Fundamentals of Machine Learning

Course 4232: Machine Learning

Dept. of Computer Science
Faculty of Science and Technology

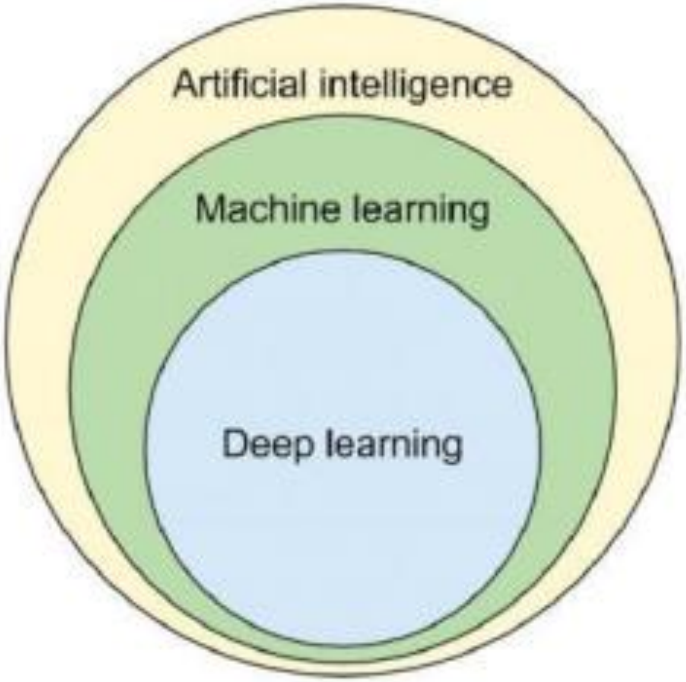
Lecturer No:	**1**	**Week No:**	**1**	**Semester:**	**Fall 23-24**
Instructor:	*Md Saef Ullah Miah (saef@aiub.edu)*				

What is learning?

- “Learning is any process by which a system improves performance from experience.” –Herbert Simon
- “Learning is constructing or modifying representations of what is being experienced.”
  –Ryszard Michalski
- “Learning is making useful changes in our minds.” –Marvin Minsky

What Is Machine Learning (ML)?

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.



Why “Learn” ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to “learn” to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

Why learn?

- Build software agents that can adapt to their users or to other software agents *or to changing environments*
  - Personalized news or mail filter
  - Personalized tutoring
  - Mars robot
- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task
  - Large, complex AI systems cannot be completely derived by hand and require dynamic updating to incorporate new information.
- Discover new things or structure that were previously unknown to humans
  - Examples: data mining, scientific discovery

Related Disciplines

The following are close disciplines:

- Artificial Intelligence
  - Machine learning deals with the learning part of AI
- Pattern Recognition
  - Concentrates more on “tools” rather than theory
- Data Mining
  - More specific about discovery

The following are useful in machine learning techniques or may give insights:

- Probability and Statistics
- Information theory

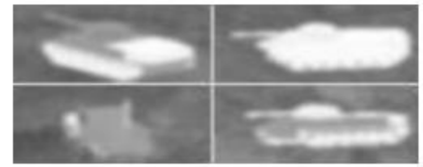
- Psychology (developmental, cognitive)
- Neurobiology
- Linguistics
- Philosophy

Data Mining

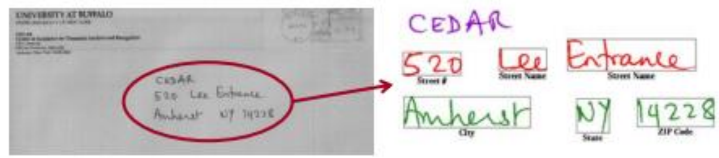
- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Control, robotics, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Spam filters, intrusion detection
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- ...

Examples of pattern recognition problems

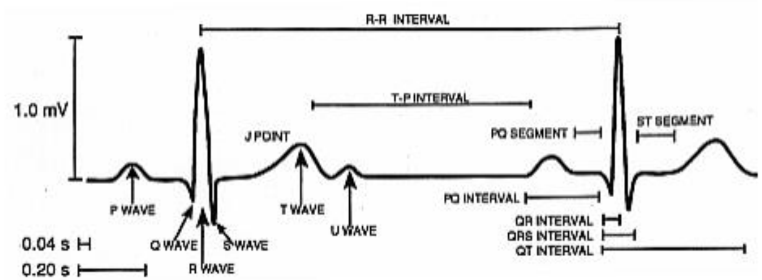
- **Machine vision**
  - Visual inspection, ATR
  - Imaging device detects ground target
  - Classification into "friend" or "foe"



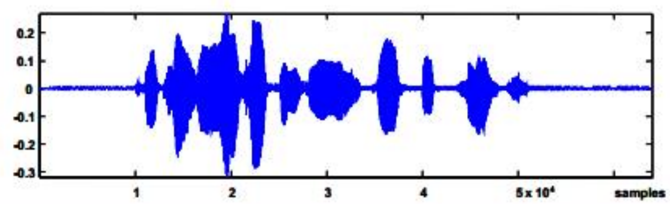
- **Character recognition**
  - Automated mail sorting, processing bank checks
  - Scanner captures an image of the text
  - Image is converted into constituent characters



- **Computer aided diagnosis**
  - Medical imaging, EEG, ECG signal analysis
  - Designed to assist (not replace) physicians
  - Example: X-ray mammography
    - 10-30% false negatives in x-ray mammograms
    - 2/3 of these could be prevented with proper analysis



- **Speech recognition**
  - Human Computer Interaction, Universal Access
  - Microphone records acoustic signal
  - Speech signal is classified into phonemes and/or words



History of Machine Learning

- 1950s
  - **Samuel’s checker player**
- 1960s:
  - **Neural networks: Perceptron**
  - **Minsky and Papert prove limitations of Perceptron**
- 1970s:
  - Expert systems and the knowledge acquisition bottleneck
  - Mathematical discovery with AM
  - Symbolic concept induction

History of Machine Learning (cont.)

- 1980s:
  - **Resurgence of neural networks (connectionism, backpropagation)**
  - Advanced decision tree and rule learning
  - Learning, planning and problem solving
  - Utility theory
  - Analogy
- 1990s
  - **Data mining**
  - **Reinforcement learning (RL)**
  - **Inductive Logic Programming (ILP)**
  - **Ensembles: Bagging, Boosting, and Stacking**

History of Machine Learning (cont.)

- 2000s
  - **Kernel methods**
    - Support vector machines
  - **Graphical models**
  - Statistical relational learning
  - Transfer learning

- Applications
  - Adaptive software agents and web applications
  - Learning in robotics and vision
  - E-mail management (spam detection)
  - …

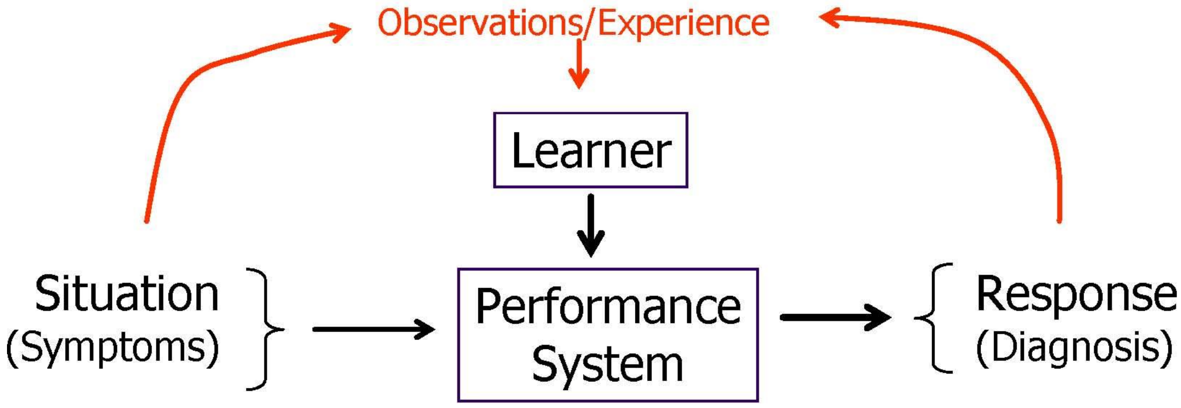
What is Machine Learning ?

- A computer program **M** is said to learn from experience **E** with respect to some class of tasks **T** and performance **P**, if its performance as measured by **P** on tasks in **T** in an environment **Z** improves with experience **E.**

- Example:
  - **T**: Cancer diagnosis
  - **E**: A set of diagnosed cases
  - **P**: Accuracy of diagnosis on new cases
  - **Z**: Noisy measurements, occasionally misdiagnosed training cases
  - **M**: A program that runs on a general purpose computer; the learner

What is Machine Learning ?

- A computer program **M** is said to learn from experience **E** with respect to some class of tasks **T** and performance **P**, if its performance as measured by **P** on tasks in **T** in an environment **Z** improves with experience **E**.



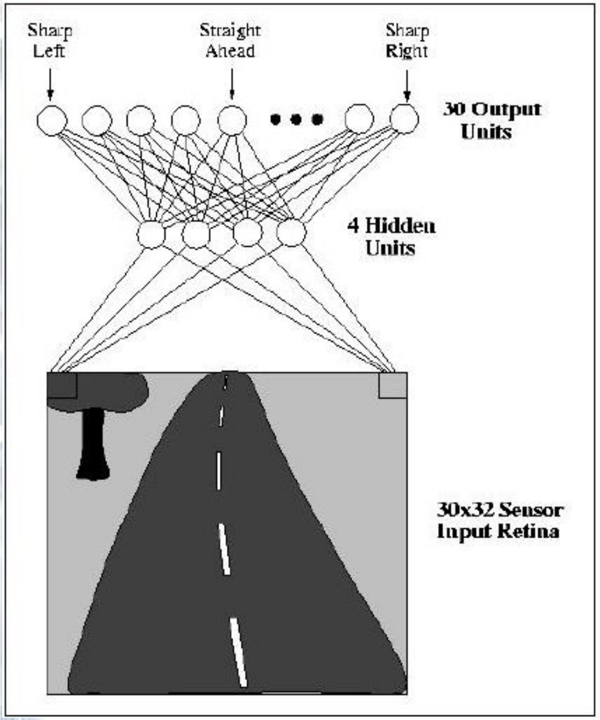
Why Machine Learning ?

- Solving tasks that required a system to be adaptive
  - Speech, face, or handwriting recognition
  - Environment changes over time

- Understanding human and animal learning
  - How do we learn a new language ? Recognize people ?
- Some task are best shown by demonstration
  - Driving a car, or, landing an airplane

- Objective of Real Artificial Intelligence:
  - “If an intelligent system–brilliantly designed, engineered and implemented– cannot learn not to repeat its mistakes, it is not as intelligent as a worm or a sea anemone or a kitten.” (Oliver Selfridge)

Tasks too Hard to Program



ALVINN [Pomerleau] drives 70 MPH on highways

Kinds of Learning

- Based on the information available
  - Association
  - Supervised Learning
    - Classification
    - Regression
  - Reinforcement Learning
  - Unsupervised Learning
  - Semi-supervised learning

- Based on the role of the learner
  - Passive Learning
  - Active Learning

Major paradigms of machine learning

- **Rote learning** – "Learning by memorization."
  - Employed by first machine learning systems, in 1950s
    - Samuel's Checkers program

- **Supervised learning** – Use specific examples to reach general conclusions or extract general rules
    - Classification (Concept learning)
    - Regression

- **Unsupervised learning (Clustering)** – Unsupervised identification of natural groups in data

- **Reinforcement learning**– Feedback (positive or negative reward) given at the end of a sequence of steps

- **Analogy** – Determine correspondence between two different representations

- **Discovery** – Unsupervised, specific goal not given

- …

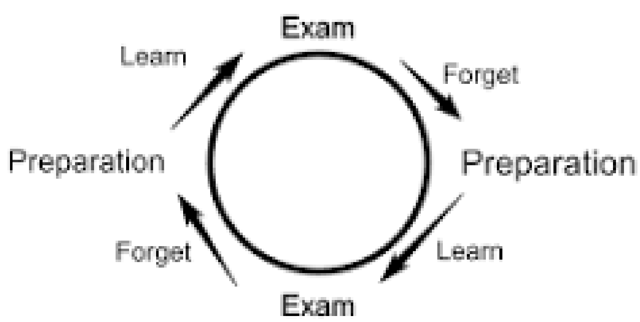
Rote Learning is Limited

- Memorize I/O pairs and perform exact matching with new inputs

- If a computer has not seen the precise case before, it cannot apply its experience

- We want computers to "generalize" from prior experience
  - Generalization is the most important factor in learning

The Rote Loop



The inductive learning problem

- Extrapolate from a given set of examples to make accurate predictions about future examples
- Supervised versus unsupervised learning
  - Learn an unknown function f(X) = Y, where X is an input example and Y is the desired output.
  - **Supervised learning** implies we are given a **training set** of (X, Y) pairs by a “teacher”
  - **Unsupervised learning** means we are only given the Xs.
  - **Semi-supervised learning**: mostly unlabelled data

Learning Associations

- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

  Example: P (sugar | tea) = 0.7

Supervised Learning

- Training experience: a set of labeled examples of the form

$$< x_1, x_2, \ldots, x_n, y >$$

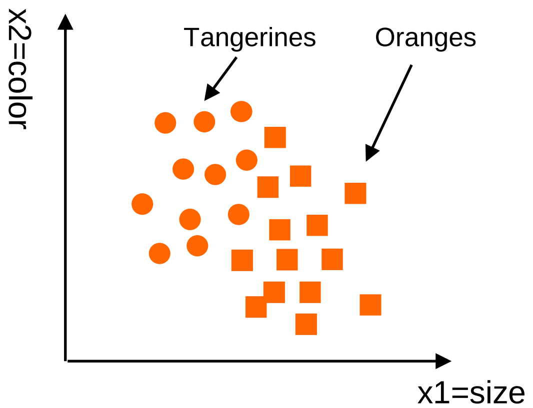
- where x_j are values for input variables and y is the output

- This implies the existence of a "teacher" who knows the right answers

- What to learn: A function $f : X_1 \times X_2 \times \ldots \times X_n \rightarrow Y$, which maps the input variables into the output domain

- Goal: minimize the error (loss function) on the test examples

Types of supervised learning

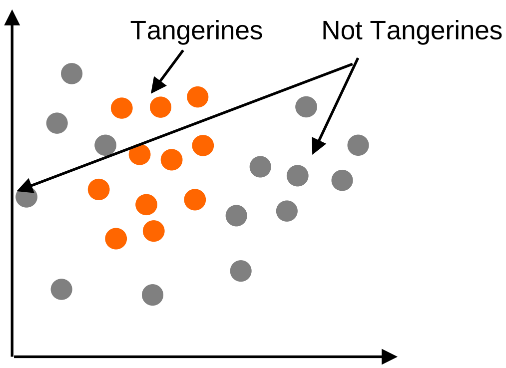


a) **Classification**:

- We are given the label of the training objects: {(x1,x2,y=T/O)}

- We are interested in classifying future objects: (x1',x2') with the correct label.

  I.e. Find y' for given (x1',x2').



b) **Concept Learning**:

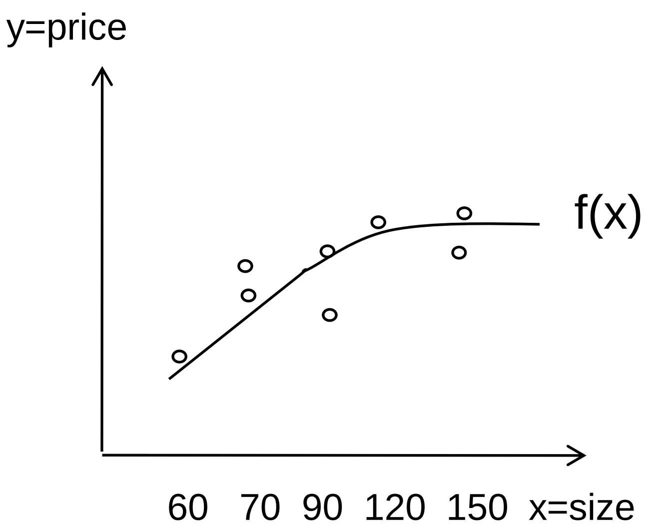
- We are given positive and negative samples for the concept we want to learn (e.g.Tangerine): {(x1,x2,y=+/-)}

- We are interested in classifying future objects as member of the class (or positive example for the concept) or not.

  I.e. Answer +/- for given (x1',x2').

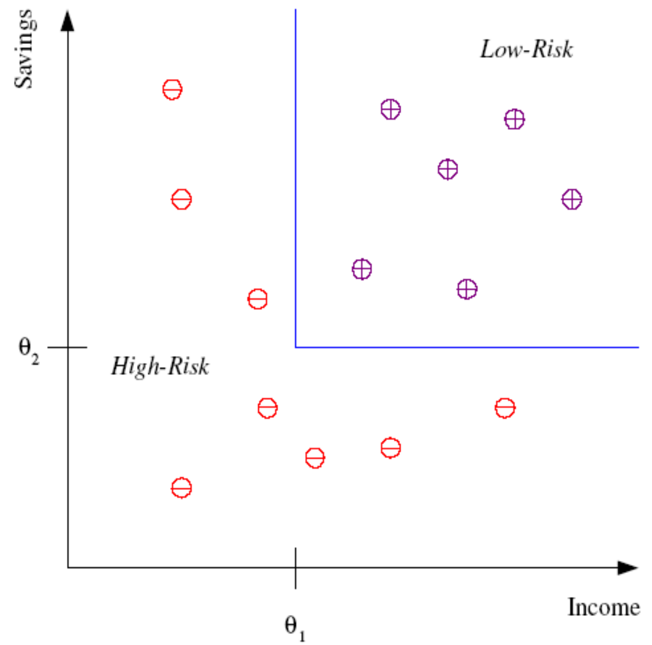
Types of Supervised Learning

- Regression
  - Target function is continuous rather than class membership
  - For example, you have some the selling prices of houses as their sizes (sq-mt) changes in a particular location that may look like this. **You may hypothesize that the prices are governed by a particular function f(x).** Once you have this function that “explains” this relationship, you can guess a given house’s value, given its sq-mt. **The learning here is the selection of this function f() .** Note that the problem is more meaningful and challenging if you imagine several input parameters, resulting in a multi-dimensional input space.



Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2
THEN low-risk ELSE high-risk

Classification: Applications

- Pattern Recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc

Face Recognition

Training examples of a person

Test images

ORL dataset,
AT&T Laboratories, Cambridge UK

Supervised Learning: Uses

- Prediction of future cases: Use the rule or model to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning “what normally happens”
- Training experience: no output, unlabeled data
- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation in CRM
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs

Reinforcement Learning

- Training experience: interaction with an environment; learning agent receives a numerical reward
  - Learning to play chess: moves are rewarded if they lead to WIN, else penalized
  - No supervised output but delayed reward

- What to learn: a way of behaving that is very rewarding in the long run - Learning a policy: A sequence of outputs

- Goal: estimate and maximize the long-term cumulative reward
- Credit assignment problem
- Robot in a maze, game playing
- Multiple agents, partial observability, ...

Passive Learning and Active Learning

- Traditionally, learning algorithms have been passive learners, which take a given batch of data and process it to produce a hypothesis or a model

- Data → Learner → Model
- Active learners are instead allowed to query the environment
  - Ask questions
  - Perform experiments
- Open issues: how to query the environment optimally? how to account for the cost of queries?

Learning: Key Steps

data and assumptions

– what data is available for the learning task?

– what can we assume about the problem?

• representation

– how should we represent the examples to be classified

• method and estimation

– what are the possible hypotheses?

– what learning algorithm to use to infer the most likely hypothesis?

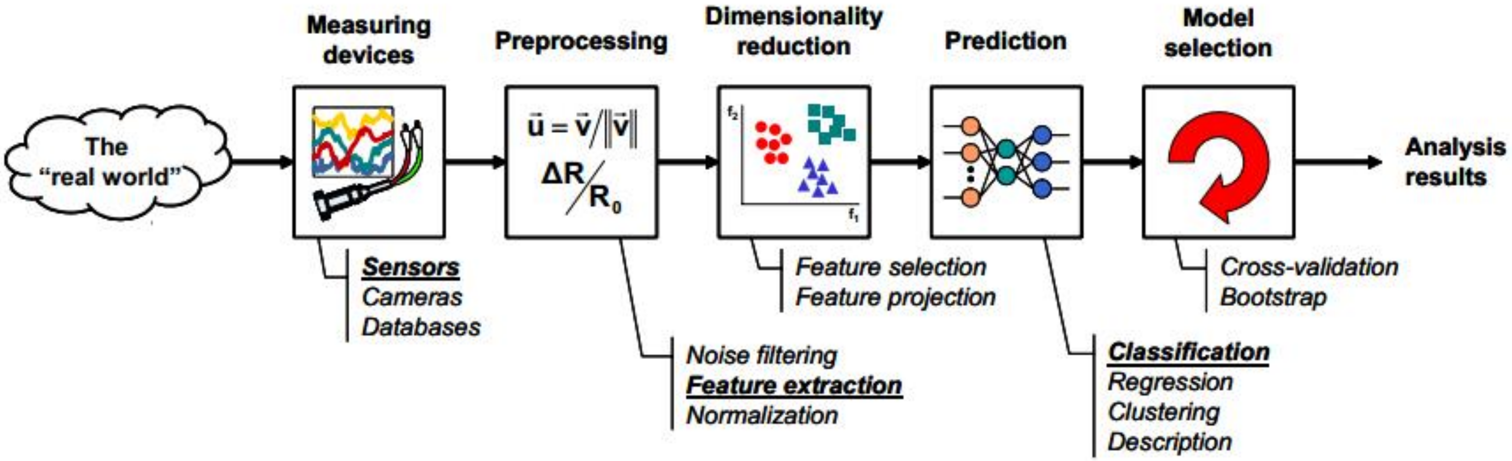
– how do we adjust our predictions based on the feedback?

• evaluation

– how well are we doing?

Components of a pattern recognition system

- **A basic pattern classification system contains**
  - A sensor
  - A preprocessing mechanism
  - A feature extraction mechanism (manual or automated)
  - A classification algorithm
  - A set of examples (training set) already classified or described

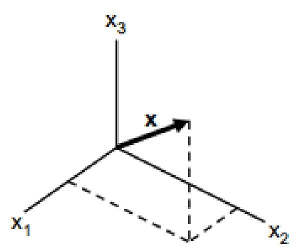


Features and patterns (1)

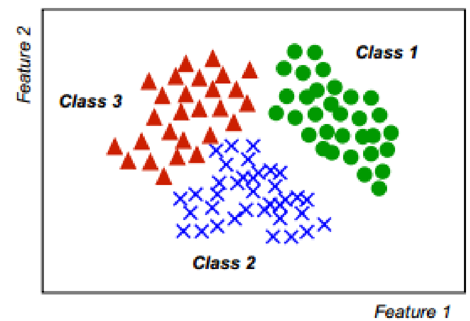
- **Feature**
  - Feature is any distinctive aspect, quality or characteristic
    - Features may be symbolic (i.e., color) or numeric (i.e., height)
  - Definitions
    - The combination of d features is represented as a d-dimensional column vector called a **feature vector**
    - The d-dimensional space defined by the feature vector is called the **feature space**
    - Objects are represented as points in feature space. This representation is called a **scatter plot**

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

Feature vector



Feature space (3D)

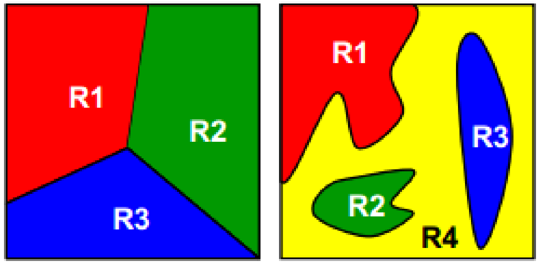


Scatter plot (2D)

- **Pattern**
  - Pattern is a composite of traits or features characteristic of an individual
  - In classification tasks, a pattern is a pair of variables $\{x,\omega\}$ where
    - ***x*** is a collection of observations or features (feature vector)
    - ***ω*** is the concept behind the observation (label)

Classifiers

- **The task of a classifier is to partition feature space into class-labeled decision regions**
  - Borders between decision regions are called **decision boundaries**
  - The classification of feature vector ***x*** consists of determining which decision region it belongs to, and assign ***x*** to this class

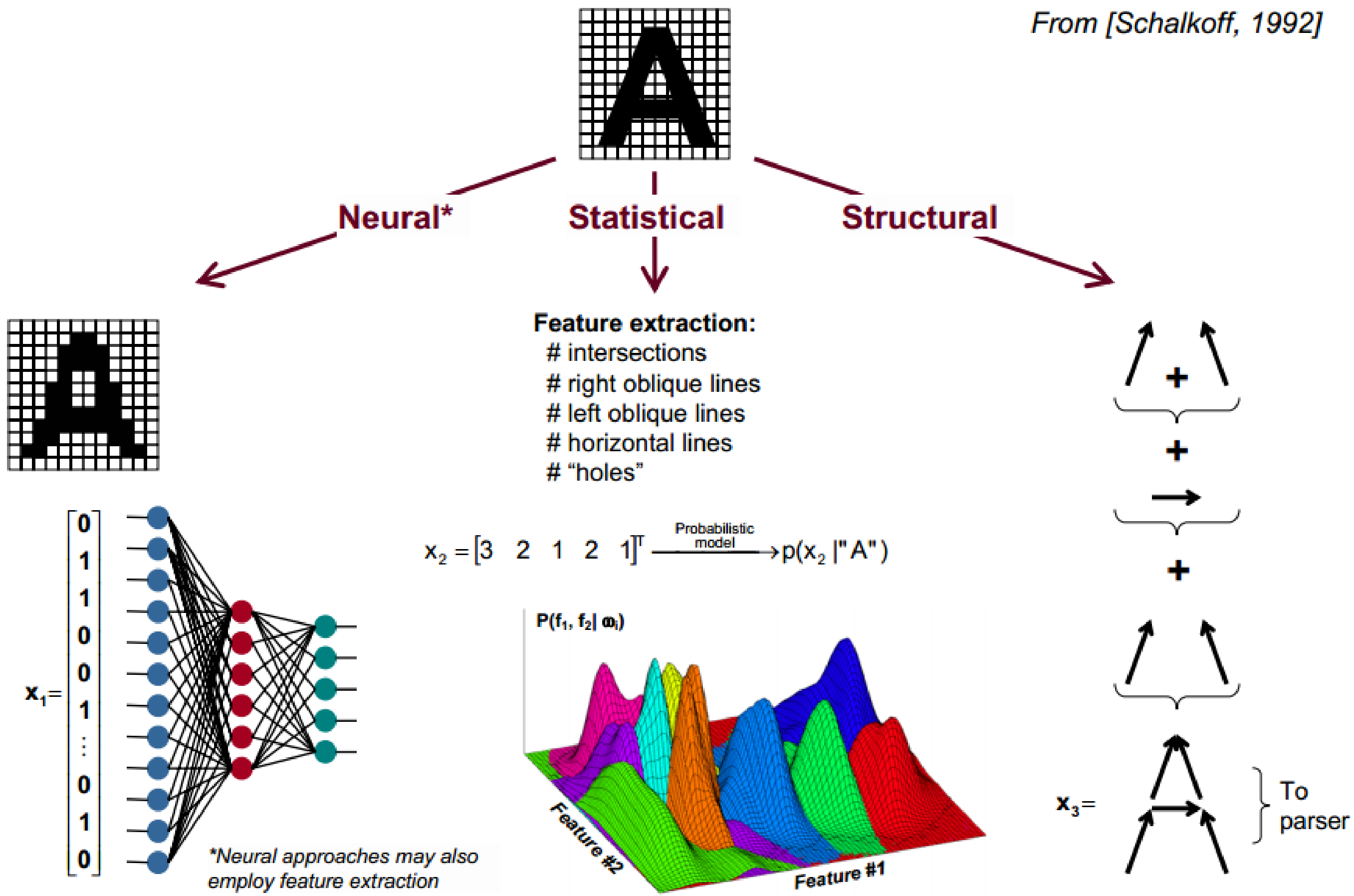


Pattern recognition approaches

- **Statistical (StatPR)**
  - Patterns classified based on an underlying statistical model of the features
    - The statistical model is defined by a family of **class-conditional probability** density functions $Pr(x|c_i)$ (Probability of feature vector $\boldsymbol{x}$ given class $\boldsymbol{c_i}$)
- **Neural (NeurPR)**
  - Classification is based on the response of a network of processing units (neurons) to an input stimuli (pattern)
    - "Knowledge" is stored in the **connectivity and strength of the synaptic weights**
  - NeurPR is a trainable, non-algorithmic, black-box strategy
  - NeurPR is very attractive since
    - it requires minimum a priori knowledge
    - with enough layers and neurons, an ANN can create **any** complex decision region
- **Syntactic (SyntPR)**
  - Patterns classified based on measures of structural similarity
    - "Knowledge" is represented by means of **formal grammars or relational descriptions** (graphs)
  - SyntPR is used not only for classification, but also for description
    - Typically, SyntPR approaches formulate hierarchical descriptions of complex patterns built up from simpler sub patterns

Example: neural, statistical and structural OCR

From [Schalkoff, 1992]



Neural approaches may also employ feature extraction

The pattern recognition design cycle (1)

- **Data collection**
  - Probably the most time-intensive component of a PR project
  - How many examples are enough?
- **Feature choice**
  - Critical to the success of the PR problem
    - "Garbage in, garbage out"
  - Requires basic prior knowledge
- **Model choice**
  - Statistical, neural and structural approaches
  - Parameter settings
- **Training**
  - Given a feature set and a "blank" model, adapt the model to explain the data
  - Supervised, unsupervised and reinforcement learning
- **Evaluation**
  - How well does the trained model do?
  - Overfitting vs. generalization

Evaluation of Learning Systems

- Experimental
  - Conduct controlled cross-validation experiments to compare various methods on a variety of benchmark datasets.
  - Gather data on their performance, e.g. test accuracy, training-time, testing-time…
  - Analyze differences for statistical significance.

- Theoretical
  - Analyze algorithms mathematically and prove theorems about their:
    - Computational complexity
    - Ability to fit training data
    - Sample complexity (number of training examples needed to learn an accurate function)

Measuring Performance

Performance of the learner can be measured in one of the following ways, as suitable for the application:

- Classification Accuracy
  - Number of mistakes
  - Mean Squared Error
  - Loss functions
- Solution quality (length, efficiency)
- Speed of performance
- …

Textbook/ Reference Materials

1. Introduction to Machine Learning (MIT Press) by Ethem Alpaydin