Course : COMP6577 – Machine Learning

Effective Period : February 2020

Introduction to Machine Learning 1

Session 01 & 02

People
Innovation
Excellence

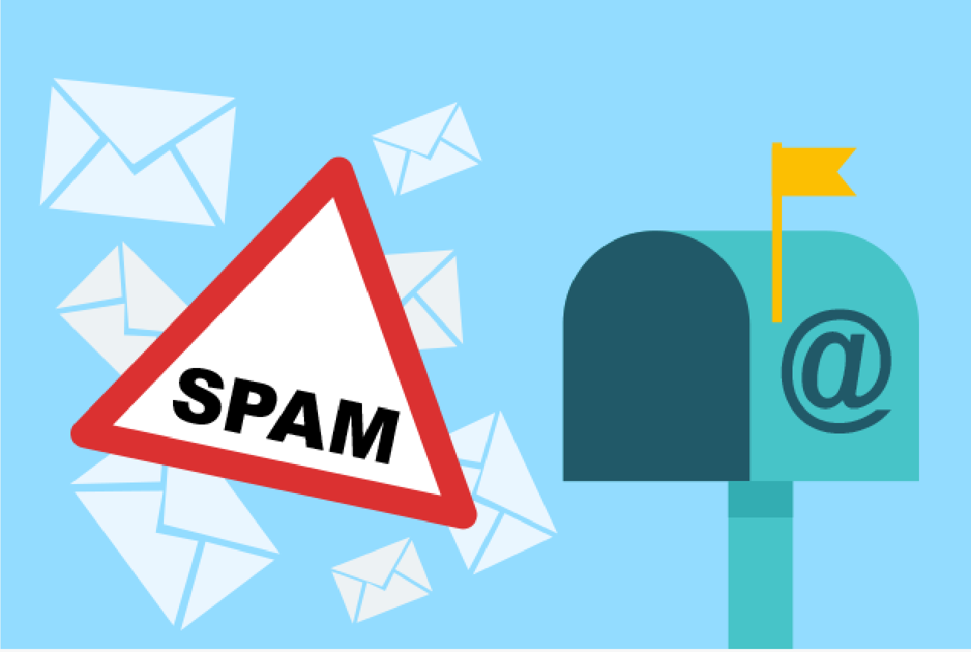
Learning Outcome

- LO1 : Student be able to explain the fundamental of machine learning concept

Outline

- What is machine learning?
- Why use machine learning?
- Machine learning area
- When use machine learning?
- Types of machine learning system
- Case Study

People
Innovation
Excellence



- The first Machine Learning application that really became mainstream, improving the lives of hundreds of millions of people, took over the world back in the 1990s: it was the spam filter. It has actually learned so well that you seldom need to flag an email as spam anymore.

What is Machine Learning?

- Machine Learning is the science (and art) of programming computers so they can learn from data.

- [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed. **(Arthur Samuel, 1959)**

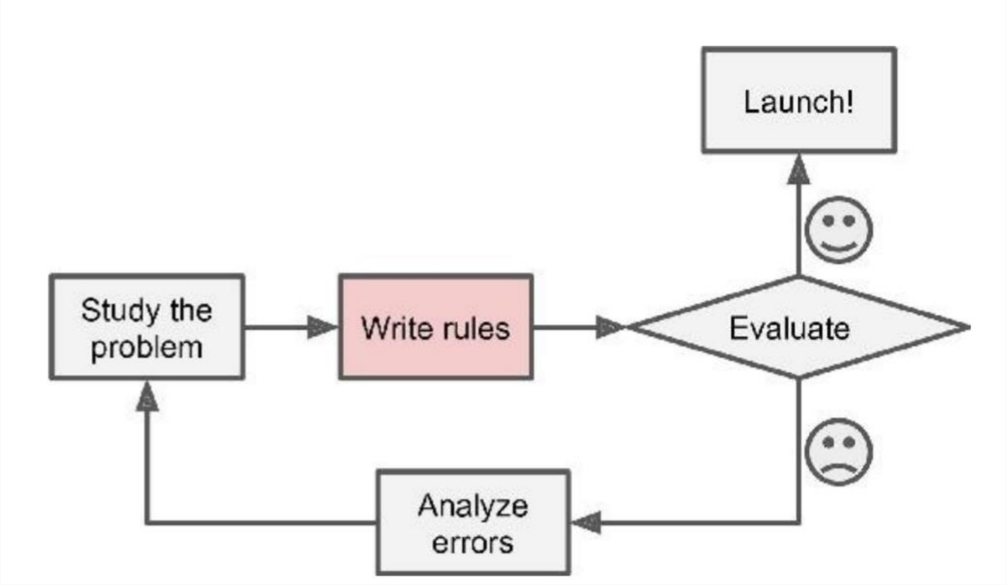
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. **(Tom Mitchell, 1997)**

Spam Filter

- Spam filter is a Machine Learning program that can learn to flag spam given examples of spam emails (e.g., flagged by users) and examples of regular (non-spam, also called “ham”) emails.
- The examples that the system uses to learn are called the **training set.** Each training example is called a **training instance** (or **sample**).
- In this case, the task **T** is to flag spam for new emails, the experience **E** is the training data, and the performance measure **P** needs to be defined; ; for example, you can use the ratio of correctly classified emails. This particular performance measure is called accuracy and it is often used in classification tasks.

Why use machine learning? (1)

- Consider how you would write a spam filter using traditional programming techniques.

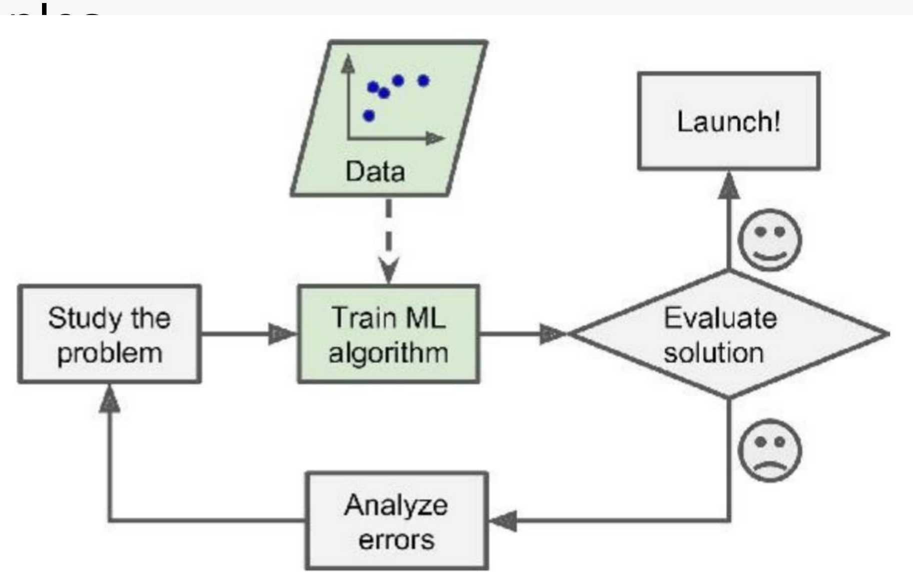


- Since the problem is not trivial, your program will likely become a long list of complex rules — pretty hard to maintain

People
Innovation
Excellence

Why use machine learning? (2)

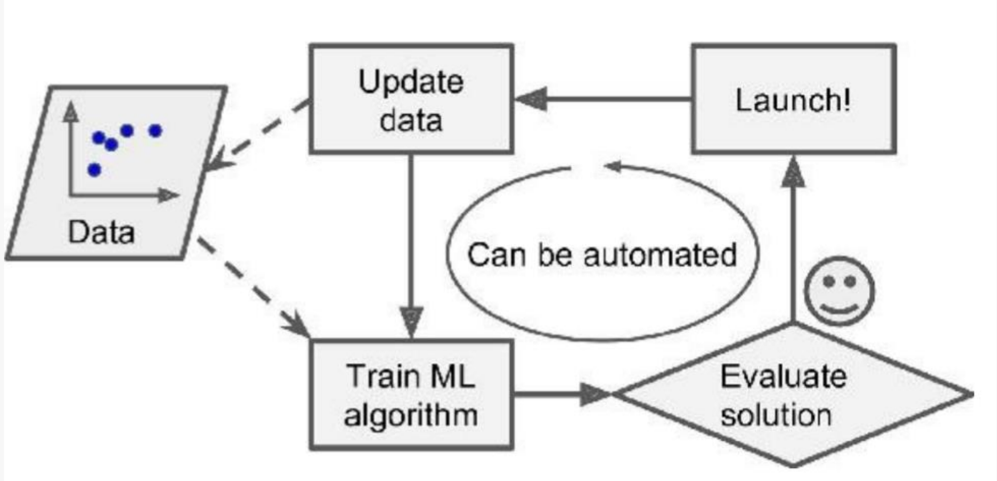
- In contrast, a spam filter based on Machine Learning techniques automatically learns which words and phrases are good predictors of spam by detecting unusually frequent patterns of words in the spam examples compared to the ham examples.



- The program is much shorter, easier to maintain, and most likely more accurate.

Why use machine learning? (3)

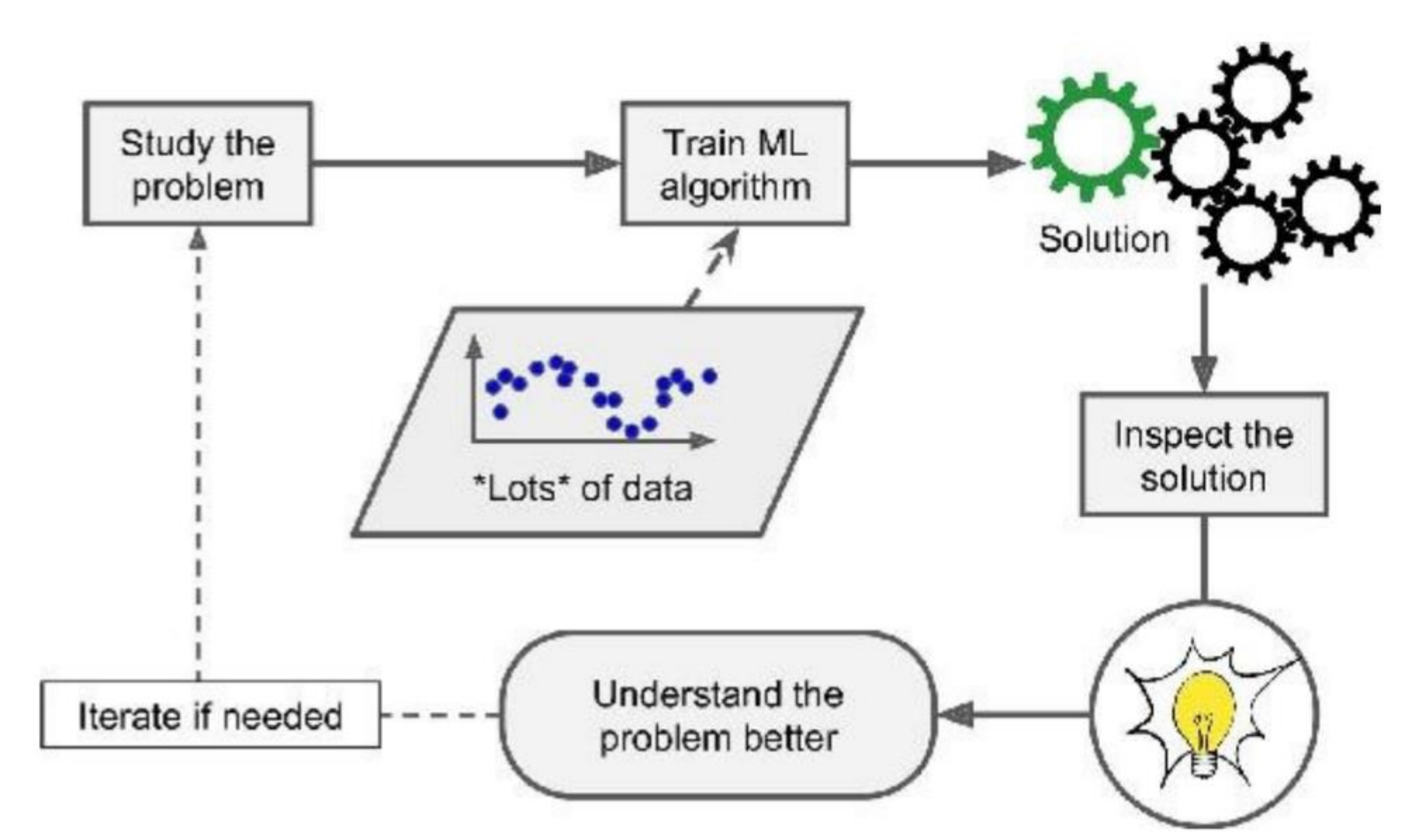
- If the case that spammers notice all their emails containing “4U” are blocked, they might start writing “For U” instead. A spam filter using traditional programming techniques would need to be updated to flag “For U” emails. (keep writing new rules forever)
- In contrast, a spam filter based on Machine Learning techniques automatically notices that “For U” has become unusually frequent in spam flagged by users, and it starts flagging them without your intervention



Machine Learning Area

- **Application cannot program by hand**
  - Problems that either are too complex for traditional approaches or have no known algorithm
  - E.g. Autonomous helicopter, handwriting recognition, speech recognition, most of Natural Language Processing (NLP), Computer Vision
- **Data Mining**
  - Applying ML techniques to dig into large amounts of data can help discover patterns that were not immediately apparent.
  - Large datasets from growth of automation/web
  - E.g. web click data, medical records, biology, engineering
- **Self-customizing program**
  - E.g. Amazon, Netflix product recommendations
- **Understanding human learning (brain, real AI)**

Machine Learning helps humans learn



When use machine learning?

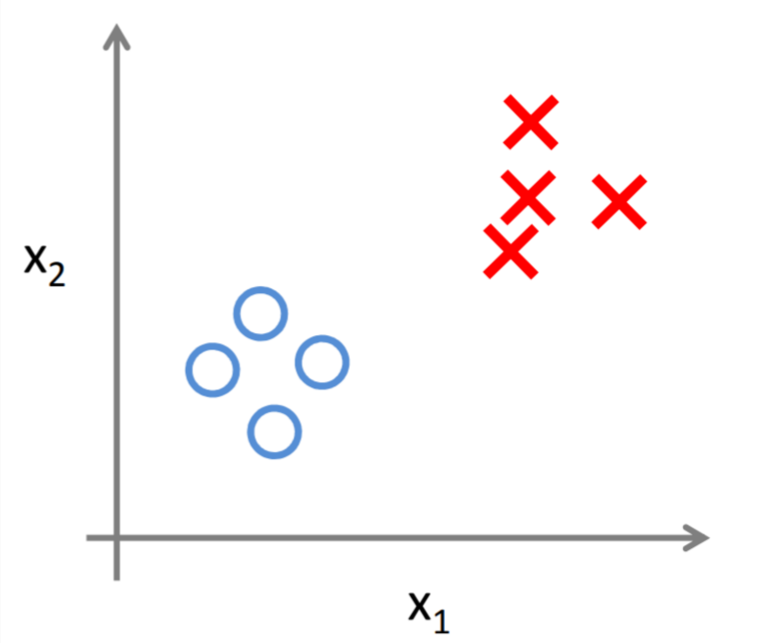
- Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

Types of Machine Learning Systems

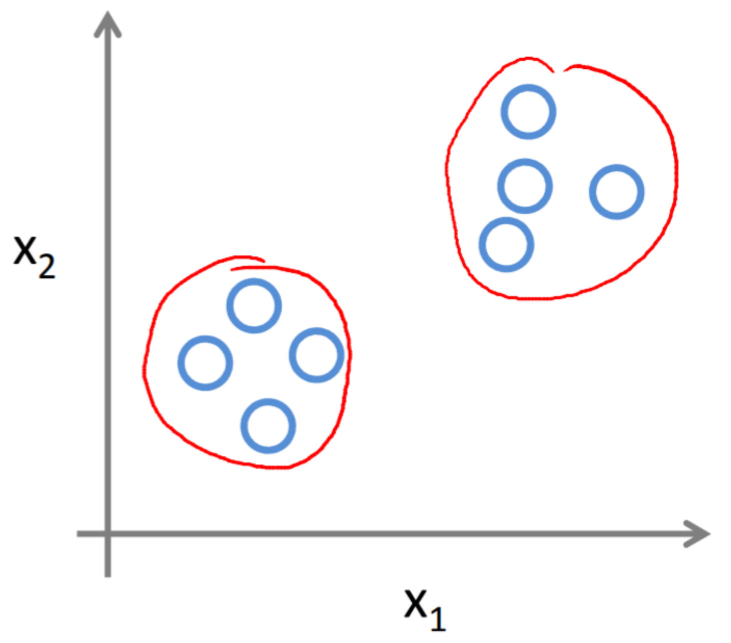
- There are so many different types of Machine Learning systems that it is useful to classify them in broad categories based on:
  1. Whether or not they are trained with human supervision (supervised, unsupervised, semisupervised, and Reinforcement Learning)
  2. Whether or not they can learn incrementally on the fly (online versus batch learning)
  3. Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)

These criteria are not exclusive; you can combine them in any way you like

Supervised Vs Unsupervised Learning



Supervised Learning



Unsupervised Learning

People
Innovation
Excellence

Supervised Learning

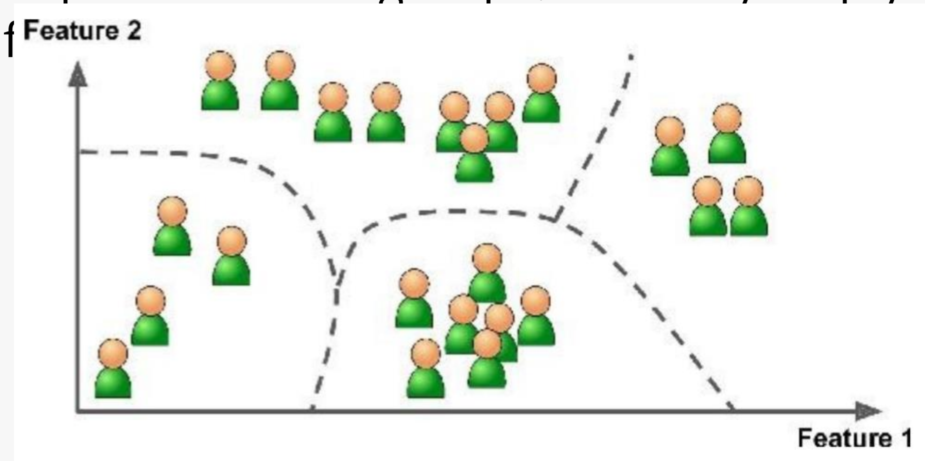
- In supervised learning, the training data you feed to the algorithm includes the desired solutions, called labels.
- A typical supervised learning task is classification. The spam filter is a good example of this.
- Another typical task is to predict a target numeric value, such as the price of a car, given a set of features (mileage, age, brand, etc.) called predictors. This sort of task is called regression .
- Note that some regression algorithms can be used for classification as well, and vice versa.
- Here are some of supervised learning algorithms:
  - k-Nearest Neighbors, Linear Regression, Logistic Regression, Support Vector Machines (SVMs), Decision Trees,and Random Forests, Neural networks

Unsupervised Learning

- In unsupervised learning, the training data is unlabeled. The system tries to learn without a teacher.
- Here are some of unsupervised learning algorithms:
  1. Clustering
     - k-Means
     - Hierarchical Cluster Analysis (HCA)
     - Expectation Maximization
  2. Visualization and dimensionality reduction
     - Principal Component Analysis (PCA)
     - Kernel PCA
     - Locally-Linear Embedding (LLE)
     - t-distributed Stochastic Neighbor Embedding (t-SNE)
  3. Association rule learning
     - Apriori
     - Eclat

1. Clustering

- Let say you have a lot of data about your blog's visitors. You may want to run a clustering algorithm to try to detect groups of similar visitors. It might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on.
- If you use a hierarchical clustering algorithm, it may also subdivide each group into smaller groups. This may help you target your posts f



2. Visualization and dimensionality reduction

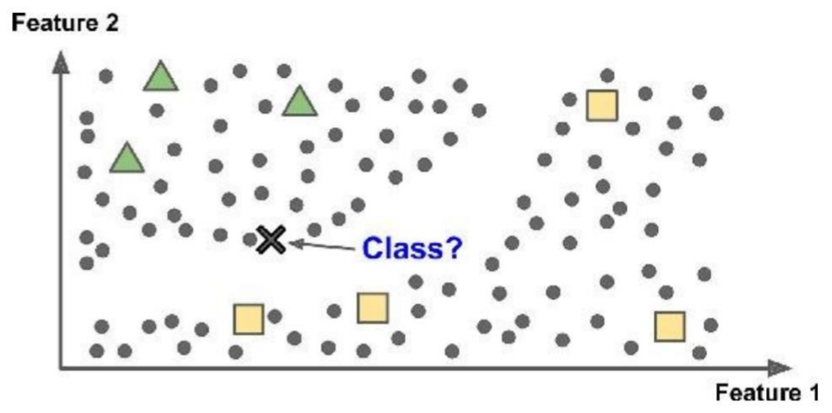
- Visualization algorithms output a 2D or 3D representation of your data that can easily be plotted. These algorithms try to preserve as much structure as they can (e.g., trying to keep separate clusters in the input space from overlapping in the visualization), so you can understand how the data is organized and perhaps identify unsuspected patterns.
- Dimensionality reduction, in which the goal is to simplify the data without losing too much information. One way to do this is to merge several correlated features into one. For example, a car's mileage may be very correlated with its age, so the dimensionality reduction algorithm will merge them into one feature that represents the car's wear and tear. This is called feature extraction.

3. Association rule learning

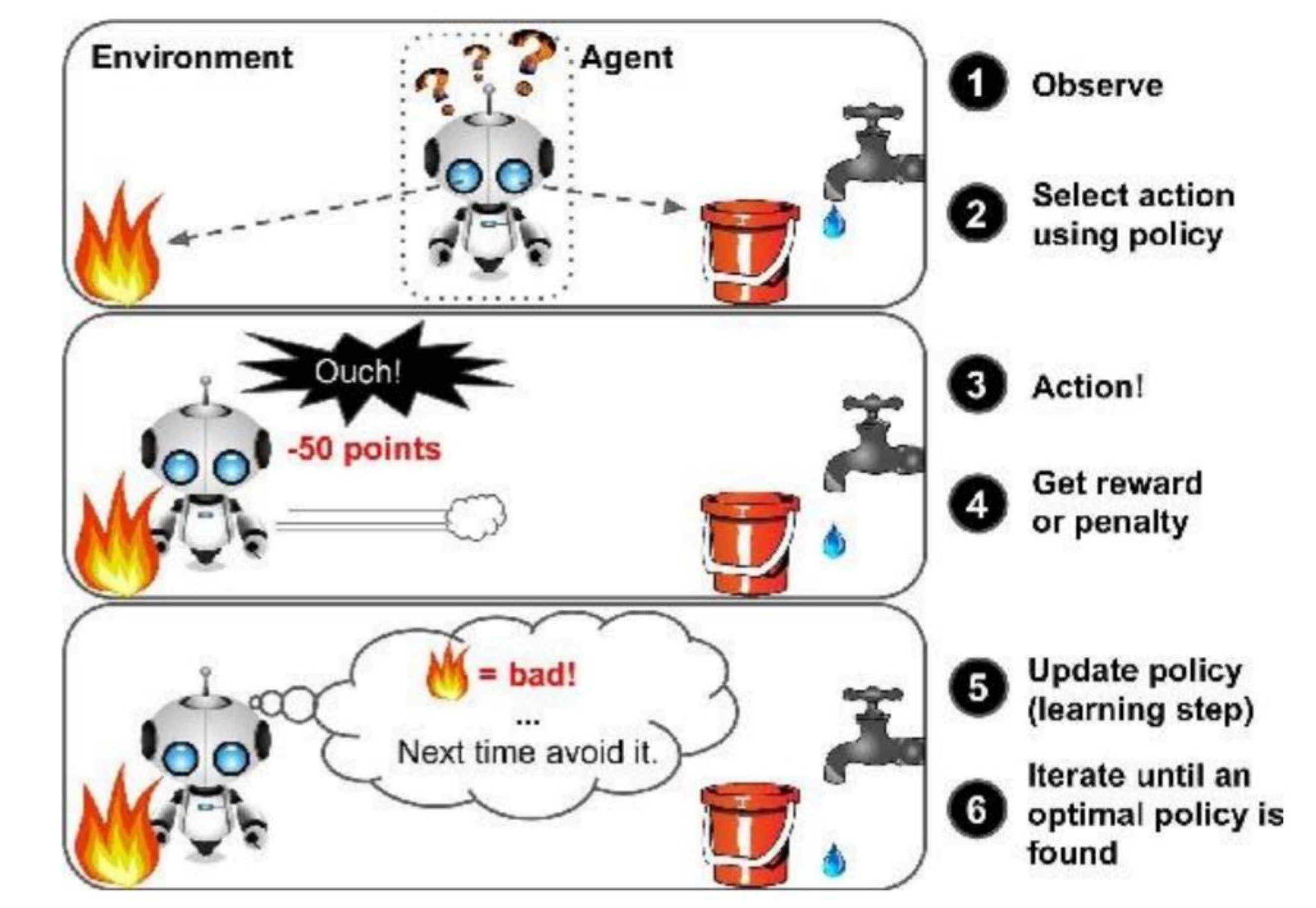
- The goal is to dig into large amounts of data and discover interesting relations between attributes. For example, suppose you own a supermarket. Running an association rule on your sales logs may reveal that people who purchase barbecue sauce and potato chips also tend to buy steak. Thus, you may want to place these items close to each other.

Semisupervised Learning

- Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data.
- Example: Some photo-hosting services, such as Google Photos. Once you upload all your family photos to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11, while another person B shows up in photos 2, 5, and 7. This is the unsupervised part of the algorithm (clustering). Now all the system needs is for you to tell it who these people are. Just one label per person, 4 and it is able to name everyone in every ph photos.



Reinforcement Learning

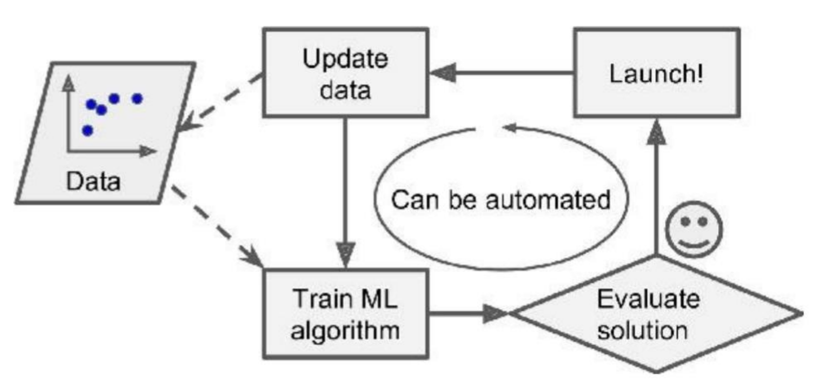

Environment
Agent
Ouch!
-50 points
= bad!
...
Next time avoid it.
1 Observe
2 Select action using policy
3 Action!
4 Get reward or penalty
5 Update policy (learning step)
6 Iterate until an optimal policy is found

Reinforcement Learning Example

- DeepMind’s AlphaGo program is also a good example of Reinforcement Learning: it made the headlines in March 2016 when it beat the world champion Lee Sedol at the game of Go. It learned its winning policy by analyzing millions of games, and then playing many games against itself. Note that learning was turned off during the games against the champion; AlphaGo was just applying the policy it had learned.

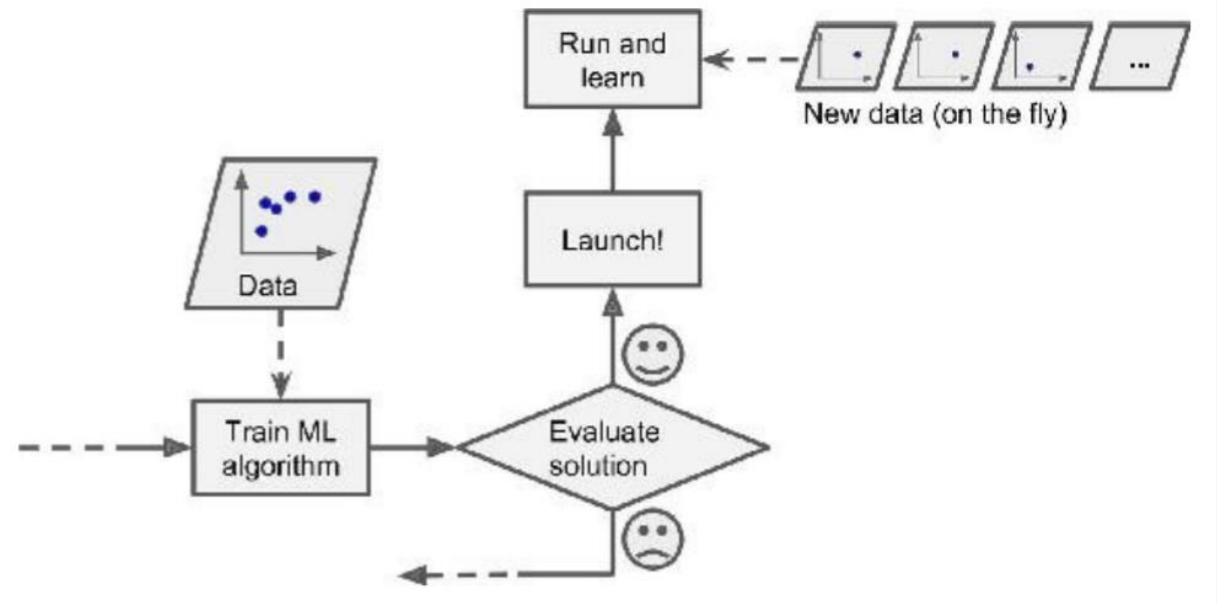
Batch Learning

- In batch learning, the system is incapable of learning incrementally: it must be trained using all the available data. This will generally take a lot of time and computing resources.
- **Offline learning**: First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned.
- If you want a batch learning system to know about new data (such as a new type of spam), you need to train a new version of the system from scratch on the full dataset (not just the new data, but also the old data), then stop the old system and replace it with the ne



Online Learning

- In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches. Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.

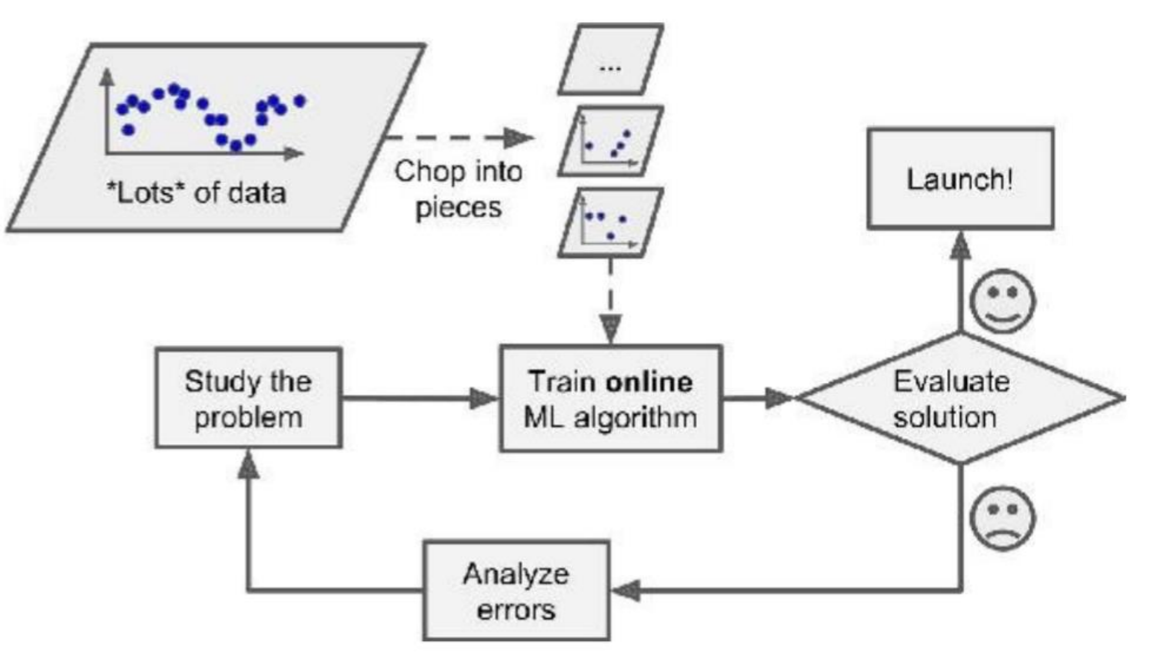


People
Innovation
Excellence

People
Innovation
Excellence

Online Learning (2)

- Advantages:
  - Train systems on huge datasets that cannot fit in one machine's main memory (this is called out-of-core learning)

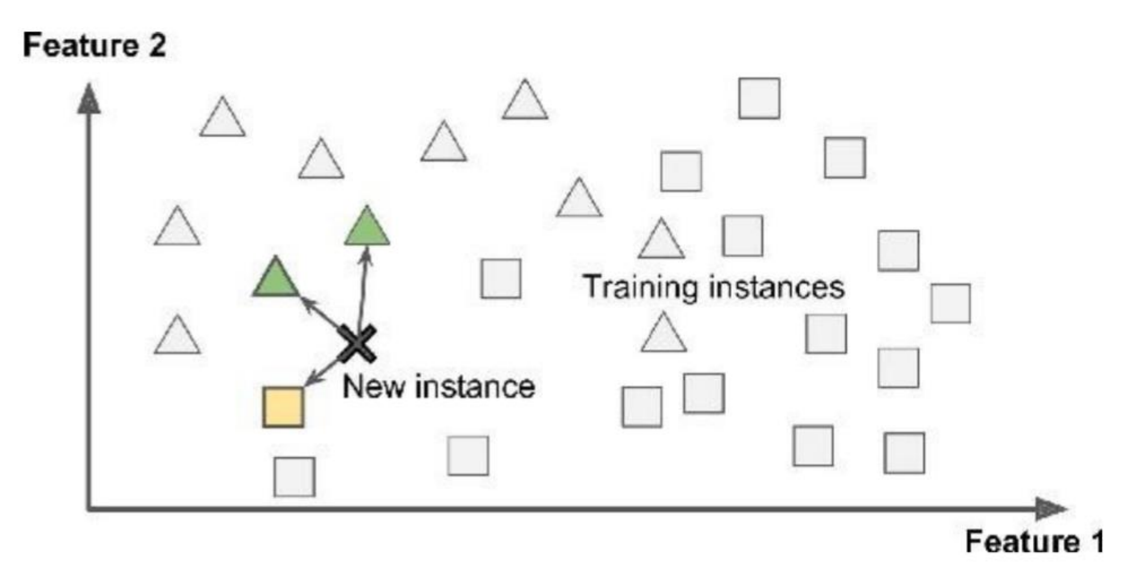


Online Learning (3)

- One important parameter of online learning systems is learning rate.
  - High learning rate: system will rapidly adapt to new data, but it will also tend to quickly forget the old data.
  - Low learning rate: system will learn more slowly, but it will also be less sensitive to noise in the new data or to sequences of nonrepresentative data points.
- A big challenge with online learning:
  - If bad data is fed to the system, the system’s performance will gradually decline.
  - Tips to reduce the risk:
    - Monitor your system closely and promptly switch learning off (and possibly revert to a previously working state) if you detect a drop in performance.
    - monitor the input data and react to abnormal data (e.g., using an anomaly detection algorithm).

Instance-Based Learning

- The system learns the examples by heart, then generalizes to new cases using a similarity measure.



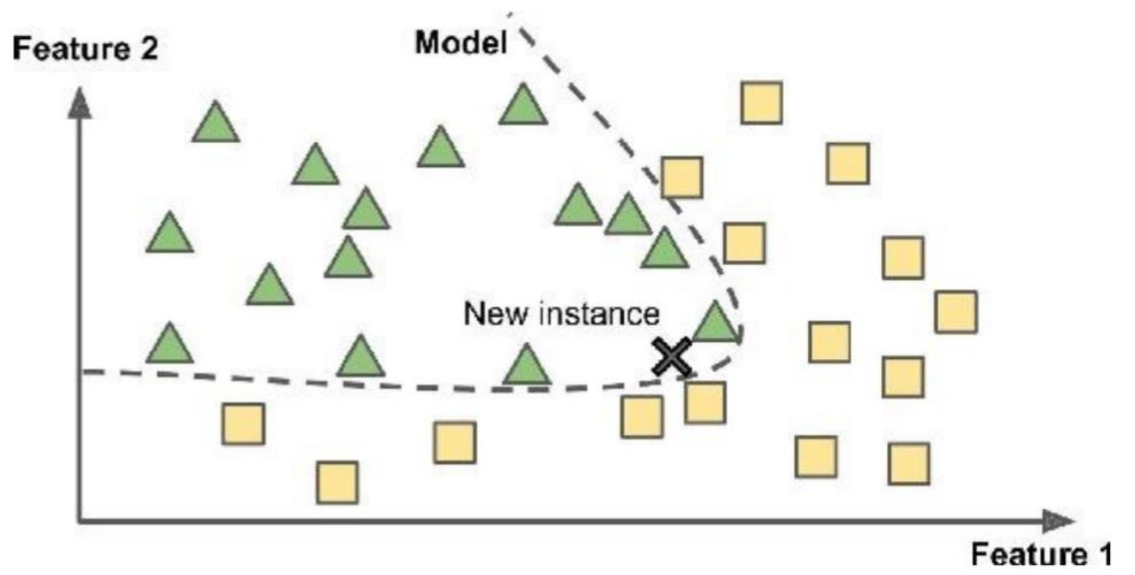
Model-based Learning

- Another way to generalize from a set of examples is to build a model of these examples, then use that model to make predictions.

1. You studied the data.
2. You selected a model.
3. You trained it on the training data (i.e., the learning algorithm searched for the model parameter values that minimize a cost function).
4. Finally, you applied the model to make predictions on new cases (this is called inference), hoping that this model will generalize well.

People
Innovation
Excellence

Model-based Learning (2)



- If all went well, your model will make good predictions. If not, you may need to use more attributes, get more or better quality training data, or perhaps select a more powerful model.

Review Question

1. How would you define Machine Learning?
2. Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?
3. What are the two most common supervised tasks?
4. Can you name four common unsupervised tasks?

Case Study

Given data of Singapore Airbnb which can be downloaded in this link

https://www.kaggle.com/jojoker/singapore-airbnb

1. Analyse the data and give a suggestion of type of machine learning system that can be applied.

2. Try in Google Collaboratory to explore the data, for instance:
   - Find minimum or maximum of a value
   - Eliminate null value
   - Replace null value to certain value

End of Session 01 & 02

References

- Aurélien Géron. (2017). 01. *Hands-on Machine Learning with Scikit-Learn and Tensorflow.* O'Reilly Media, Inc..LSI: 978-1-491-96229-9. Chapter 1.
- Sergios Theodoridis. (2015). *Machine Learning: a Bayesian and Optimization Perspective.* Jonathan Simpson. ISBN: 978-0-12-801522-3. Chapter 1.
- Coursera Machine Learning by Prof. Andrew Ng. Week 1.
- https://www.kaggle.com/jojoker/singapore-airbnb

People
Innovation
Excellence