COMP4388: **MACHINE LEARNING**

Week 2

Dr. Radi Jarrar
Department of Computer Science

MACHINE LEARNING

What is Machine Learning?

- The term 'Machine Learning' came into wide-spread use following a workshop by that name in 1980
- Machine learning: and artificial intelligence approach, R. Michalski, J. Carbonell, and T. Mitchell
- Later, identified as a research field in its own

What is Machine Learning?

- The task of building knowledge and storing it in some form in the computer (such as algorithm or mathematical model) that can help detect patterns
- It is concerned with predicting a particular outcome given some data
- Training set is used to train the algorithm
- The algorithm, typically, has a number of parameters that are learnt from the data

What is Machine Learning?

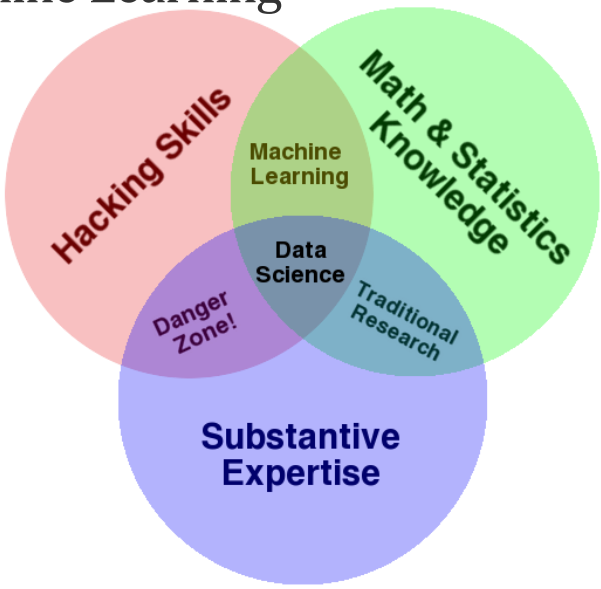
- Machine Learning algorithms are applied to situations where it is very challenging to define rules
  - Image classification
  - Face detection
  - Speech recognition
  - Financial prediction

What is Machine Learning? (TM)

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E
- Machine Learning:
  - Study of algorithm that
    - improve their performance **P**
    - at some task **T**
    - with experience **E**

- well-defined training task: <P,T,E>

Machine Learning



What is Machine Learning?

- Machine Learning exists at the intersection of Mathematics & Statistics with Software Engineering & Computer

- **Machine Learning** is concerned with *teaching* computers something about the world, so that they can think more clearly about the world in order to make better decisions

What is Machine Learning?

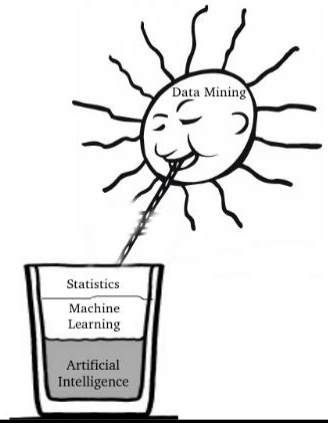
- The learning occurs by extracting as much information from the data as possible through algorithms that parse data and distinguish signal from noise
- The algorithm is capable to distinguish patterns (signals in specific format), it decides that everything else that is left over is noise

What is Machine Learning?

- Machine Learning techniques often referred as Pattern Recognition Algorithms
- Observing data, learning from it, then automate the process of recognition is the backbone of Machine Learning

Machine Learning and other disciplines

- Machine Learning is related to other disciplines such as Artificial Intelligence, Statistics, and Data Mining
- They are still have some differences



ML & Artificial Intelligence (AI)

- AI is the broad family of Machine Learning
- AI is the study of how to create intelligent agents
- How to program the computer to behave and perform as a intelligent agent (i.e., a human)
- This may not involve training or learning from data!
- A large area within AI is ML
  - Involves the study of algorithms that can extract information automatically and learn patterns from

ML & Statistics

- Statistics is concerned with learning something interpretable from data; whereas Machine Learning is concerned with turning data into something practical and usable
- Machine learning is concerned with teaching *computers* something about the world, so that they can use that knowledge to perform other tasks. Statistics is more concerned with developing tools for teaching *humans* something about the world, so that they can think more clearly about the world in order to make better decisions.

ML & Data Mining

- Often confused!
- Machine learning can be seen as a pre-requisite for Data mining
- Machine Learning focuses on prediction based on 'known' properties learnt from the training data
- Data Mining is the identification of correlations and patterns within data. It focuses on discovery of 'unknown' properties and patterns of data

Example 1: Spam Filter

- Email clients identify spam emails (i.e., Junk emails) using Machine Learning algorithms (**spam filter**)
- Previously relied on Hand-coded patterns to identify spams emails
- Issues
  - Hard to maintain
  - Offers insufficient flexibility: one's man spam, is another man's ham!

Example 1: Spam Filter

- Input: email
- Output: spam/ham
- Setup:
  - A large collection of emails is 'annotated' or 'labeled' as spam or ham
  - Annotation is done manually!
- Task
  - Predict on new emails whether they are spam or ham
- Features
  - Attributes that are used to make spam or ham decisions
  - E.g., Words: FREE!, Donation, Visa, Prize
    Text patterns: BLOCK LETTERS, $, …

Example 1: Spam Filter (2)

- Input: email
- Output: spam/ham
- Setup:
  - A large collection of emails is 'annotated' or 'labeled' as spam or ham
  - Annotation is done manually!
- Task
  - Predict on new emails whether they are spam or ham
- Features
  - Attributes that are used to make spam or ham decisions
  - E.g., Words: FREE!, Donation, Visa, Prize Text patterns: BLOCK LETTERS, $, …

Dear Sir,

Please note that by replying to this email you will get a huge prize up to 100,00$ IN CASH! PLEASE REPLY THIS EMAIL FOR FREE!

Hi there,

As discussed on the phone, I will do my best to continue on the matter discussed.

Sorry for any delay! Cheers!

Example 1: Spam Filter (2)

- Input: email
- Output: spam/ham
- Setup:
  - A large collection of emails is 'annotated' or 'labeled' as spam or ham
  - Annotation is done manually!
- Task
  - Predict on new emails whether they are spam or ham
- Features
  - Attributes that are used to make spam or ham decisions
  - E.g., Words: FREE!, Donation, Visa, Prize Text patterns: BLOCK LETTERS, $, …

Dear Sir,

Please note that by replying to this email you will get a huge prize up to 100,00$ IN CASH! PLEASE REPLY THIS EMAIL FOR FREE!

SPAM!

Hi there,

As discussed on the phone, I will do my best to continue on the matter discussed.

Sorry for any delay! Cheers!

Example 1: Spam Filter (2)

- Input: email
- Output: spam/ham
- Setup:
  - A large collection of emails is 'annotated' or 'labeled' as spam or ham
  - Annotation is done manually!
- Task
  - Predict on new emails whether they are spam or ham
- Features
  - Attributes that are used to make spam or ham decisions
  - E.g., Words: FREE!, Donation, Visa, Prize
    Text patterns: BLOCK LETTERS, $, …

Dear Sir,

Please note that by replying to this email you will get a huge prize up to 100,00$ IN CASH! PLEASE REPLY THIS EMAIL FOR FREE!

SPAM!

Hi there,

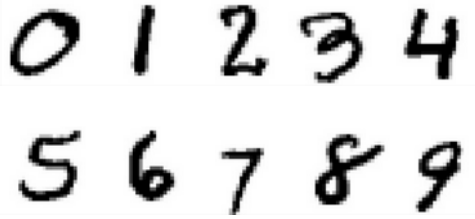
As discussed on the phone, I will do my best to continue on the matter discussed.

Sorry for any delay! Cheers!

Ham!

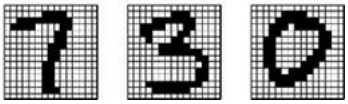
Example 2: Handwritten digit recognition

- Input: images of handwritten digits
- Output: the digit in images
- Setup:
  - A large collection of image examples 'labeled' with the correct numbers
- Task
  - Predict on new images with their numbers
- Features
  - Image pixels!



Example 2: Handwritten digit recognition

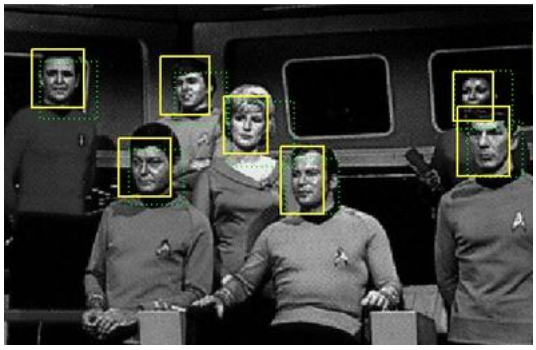
- Each image is 15 x 15 pixels



- The input 'features' of each image can be represented as a vector $x \in \mathcal{R}^{225}$
- The classification problem is a function f(x) such that $f : x \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$

Example 3: Face detection

- Detect faces within images
- Detection might specify frontal or side-face

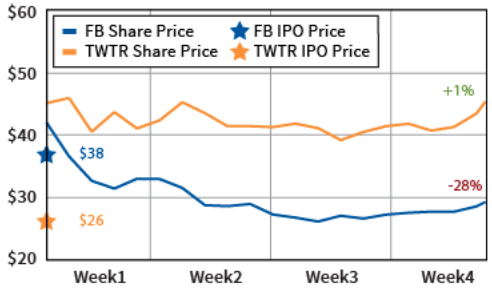


Example 4: Text classification

- Classify text documents into pre-defined categories
- Documents might be classified into a single or multiple categories
- Helps in
  - Archiving old texts
  - Classifying news
  - . . .

Example 5: Stock price prediction

- Predict stock price at future
- This task is a bit different as the output is continuous valued (rather than category or discrete value); i.e., Regression

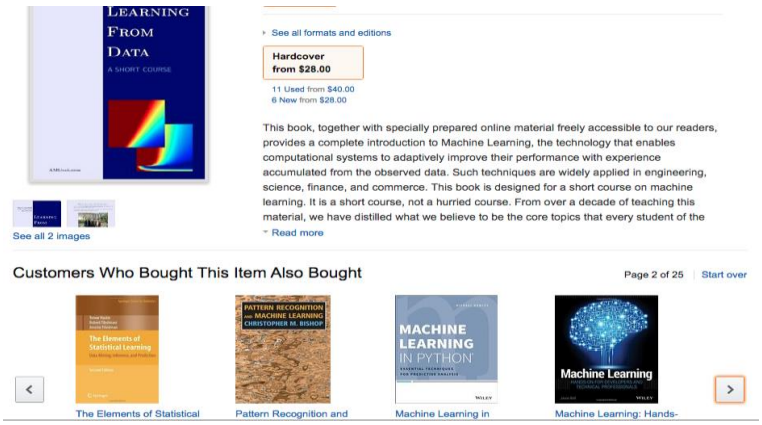


Example 6: Image classification

- Classify images based on their contents (e.g., indoor vs. outdoor)
- Training annotated images are used to generate the classifier

Example 7: Recommendation systems

- Give users suggestions on items based on previous data (history) from other users (or the same user!)



FEATURES AND TYPES OF LEARNINGS

Features

- Machine learning algorithms receives data and it manages to classify data
- There is a “garbage in; garbage out” aspect to machine learning
- Meaning, the data should be good, clean, and representative in order to make the algorithm learn how to apply hypothesis and learn data
- Feature engineering is crucial to machine learning

Features (2)

- Can be seen as the language that we use to describe certain object
- E.g., email, image, historical stock information,…
- Through proper methods/algorithms, features are extracted from the objects from which the learner is being trained
- They determine the much of the success of the model: a model is only as good as its features

Features (3)

- E.g., imagine you have a 100x100 pixel image, an easy feature representation of the image could be a 30,000 dimensional feature vector
- Each dimension corresponds to the red, blue, or green component of some pixel in the image
- The first element in the vector could be the red element of the first vector, the second is the blue of the same pixel, and so on
- This is the pixel representation of the image

Features (4)

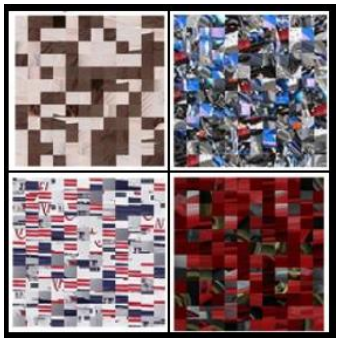
- Any problems in that representation?

Features (4)

- Any problems in that representation?
- It actually does not consider locality information in am image
- The learner doesn't care about features, they care about feature values

Features (5)

- The spatial location problem can be taken care of by using patch representation (i.e., an image is represented by several rectangular blocks each represents a point of interest)



Features (6)

- Text (as well images) can be represented as Bags-of-Words, which considers documents as collection words regardless of their location in the document
- The terms that occur the highest will be considered as features and their values is how many times those words occurred in a document

Features (7)

- A good representation of features should be considered
- Noisy/irrelevant features should be eliminated
- Irrelevant feature: a feature that is completely uncorrelated to the prediction task
- A feature “F” whose expectation does not depend on a specific label, might be irrelevant
- E(f | y) = E(f)
- For instance, using the feature “gender” to predict whether the course review is positive or negative

Features (8)

- Another issue is data redundancy
- This is the case if two features are highly correlated
- These features might be harmful to learners (especially when using Decision Trees as they tend to select the features that are highly correlate to the class labels)
- Different metrics can be used such as mutual information, correlation coefficient scores, …

Features (9)

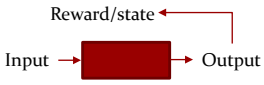
- Missing values is another problem
- Some learner do not accept missing values
- Easiest method is to remove rows with missing features (however, this might remove some other good values of other features!)
- Impute missing values:
  - Add a constant value (such as 0) to distinguish from other features
  - Add a value from another randomly selected record
  - Add the mean, median or the mode value for the column
  - Use algorithms that support Missing values!

Features (10)

- Another factors to consider when working with features including Normalisation, Scaling, and Prunning

Types of Machine Learning

Supervised	Unsupervised	Reinforcement
• Task driven • Data with defined output are provided • Classification (image classification) and regression (forecasting)	• Data driven • Machine understands data (identifies patterns/structures) • Clustering (customer segmentation)	• Algorithms learns to react to an environment • Game AI and Robot navigation



Supervised learning

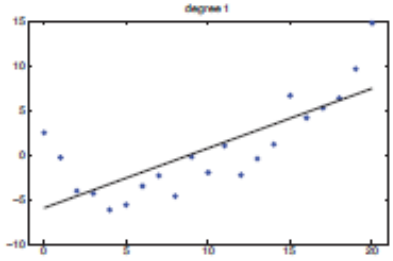
- A model creation process where the model describes the relationship between a set of features (attributes) and a predefined variable called target class (also called label)
- The task of the model is to estimate the value of the target variable as a function of the features
- In supervised learning, we have some target quantity that we would like to predict

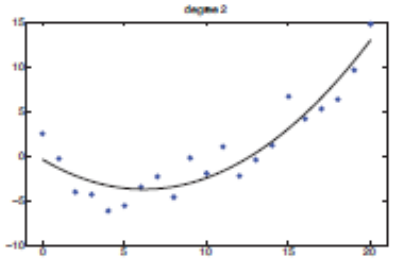
Classification and Regression

- Classification: the task of classifying some input features into classes (categories)
- As the name implies, the output of classification is categorical ('class' name)
- Image classification, Face detection and recognition, Handwritten recognition, Speech recognition, Document classification, …

Classification and Regression (2)

- Regression: the output variables of a regression problem is a continuous value (i.e., predict the price of stocks in the future)





Classification and Regression (3)

- Examples on regression:
  - Predict the stock market given current market conditions
  - Predict the prices of real-estate given some data about it
  - Predict the age of a viewer watching a video on YouTube
  - Predict the weather given some measures (temperature, humidity, wind, …)
  - Predict the temperature inside a building using weather data, time, door sensors, …

Model

- A Model is a simplified representation of reality created to serve a purpose
- Simplified based on assumptions of what is and what is not important for the specific purpose
- Sometimes based on constraints on information

Binary classification vs. Multi-class classification

- The outcome of a binary classification problem is two classes
- Usually the classes are encoded as {-1, 1}
- A multiclass classification problem has many target classes
- For example, image scene classification problem where the outcome of the classification is a class label of the scene

Prediction

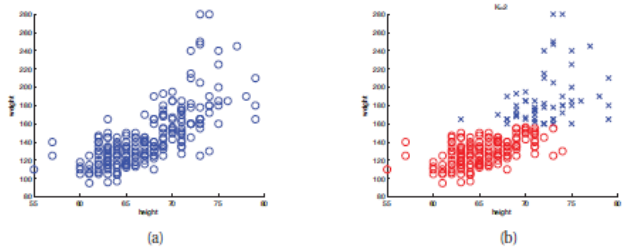
- Prediction means to forecast a future event
- In the context of Machine Learning and Data Science, it means to estimate an unknown value
- Predictive models are built using historical data
- For instance, a predictive model spam filtering estimates whether a given email is spam or not

Unsupervised learning

- In unsupervised learning, the training data consists of input vectors without any corresponding target values
- The goal might be, for instance, to discover similar examples within the data (clustering such as K-means algorithm)
- Discover patterns/structures hidden inside the data
- This is referred to as Knowledge Discovery

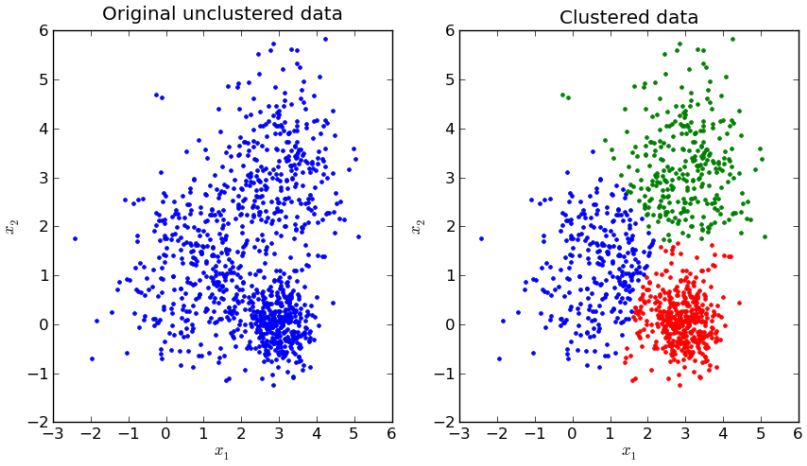
Unsupervised learning

- The tasks here are generally formulated as density estimations in order to build the models
- More widely applicable than supervised learning as there is not required human annotation of the data as in supervised learning

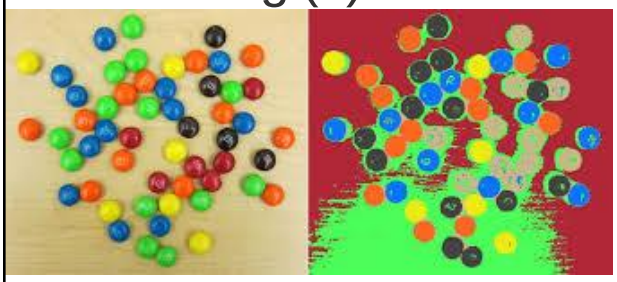


(a) The height and weight of some people. (b) A possible clustering using $K = 2$ clusters.

Clustering



Clustering (2)



Datasets

- The dataset is a matrix of variables (features, attributes) that represent the observations of the real world
- Each row contains a set of attributes (features, variables)
- Each row can be seen as an instance and is referred to as a feature vector

Training/test data

- The training data is the data that is used as input for the learning algorithm for inducing the model
- It is the data (experience) that is used by the model to build its hypothesis (whether classification or regression)
- The test dataset is the subset of the data that is used to measure the accuracy of the model