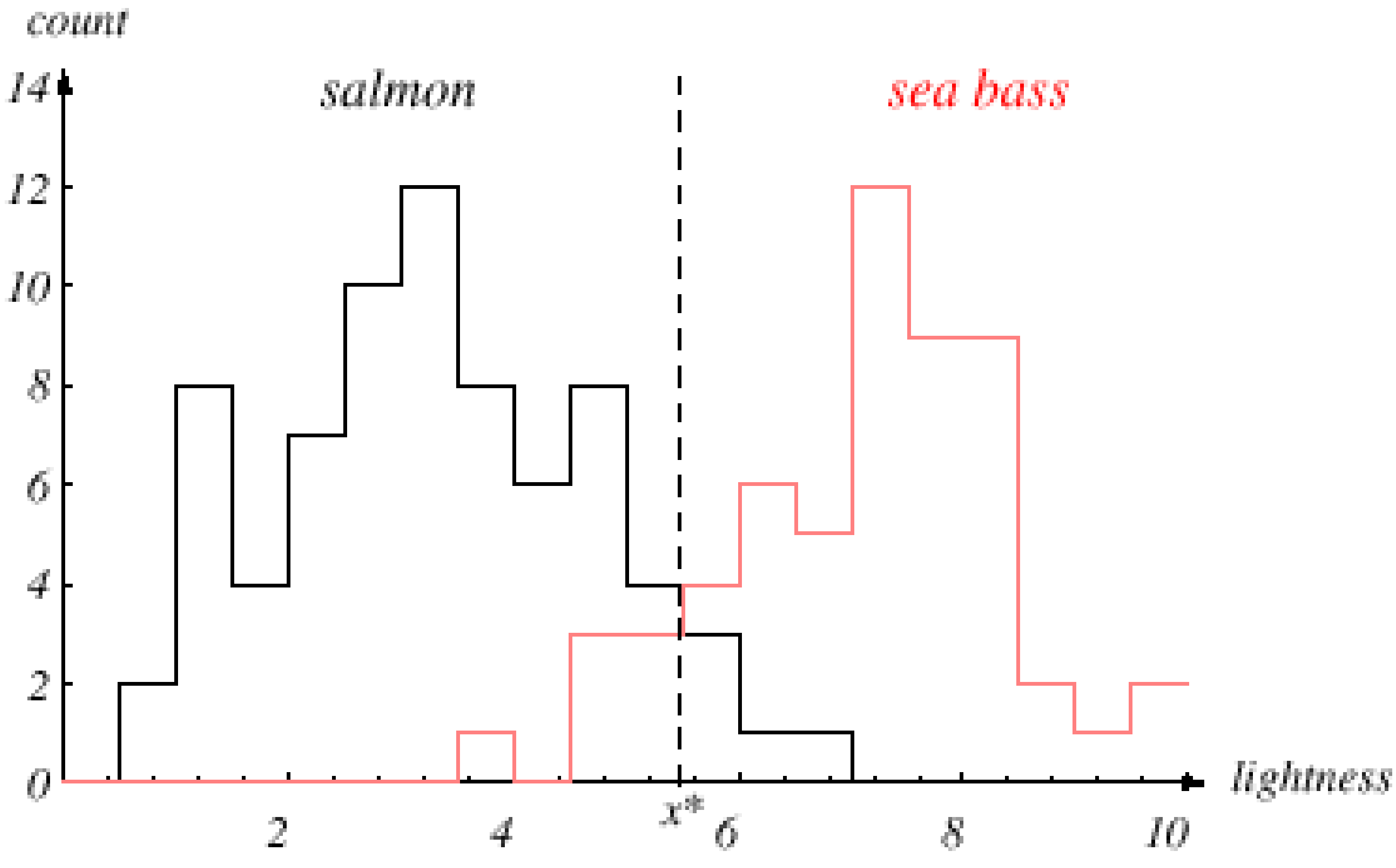
- Classification
  - Select the length of the fish as a possible feature for discrimination



salmon
sea bass
count
22
20
18
16
12
10
8
6
4
2
0
5
10
15
20
25
length
l*

The length is a poor feature alone!

Select the lightness as a possible feature.


count
salmon
sea bass
14
12
10
8
6
4
2
0
2
4
x*
6
8
10
lightness

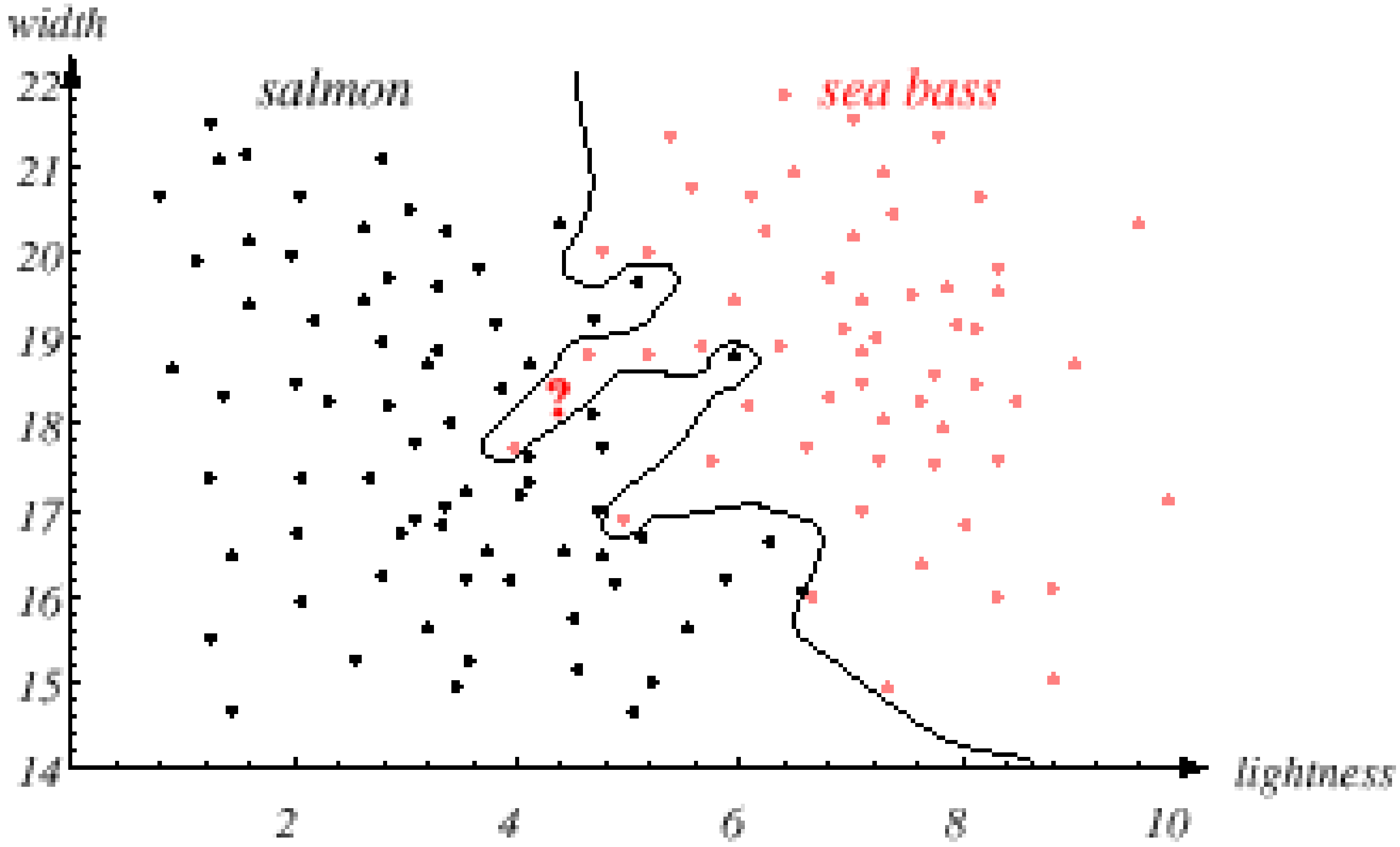
- Adopt the lightness and add the width of the fish as a new feature

$$\text{Fish} \Longrightarrow \boldsymbol{x}^T = [x_1, x_2]$$

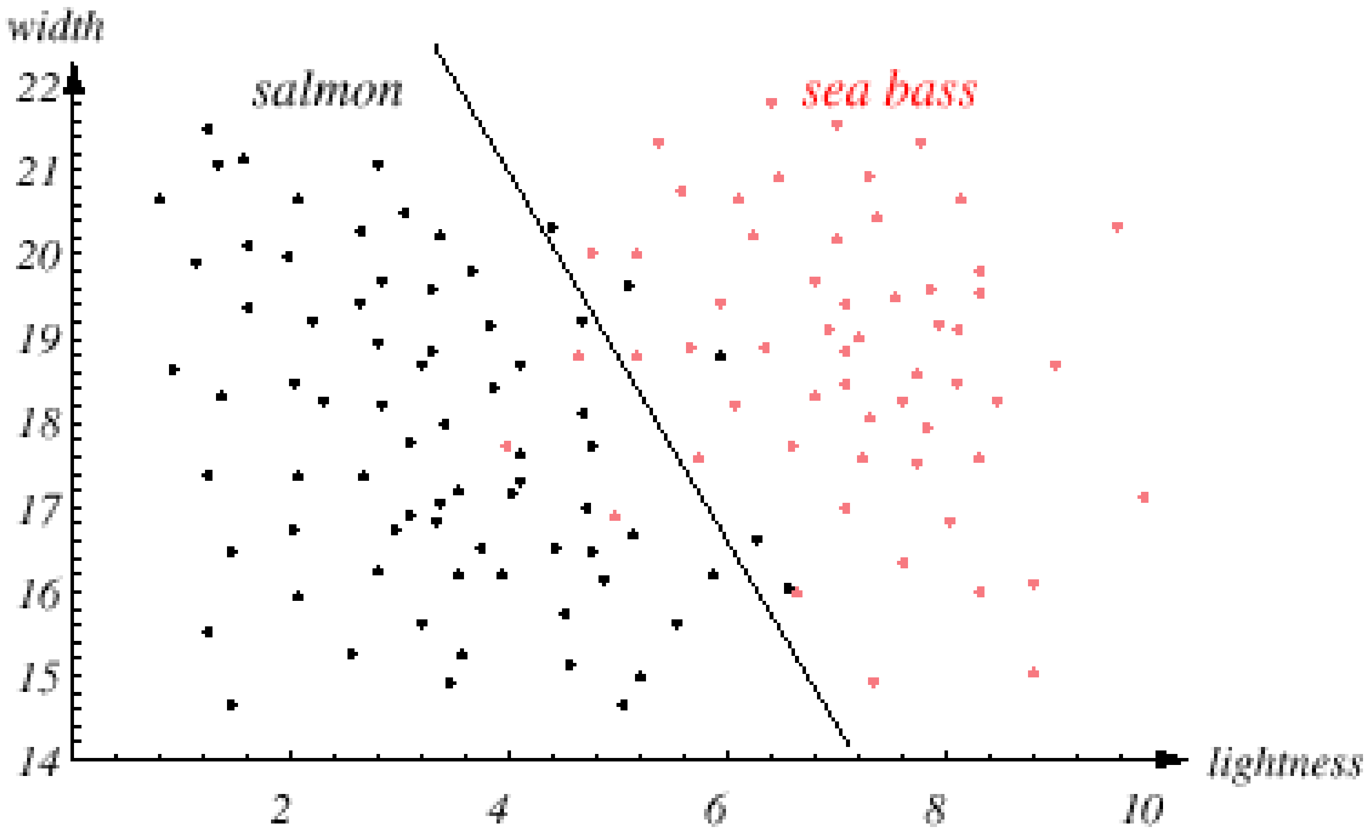
Lightness **Width**

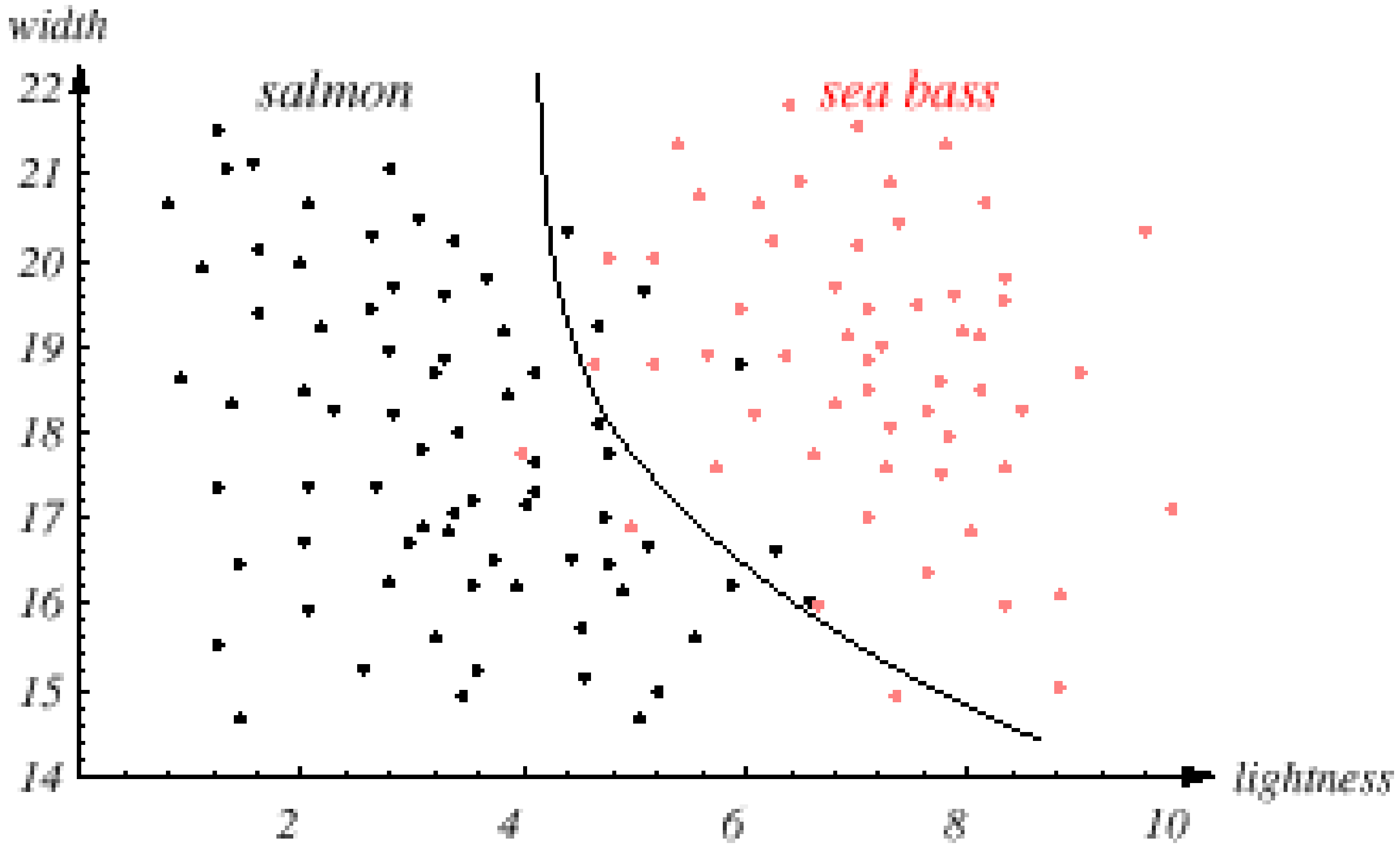
- We might add other features that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding such “noisy features”

- Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:

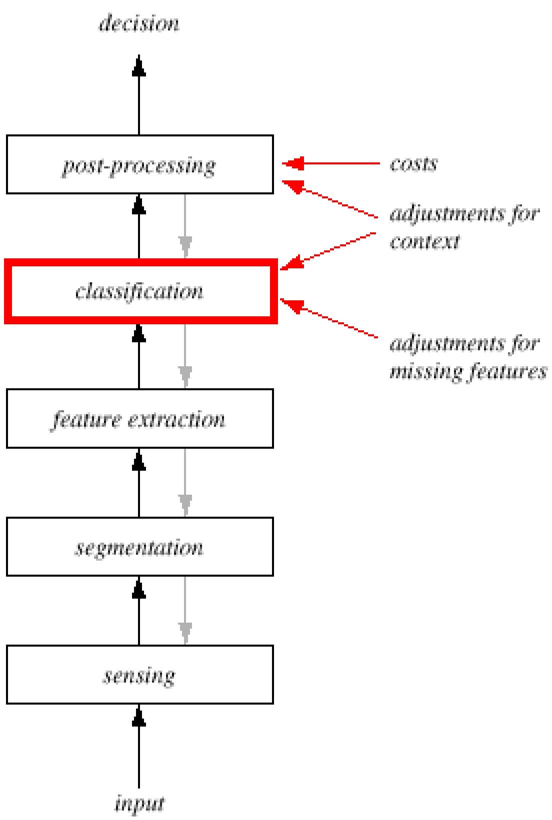


Use simple models to complicated ones : Occams razor


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


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

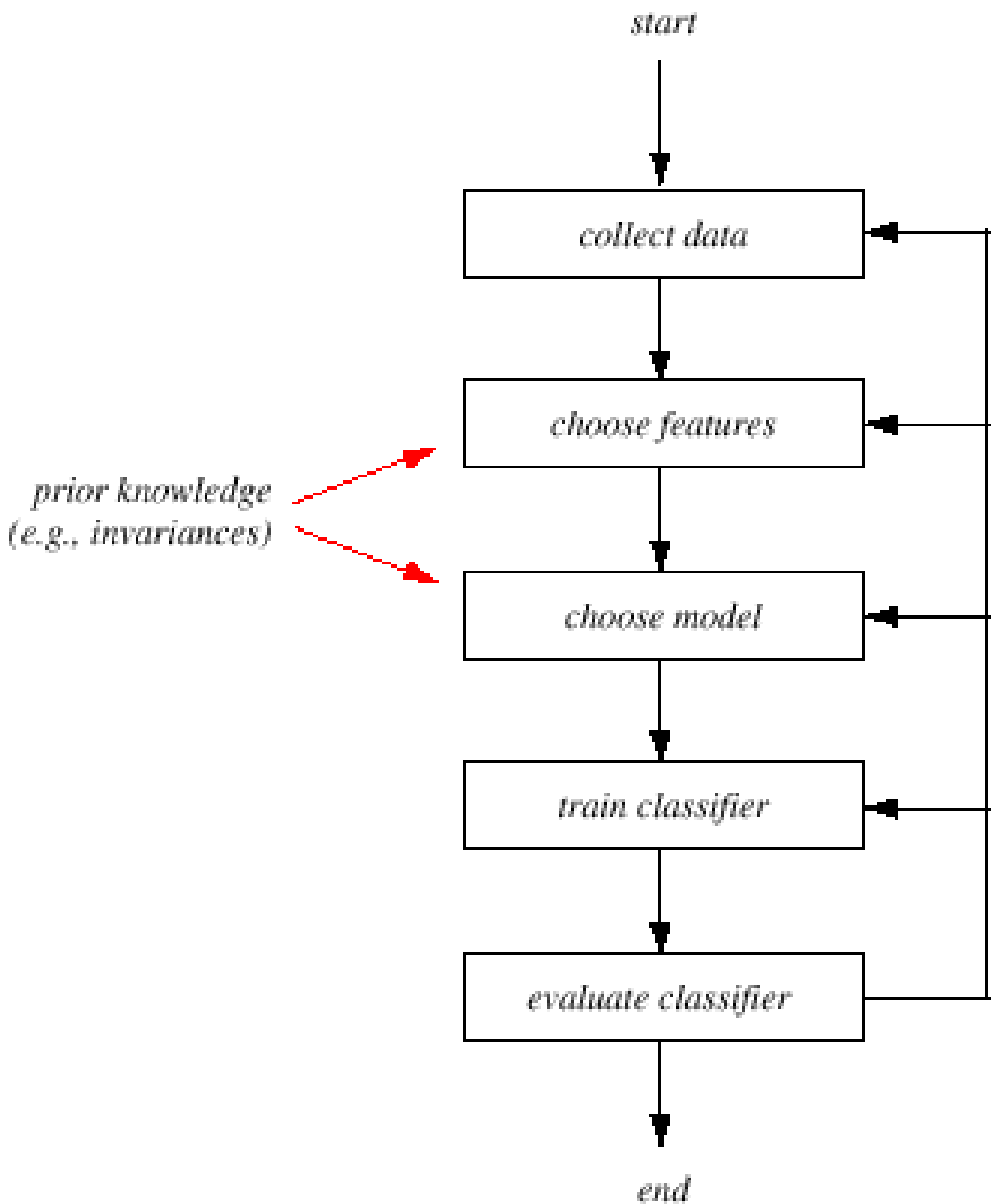
- Sensing
  - Use of a transducer (camera or microphone)
- Segmentation and grouping
  - Patterns should be well separated and should not overlap


decision
post-processing
costs
adjustments for
context
classification
adjustments for
missing features
feature extraction
segmentation
sensing
input

- Feature extraction
  - Discriminative features
  - Invariant features with respect to translation, rotation and scale.
- Classification
  - Use a feature vector provided by a feature extractor to assign the object to a category
- Post Processing
  - Exploit context input dependent information other than from the target pattern itself to improve performance

The Design Cycle

- Data collection
- Feature Choice
- Model Choice
- Training
- Evaluation
- Computational Complexity


start
collect data
choose features
choose model
train classifier
evaluate classifier
end
prior knowledge
(e.g., invariances)

- Data Collection
  - How do we know when we have collected an adequately large and representative set of examples for training and testing the system?

- Feature Choice

  - Depends on the characteristics of the problem domain. Simple to extract, invariant to irrelevant transformation insensitive to noise.

- Model Choice
  - Unsatisfied with the performance of our fish classifier and want to jump to another class of model

- Training
  - Use data to determine the classifier. Many different procedures for training classifiers and choosing models

- Evaluation

  - Measure the error rate (or performance and switch from one set of features to another one

- Computational Complexity
  - What is the trade-off between computational ease and performance?
  - (How an algorithm scales as a function of the number of features, patterns or categories?)

Learning paradigms

- Supervised learning
  - A teacher provides a category label or cost for each pattern in the training set
- Unsupervised learning
  - The system forms clusters or “natural groupings” of the input patterns

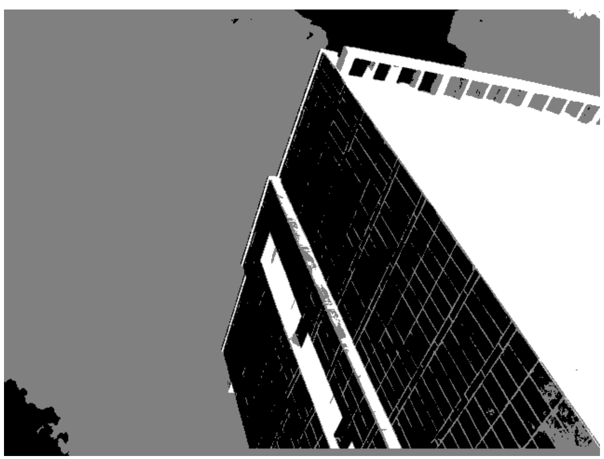
Unsupervised Learning

- The system forms clusters or “natural groupings” of the input patterns….

- Clustering is often called an **unsupervised learning** task as no class values denoting an a priori grouping of the data instances are given

Segmentation of an image into *k* clusters by a popular iterative algorithm called *k* Means Algorithm.

Original image



Segmented image using
k Means Clustering
(k=3)

Reinforcement learning

- **Reinforcement learning** is an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take *actions* in an *environment* so as to maximize some notion of cumulative *reward*.

Semi-supervised learning

- **Semi-supervised learning** is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

- It falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data).

Learning paradigms

- Supervised learning
  - A teacher provides a category label or cost for each pattern in the training set
- Unsupervised learning
  - The system forms clusters or “natural groupings” of the input patterns

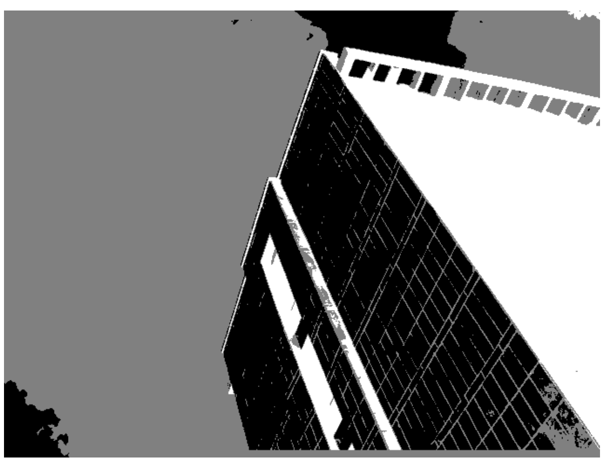
Unsupervised Learning

- The system forms clusters or “natural groupings” of the input patterns….

- Clustering is often called an **unsupervised learning** task as no class values denoting an a priori grouping of the data instances are given

Segmentation of an image into *k* clusters by a popular iterative algorithm called *k* Means Algorithm.

Original image



Segmented image using
k Means Clustering
(k=3)

Reinforcement learning

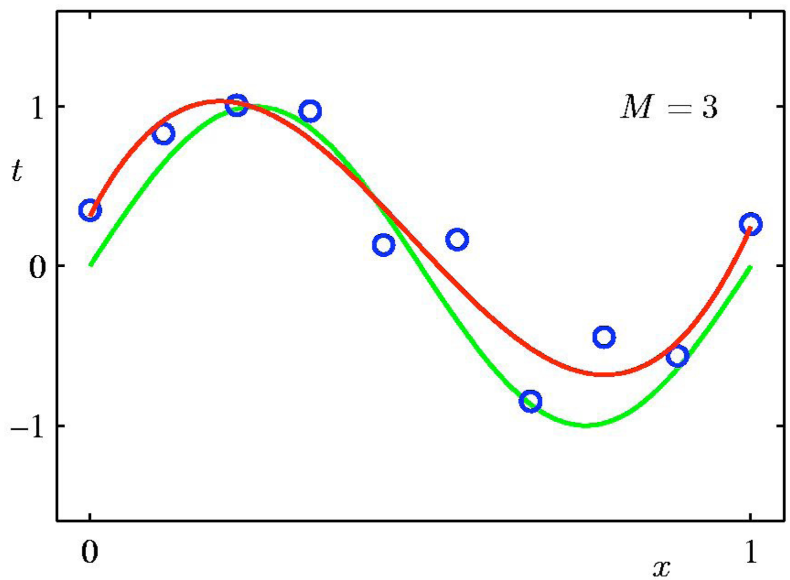
- **Reinforcement learning** is an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take *actions* in an *environment* so as to maximize some notion of cumulative *reward*.

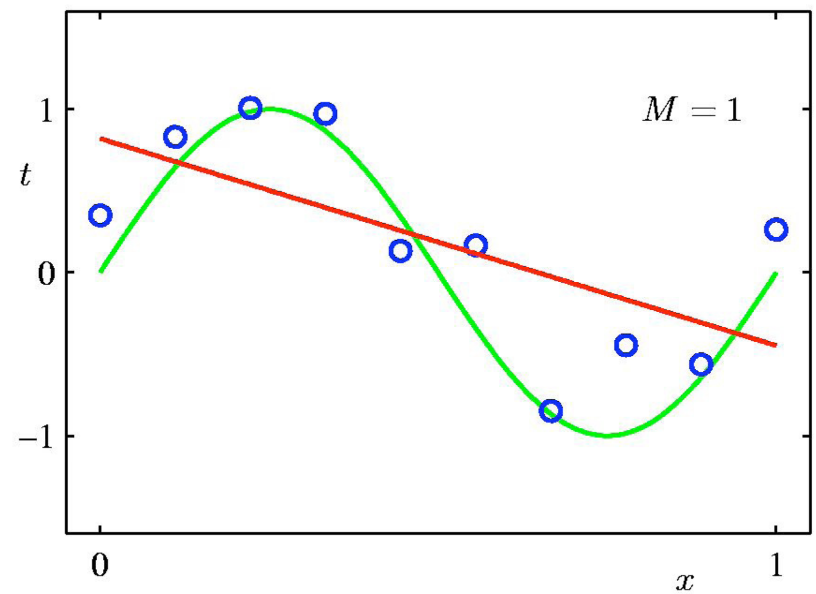
Semi-supervised learning

- **Semi-supervised learning** is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

- It falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data).

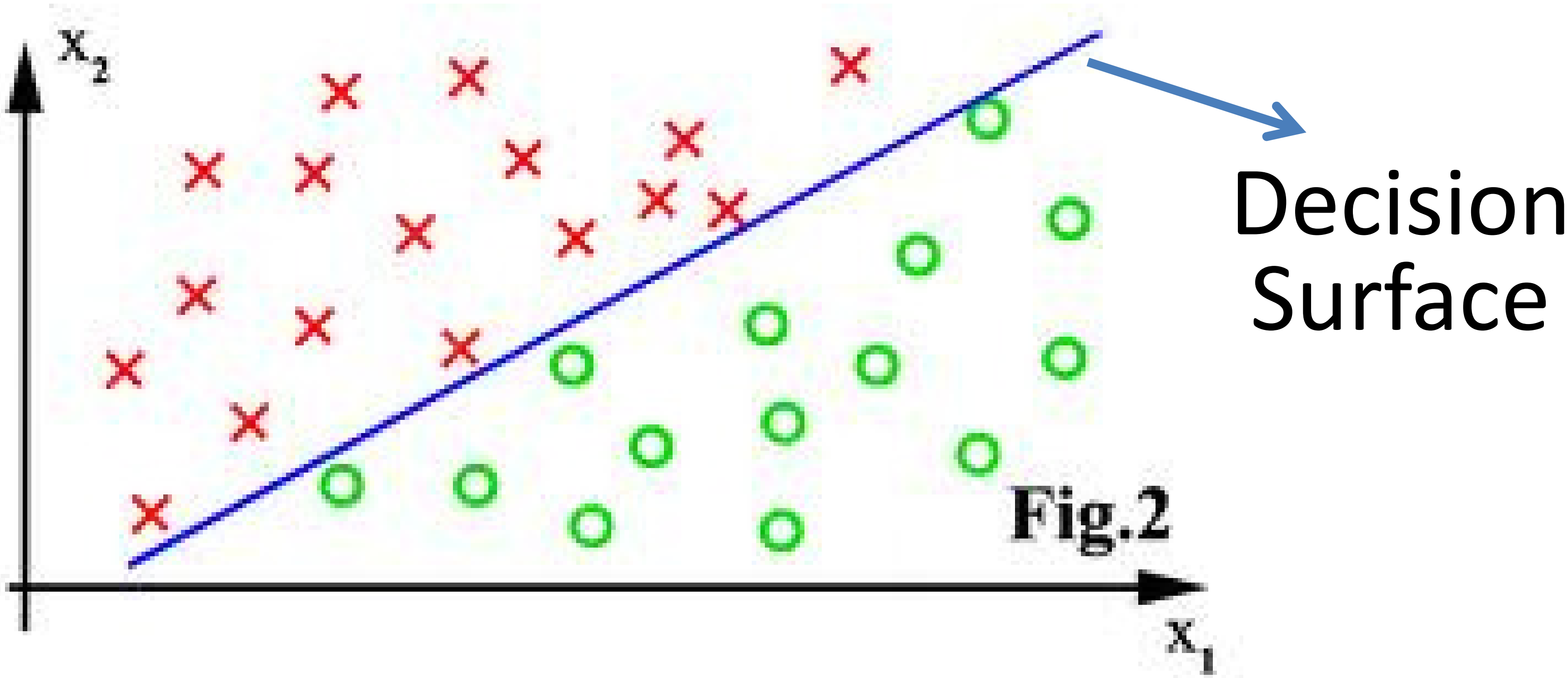
Regression





Similar to Curve Fitting Problem to a set of points…..

Classifier



Division of feature space to distinct regions by decision surfaces

Empirical Risk Minimization

- Every classifier / regressor does what is called as - `empirical risk minimization'

- Learning pertains to coming up with an architecture that can minimize a risk / loss function defined on the training /empirical data.

No- free lunch theorem

- There ain't such thing as free lunch --→ It is impossible to get nothing for something !

- In view of the no-free-lunch theorem it seems that one cannot hope for a classifier that would perform best on all possible problems that one could imagine.

Classifier taxonomy

- Generative classifiers
- Discriminative classifiers

- Types of generative classifier

[a] Parametric

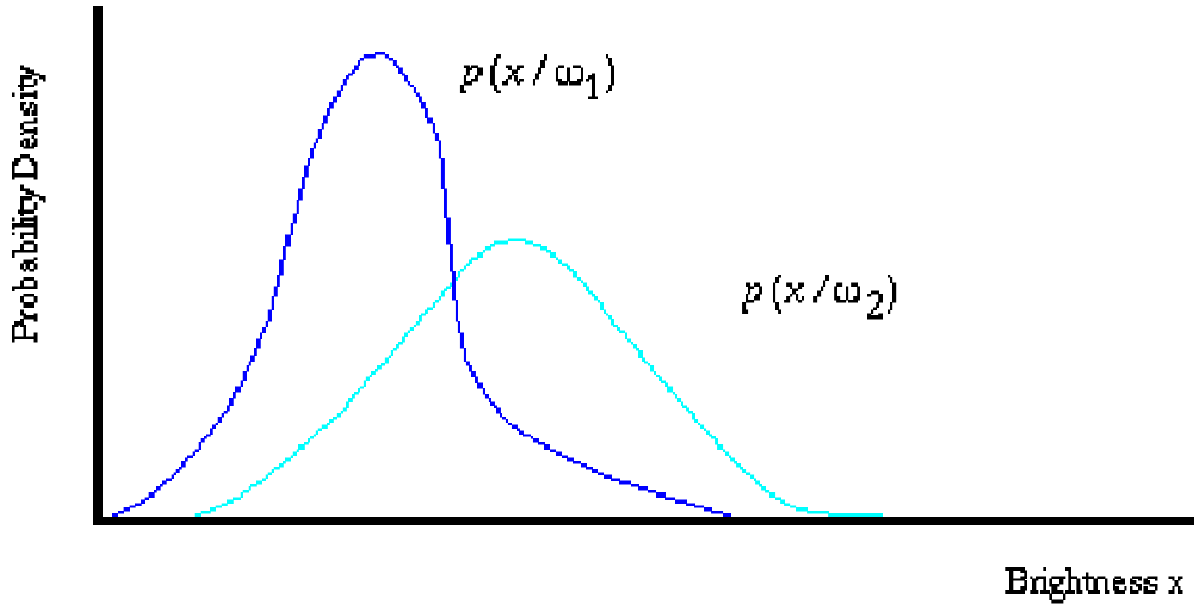
[b] Non-parametric

Generative classifier

- Samples of training data of a class assumed to come from a probability density function (class conditional pdf)

- If the form of pdf is assumed , such as uniform, gaussian, rayleigh, etc …one can estimate the parameters of the distribution.

- → Parametric classifier



Class conditional Density : pdf built using infinite samples of a given pattern / class.

In this figure, we have 2 pdf s corresponding to 2 classes w1 and w2 .

Feature x 'brightness' is used to construct the pdfs.

Generative classifier

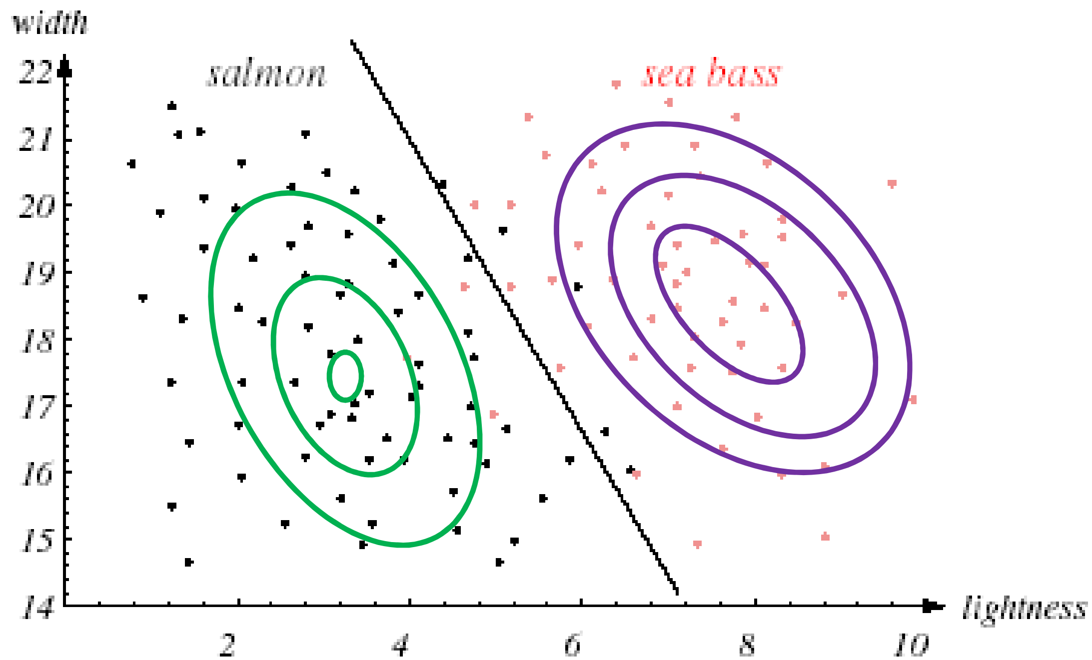


FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- One can as well assume to use the training data to build a pdf -→ Non parametric approach

- Discriminative classifier → No such assumption of data being drawn from an underlying pdf. Models the decision boundary by adaptive gradient descent techniques.

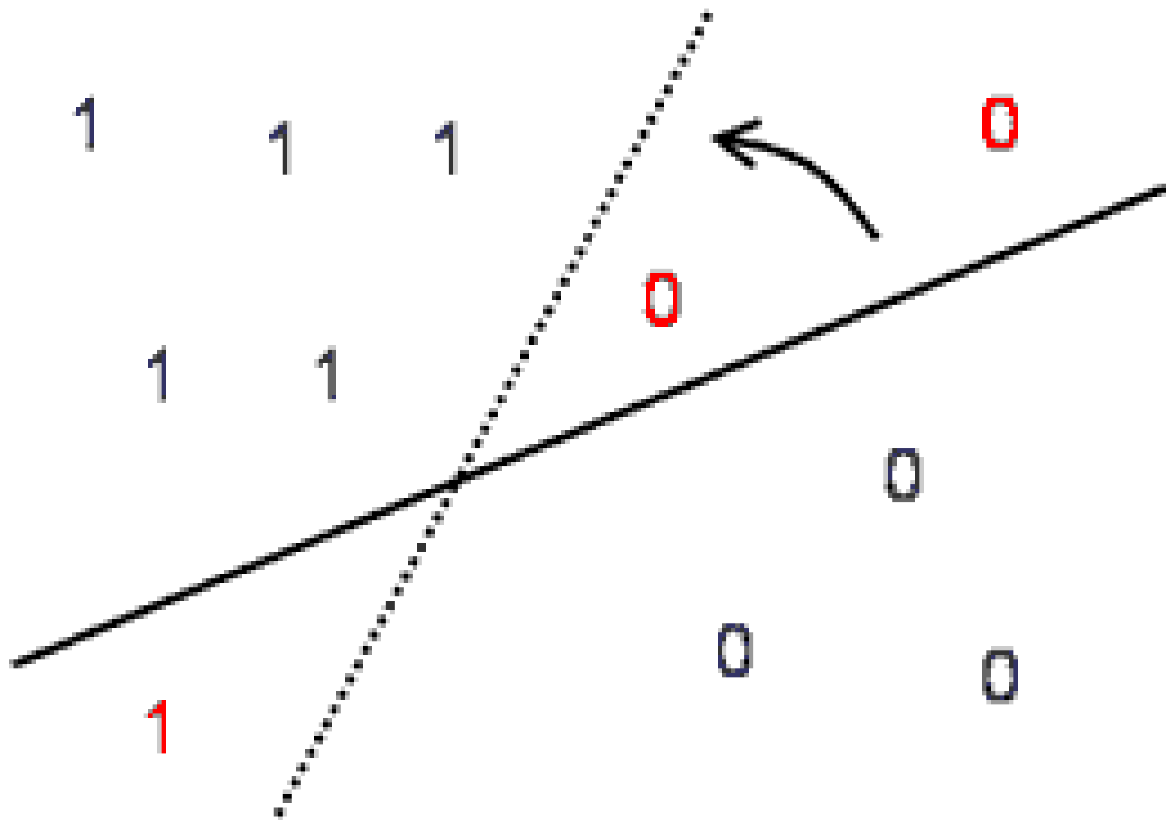
Discriminative Classifier

- Start with initial weights that define the decision surface
- Update the weights based on some optimization criterion….

- No need to model the distribution of samples of a given class…..class conditional density concept not required!

- Neural nets (such as MLP, Single layer perceptron, SVMs) fall in the category of discriminative classifiers.

Discriminative classifier



Linearly separable data

$w_1x_1+w_2x_2+b>0$

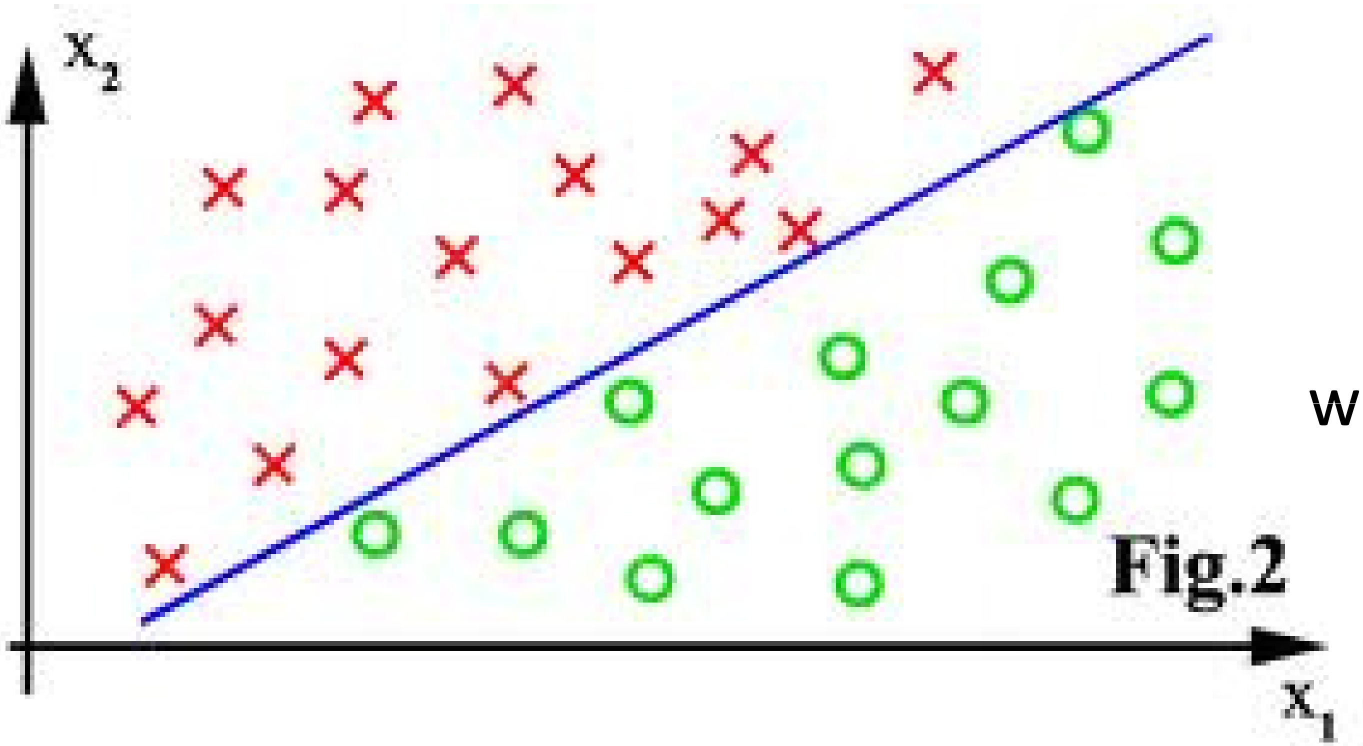


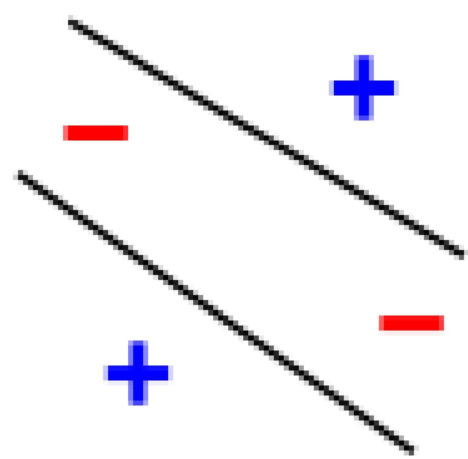
Fig.2

$w_1x_1+w_2x_2+b<0$

Linearly separable data

Separating line： $w_1x_1+w_2x_2+b=0$

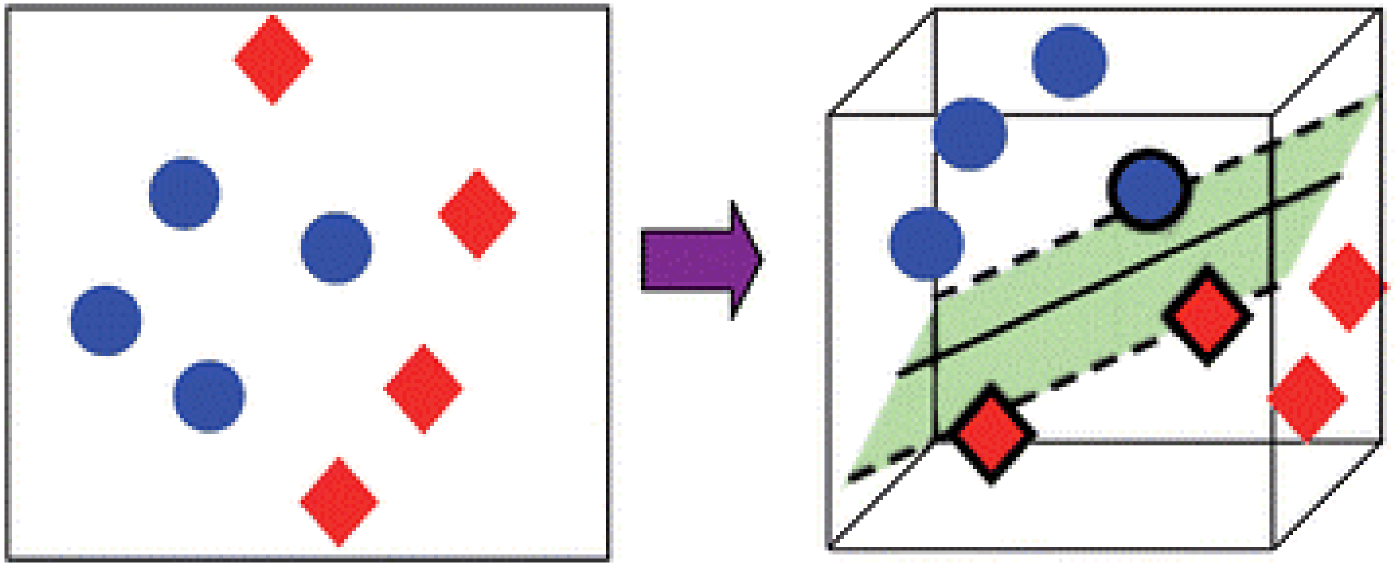
Non- linearly separable data



Covers Theorem

- The theorem states that given a set of training data that is not linearly separable, one can transform it into a training set that is linearly separable by mapping it into a possibly higher-dimensional space via some non-linear transformation.

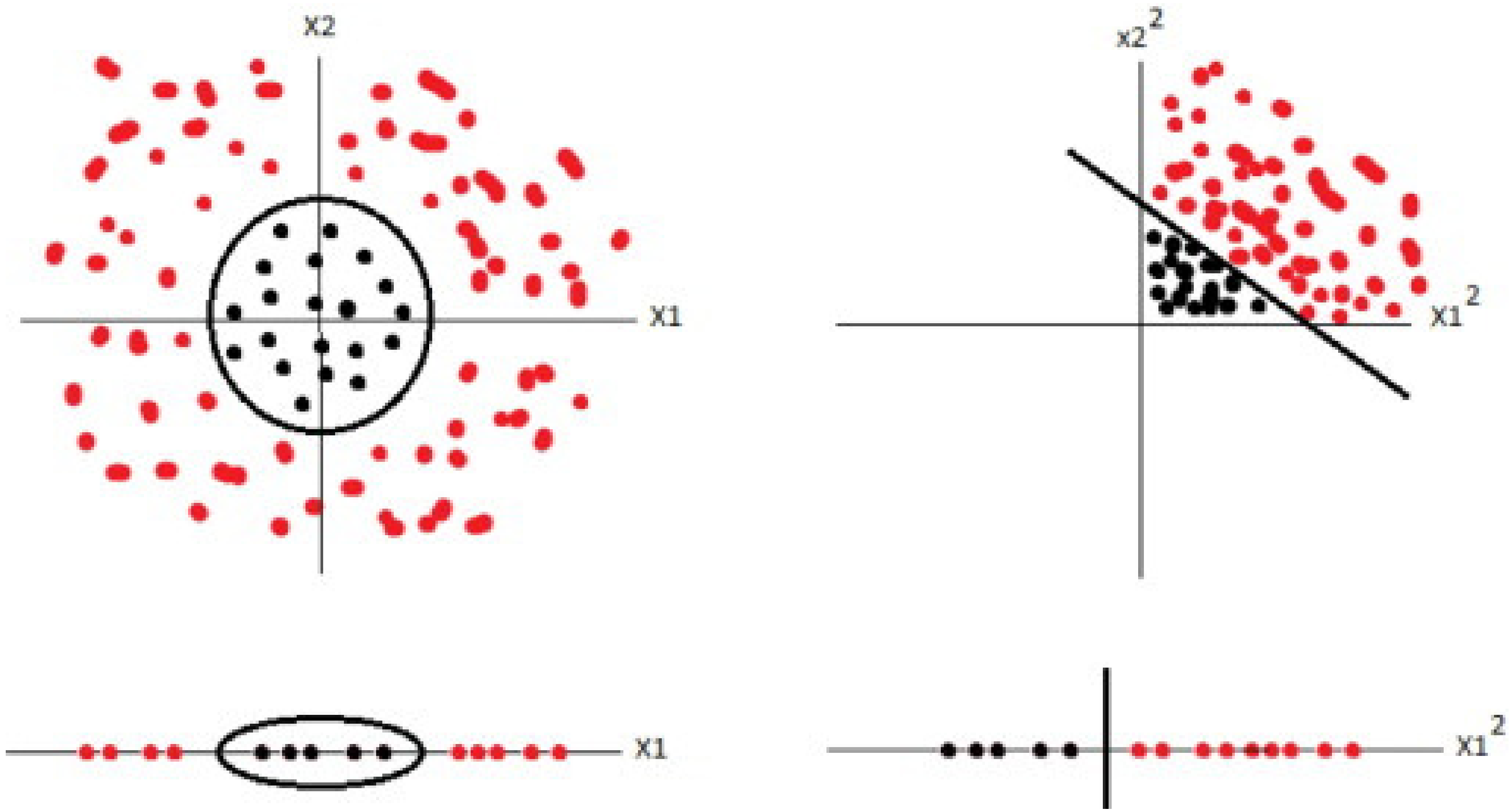
Cover's Theorem



(a) (b)

The samples of the original data is in 2D. After a non-linear transformation , it becomes linearly separable in three dimensions as shown in (b).

Cover's Theorem



Evaluation Metric

Consider scenario wherein a patient is screened for a disease.

Yes : Healthy
No: Diseased

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

TP : True positive
FN : False negative
TN : True Negative
FN : False Negative