Basics: Machine Learning

Lecture 6

Dr. Nouman M Durrani

Contents of this Week

- Introduction to Machine Learning
  - ML Examples
- What is a Model?
- Types of Machine Learning
  (a) Supervised Learning
    - Examples
    - Workflow
    - Classification
    - Regression
    - Examples

  (b) Unsupervised Learning
    - Examples
    - Workflow
    - Clustering
    - Association
    - Examples
  (c) Reinforcement Learning
    - Examples

- The clustering Problem
- **k-means Clustering**
- k-means Clustering Solved Example
- Examples
- k-means++
- **kNN Classification**
- kNN Classification Solved Example

Week 9

- Hierarchical clustering
- Naive Bayes Classifier

Machine Learning

- The term Machine Learning was coined by Arthur Samuel in 1959:

"Machine Learning algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed".

- In 1997, Tom Mitchell gave a mathematical and relational definition that:

"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".

ML Examples

Family-friendly hotels in Istanbul — See all

Hotel Amira Istanbul — 4,017 Reviews — Istanbul, Turkey — "They were personable, funny, very helpful, and provided the **total package** in terms of the accommodation."

Muyan Suites — 1,149 Reviews — Istanbul, Turkey — "... card to travel to takism... Overall just perfect and awesome place thatz Muyan suites for a best **vacation** in istanbul.. Wish to go back again n stay only with thix small lil family of ours now... .. Thank ..."

Hotel Yasmak Sultan — 1,842 Reviews — Istanbul, Turkey — "This hotel really is the **whole package** -- comfortable, clean, superbly located, has a lovely rooftop restaurant, and even has its own hamam."

White House Hotel Istanbul — 4,593 Reviews — Istanbul, Turkey — "Amazing and very clean Hotel !!! Have a great sleep and nicely **holiday**!!!!! Amazing stuff very helpful when you need something !!! Amazing manager , very professional !!! Thank you very much to all team "

Luxury hotels in Istanbul — See all

Example 1:

- Suppose you decide to check out trip offers for a vacation

- You browse through the travel agency website and search for a hotel

- When you look at a specific hotel, just below the hotel description there is a section titled "You might also like these hotels".

- This is a common use case of Machine Learning called "Recommendation Engine"

  - Many data points were used to train a model in order to predict what will be the best hotels to show you under that section, based on a lot of information they already know about you

Example: Netflix
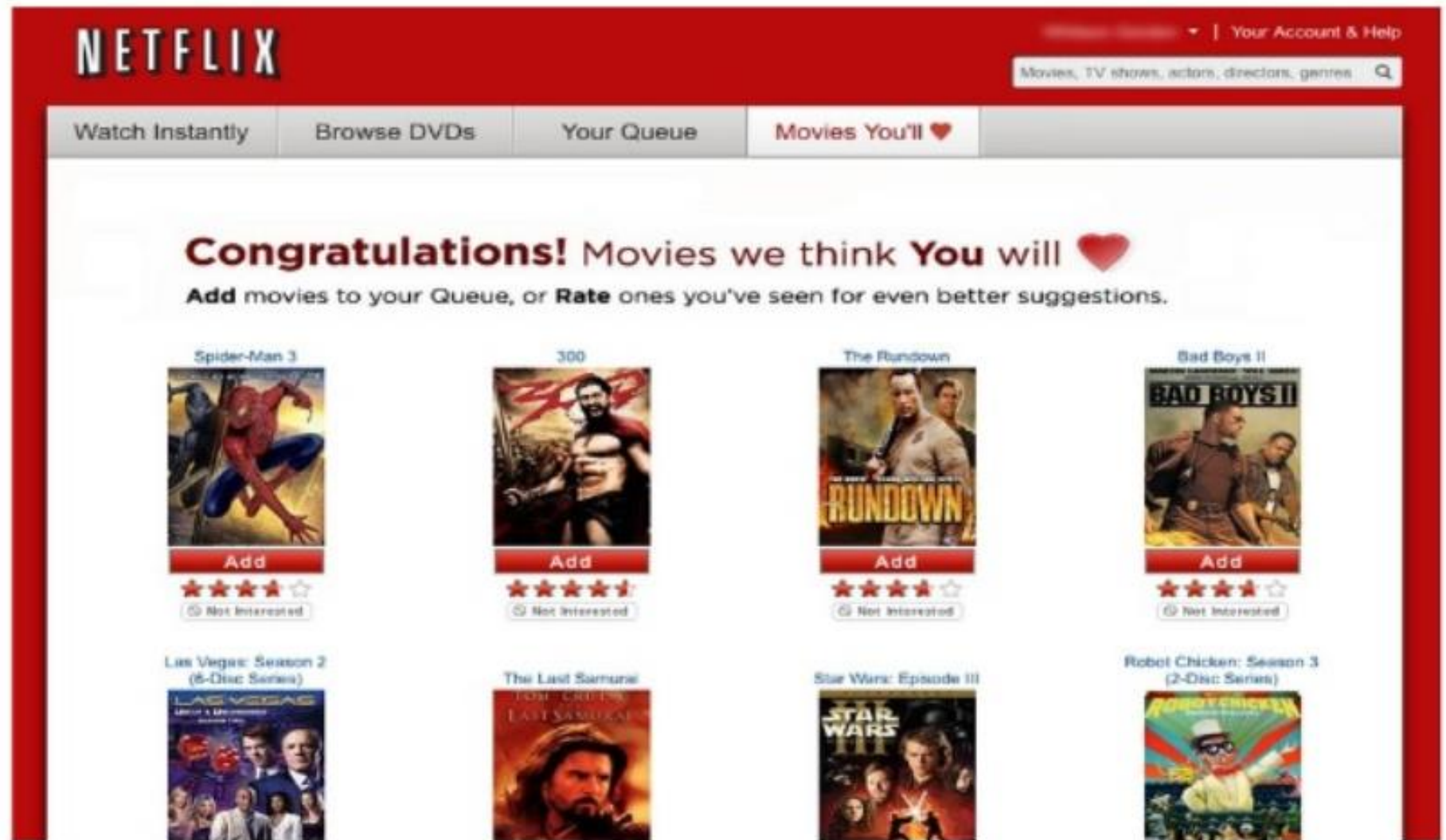
ML Examples

Example 2:

- Program to predict traffic patterns at a busy intersection (task T)

- Run it through a machine learning algorithm with data (task T) about past traffic patterns (experience E) and, if it has successfully "learned", it will then do better in predicting future traffic patterns (performance measure P).

ML Examples

Example 3: Learn to detect SPAM

- T: Distinguish between SPAM and Non-SPAM
- P: % of emails correctly classified
- E: Labeled emails from your friend Abdullah

Examples of machine learning problems

- Medical diagnoses: ML is trained to recognize cancerous tissues
  - "Is this cancer?"

- Graph Processing:
  - "Which of these people are good friends with each other?"

- Recommender Systems:
  - "Will person X likes movie Y?"

- Financial industry and trading —fraud investigations and loan sanction

- Speech Recognition:
  - "Is this his/her voice?" (voice searches, voice dialing, call routing, and appliance control)

Such problems are excellent targets for Machine Learning, and in fact machine learning has been applied to such problems with great success.

What is a Model

- A **model** is a mathematical formula with a number of parameters that need to be learned from the data
  - Fitting a model to the data is a process known as model training

- Example: Consider a one feature/variable linear regression, where the goal is to fit a line (described by the equation $y = ax + b$) to a set of distributed data points.

- Once the model training is completed we get a model equation $y = 2x + 5$.
  - Then for a set of inputs [1, 0, 7, 2, …] we would get a set of outputs [7, 5, 19, 9, …].

Types of Machine Learning

Machine learning can be classified into 3 types of algorithms.
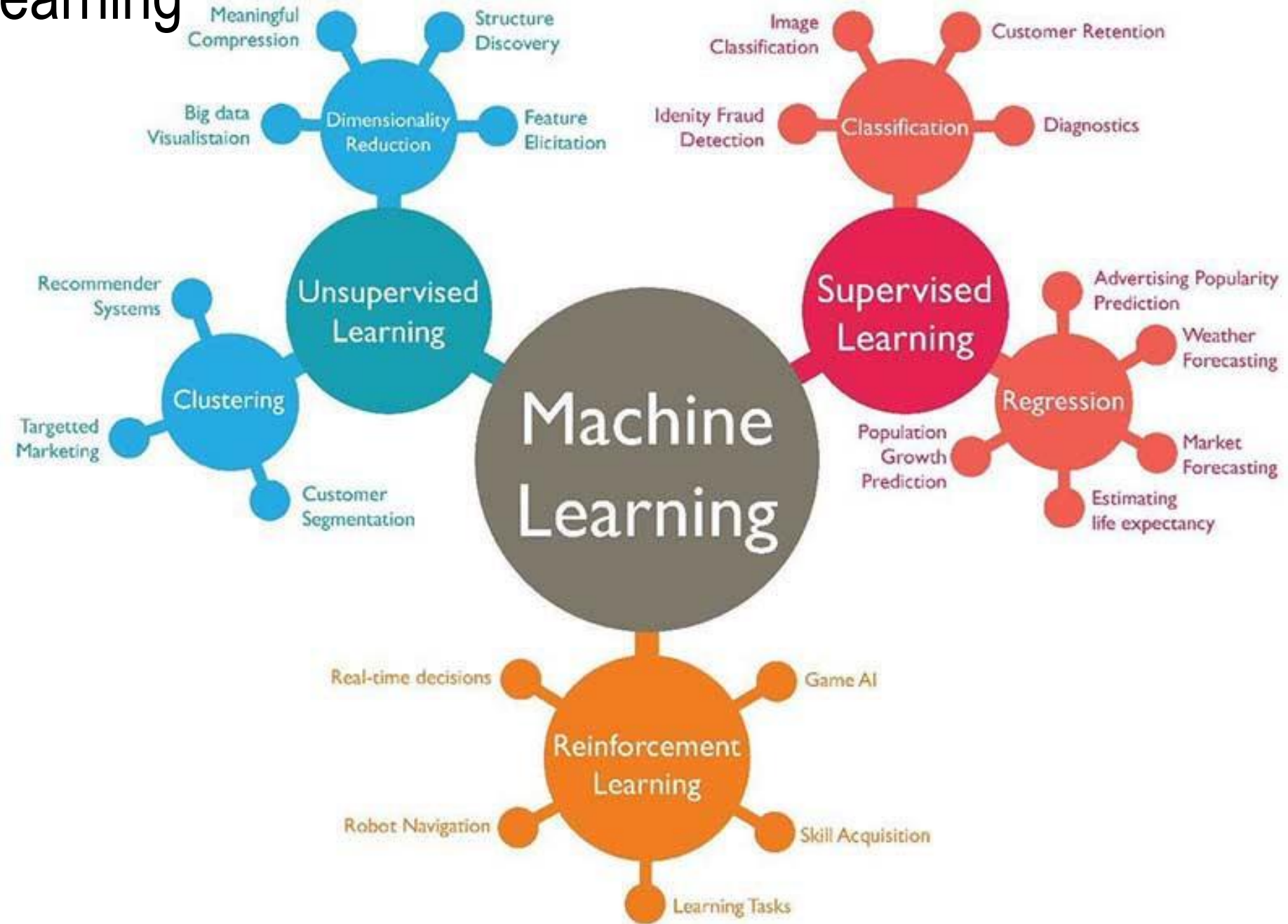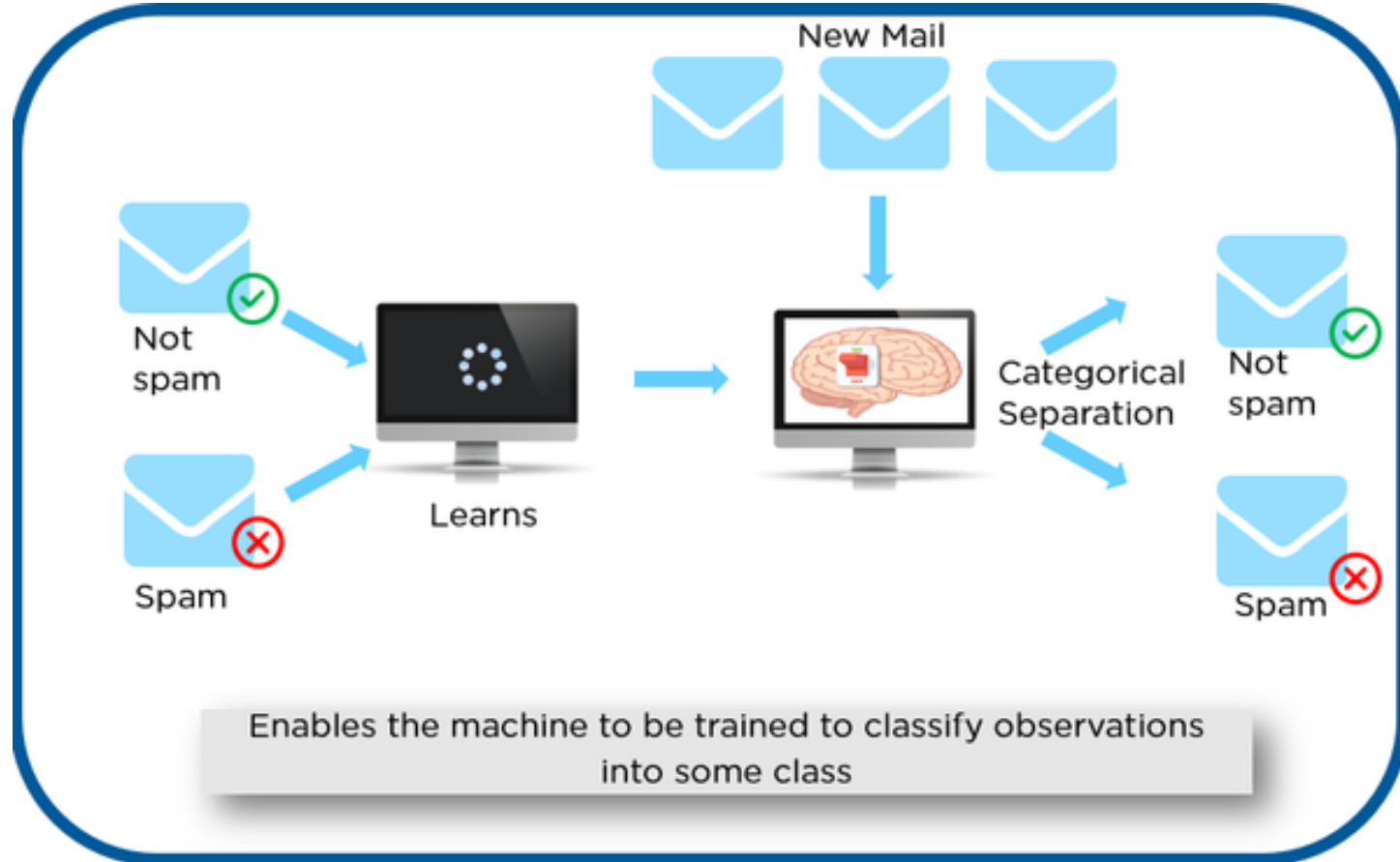
Supervised Learning

- Supervised learning is a learning model built to make prediction, given an unforeseen input instance.

- A supervised learning algorithm takes a known set of input dataset and its known responses to the data (output) to learn the model.

- A learning algorithm then trains a model to generate a prediction for the response to new data or the test dataset.

- Supervised learning uses classification algorithms and regression techniques to develop predictive models.

- The algorithms include linear regression, logistic regression, neural networks, decision tree, Support Vector Machine (SVM), random forest, naive Bayes, and k-nearest neighbor.
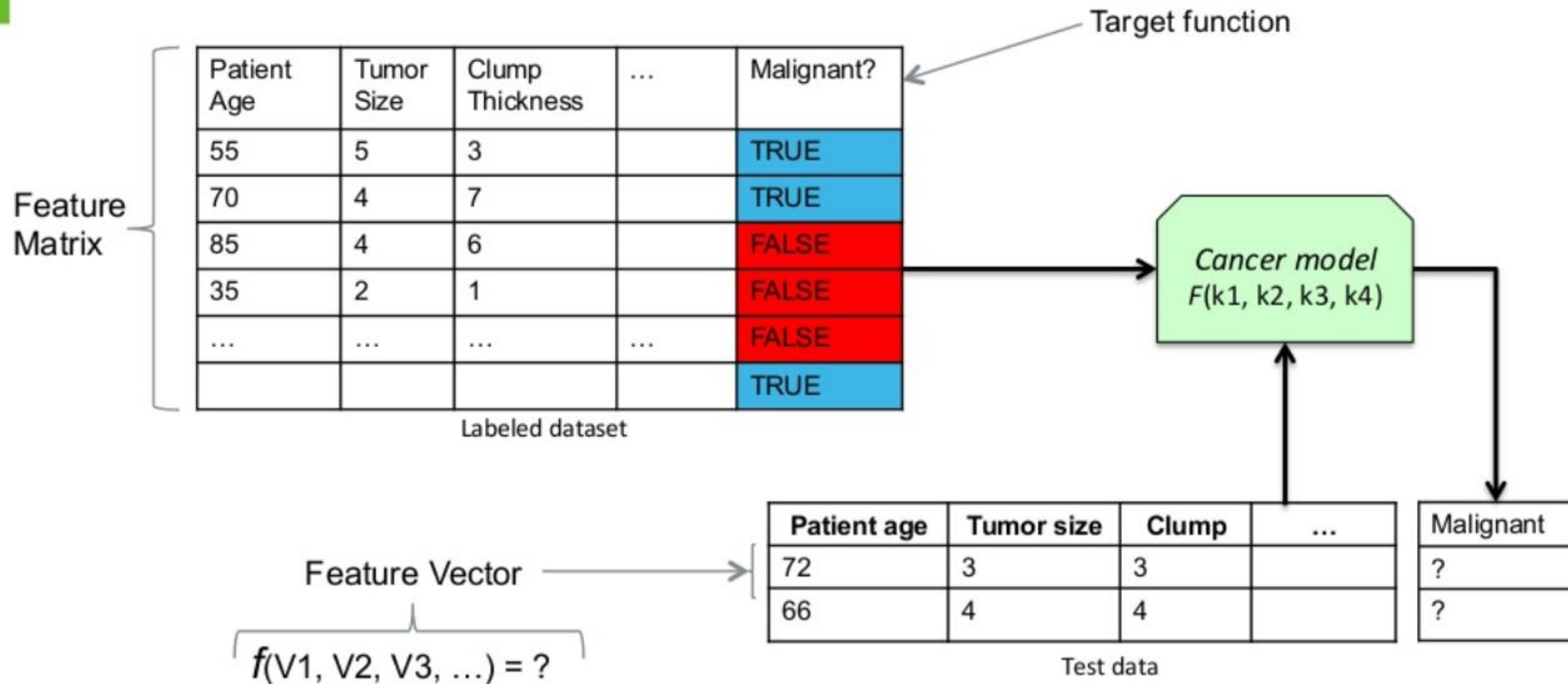
Supervised Learning

For example:

Spam filtering where Large number of email messages are labelled as either:

- spam

- non-spam

- New email message will then be classified as spam or non-spam

New Mail

Not spam

Spam

Learns

Categorical Separation

Not spam

Spam

Enables the machine to be trained to classify observations into some class

Supervised Learning: learn from examples

Target function

Patient Age	Tumor Size	Clump Thickness	...	Malignant?
55	5	3		TRUE
70	4	7		TRUE
85	4	6		FALSE
35	2	1		FALSE
...	...	...	...	FALSE
				TRUE

Feature Matrix

Labeled dataset

Cancer model
$F(k1, k2, k3, k4)$

Feature Vector

$f(V1, V2, V3, …) = ?$

Patient age	Tumor size	Clump	...
72	3	3	
66	4	4	

Test data

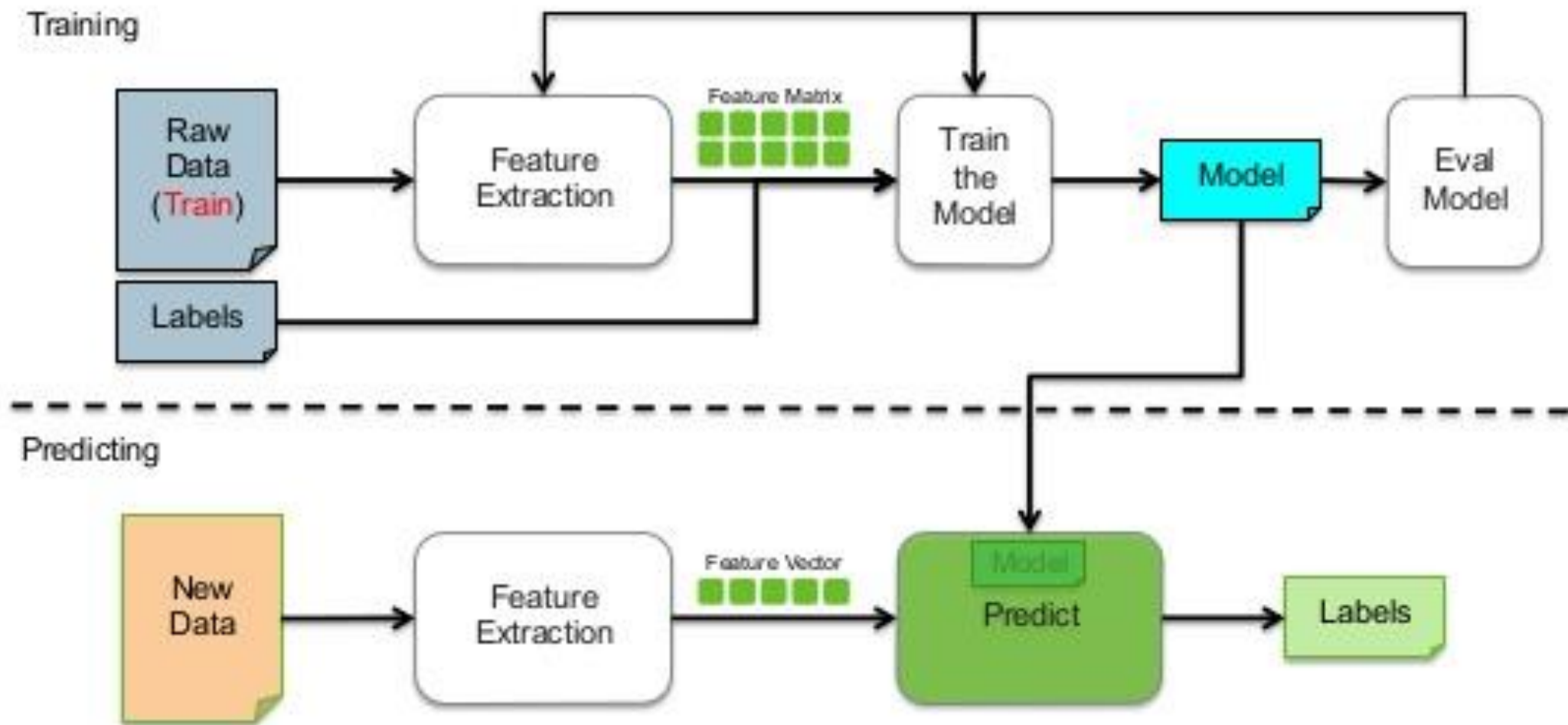| Malignant |
|---|
| ? |
| ? |

Supervised Learning

Training Images

Test Image

Supervised Learning Workflow

Classification

Classification learns from existing categorizations and then assigns unclassified items to the best category.

- Classification models classify input data into categories and predicts discrete responses

- Classification is recommended if the data can be categorized, tagged, or separated into specific groups or classes

- **Classification Examples:**
  - Bank credit scoring
  - Medical imaging
  - Speech recognition
  - To recognize letters and numbers in Handwriting
  - To check whether an email is genuine or spam
  - To detect whether a tumor is benign or cancerous

- Classification Algorithms:
  - k-nn, Decision Trees, Random Forest, SVM, Neural Network…

Classification Algorithms

- Classification algorithms attempt to estimate the mapping function (f) from the input variables (x) to discrete or categorical output variables (y).
  - In this case, y is a category that the mapping function predicts.

- **For example**, when provided with a dataset about houses, a classification algorithm predict whether the prices for the houses "sell more or less than the recommended retail price"

- **For example**, in a banking application, the customer who applies for a loan may be classified as a safe and risky according to his/her age and salary. The constructed model can be used to classify new data

Classification: predicting a category
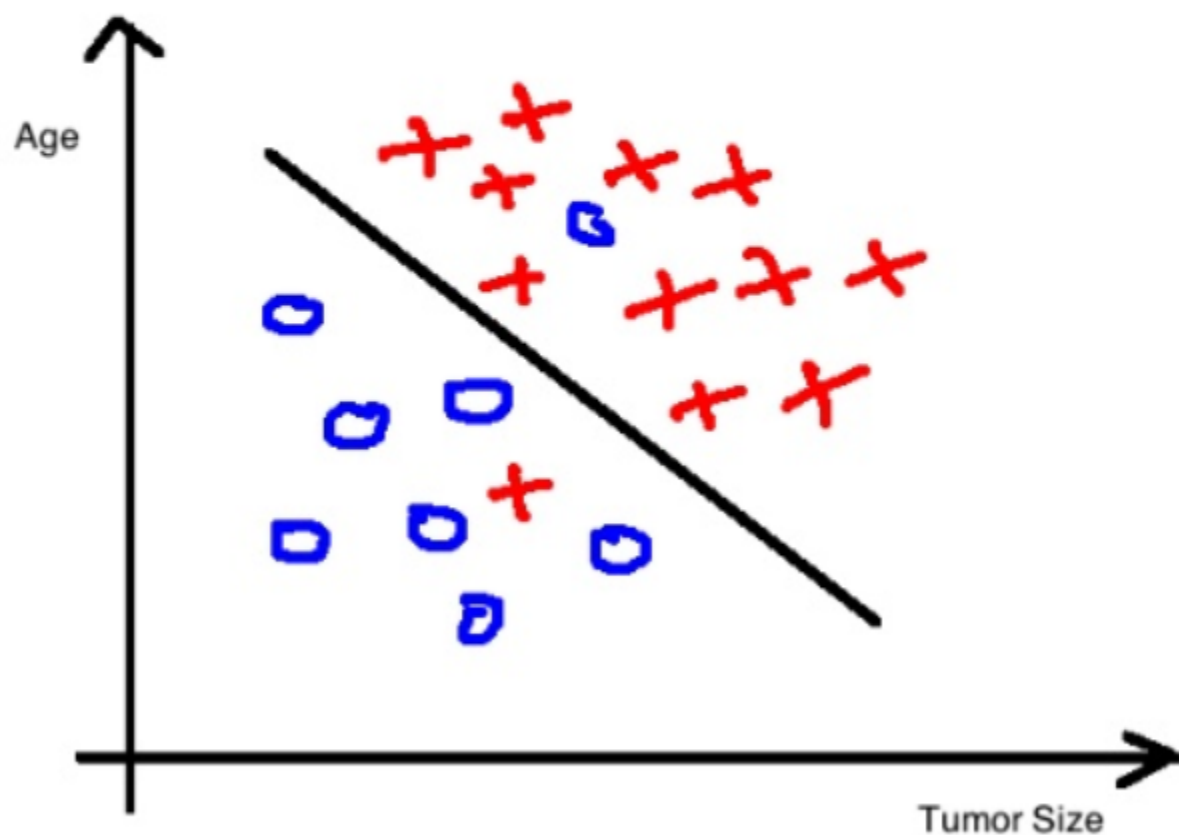
Some techniques:
- Naïve Bayes
- Decision Tree
- Logistic Regression
- SGD
- Support Vector Machines
- Neural Network
- Ensembles

Basics: Regression Algorithms

Regression techniques predict continuous responses

Regression techniques predict a continuous-valued attribute associated with an object

- Regression algorithms attempt to estimate the mapping function (f) from the input variables (x) to numerical or continuous output variables (y).
  - In this case, y is a real value, which can be an integer or a floating point value.
  - Therefore, regression prediction problems are usually quantities or sizes.

- For example, when provided with a dataset about houses, and you are asked to predict their prices, that is a regression task because price will be a continuous output.

- Regression algorithms include linear regression, Ensembles, Support Vector Regression (SVR), and regression trees.

Regression Examples:

- A linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data

- For example, a data is collected about how happy people are after getting so many hours of sleep.
  - In this dataset, sleep and happy people are the variables.

- Other Examples:
  - Drug Response, Stock Price, …

Regression: predict a continuous value
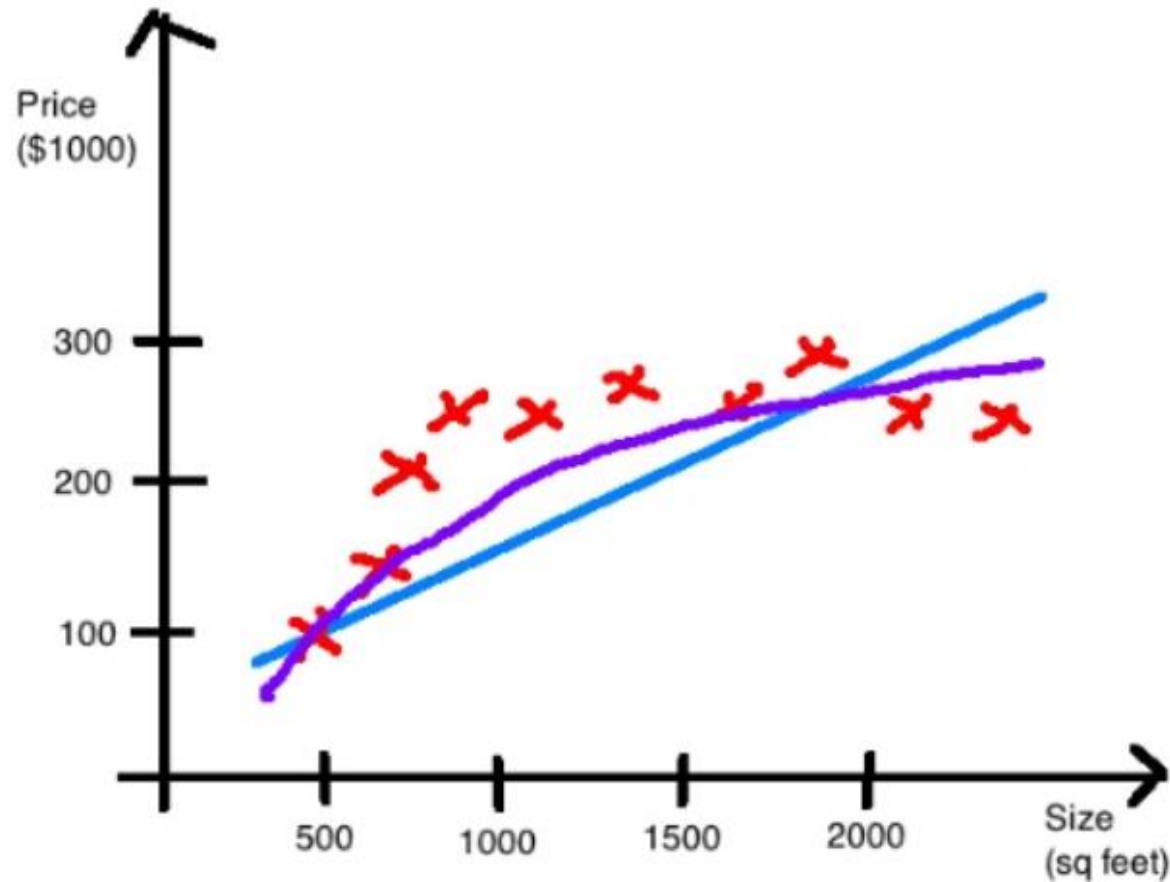
Some techniques:

- Linear Regression / GLM
- Decision Trees
- Support vector regression
- SGD
- Ensembles

Unsupervised Learning

In unsupervised learning the training data comprises examples of input vectors WITHOUT any corresponding target variables.

- In unsupervised learning, the algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own.
  - In an unsupervised learning you only have input data (X) and no corresponding output variables.
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Unsupervised Learning

Examples: Speech recognition, document clustering, and image compression.

- In document clustering, the aim is to group documents into various reports of politics, entertainment, sports, culture, heritage, art, and so on.

- Fraud Detection: Identify groups of motor insurance policy holders with a high average claim cost

- Social Networks: Recognize communities within large groups of people

Clustering: detect similar instance groupings
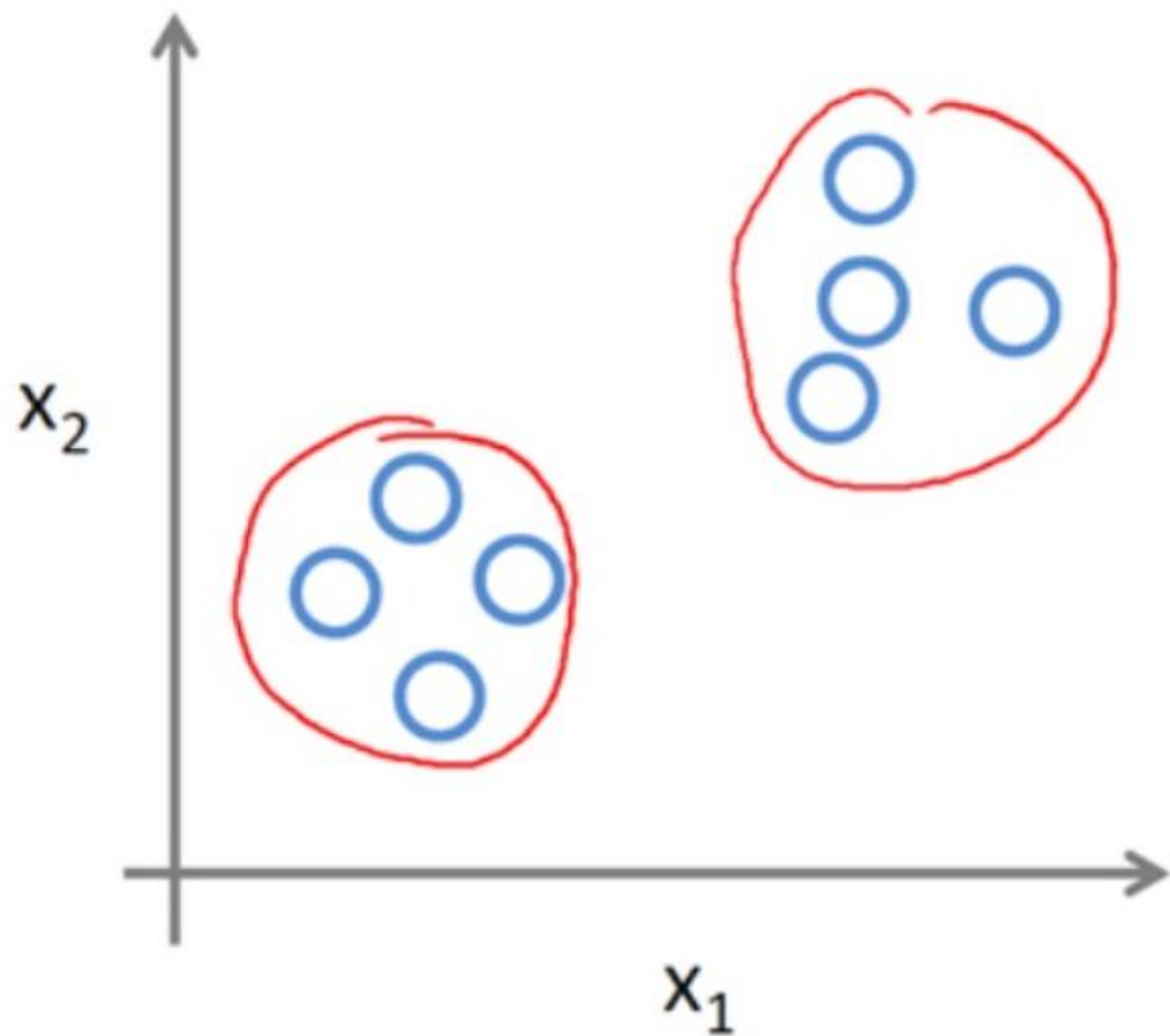
Some techniques:
- k-means
- Spectral clustering
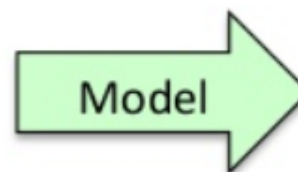- DB-scan
- Hierarchical clustering

Unsupervised Learning: detect natural patterns

Age	State	Annual Income	Marital status
25	CA	$80,000	M
45	NY	$150,000	D
55	WA	$100,500	M
18	TX	$85,000	S
…	…	…	…

Model → Naturally occurring (hidden) structure

No labels

Hortonworks

Unsupervised Learning

Unsupervised Learning Model

Training Text, Documents, Images, etc.

Feature Vectors

Machine Learning Algorithm

New Text, Document, Image, etc.

Feature Vector

Predictive Model

Likelihood or Cluster ID or Better Representation
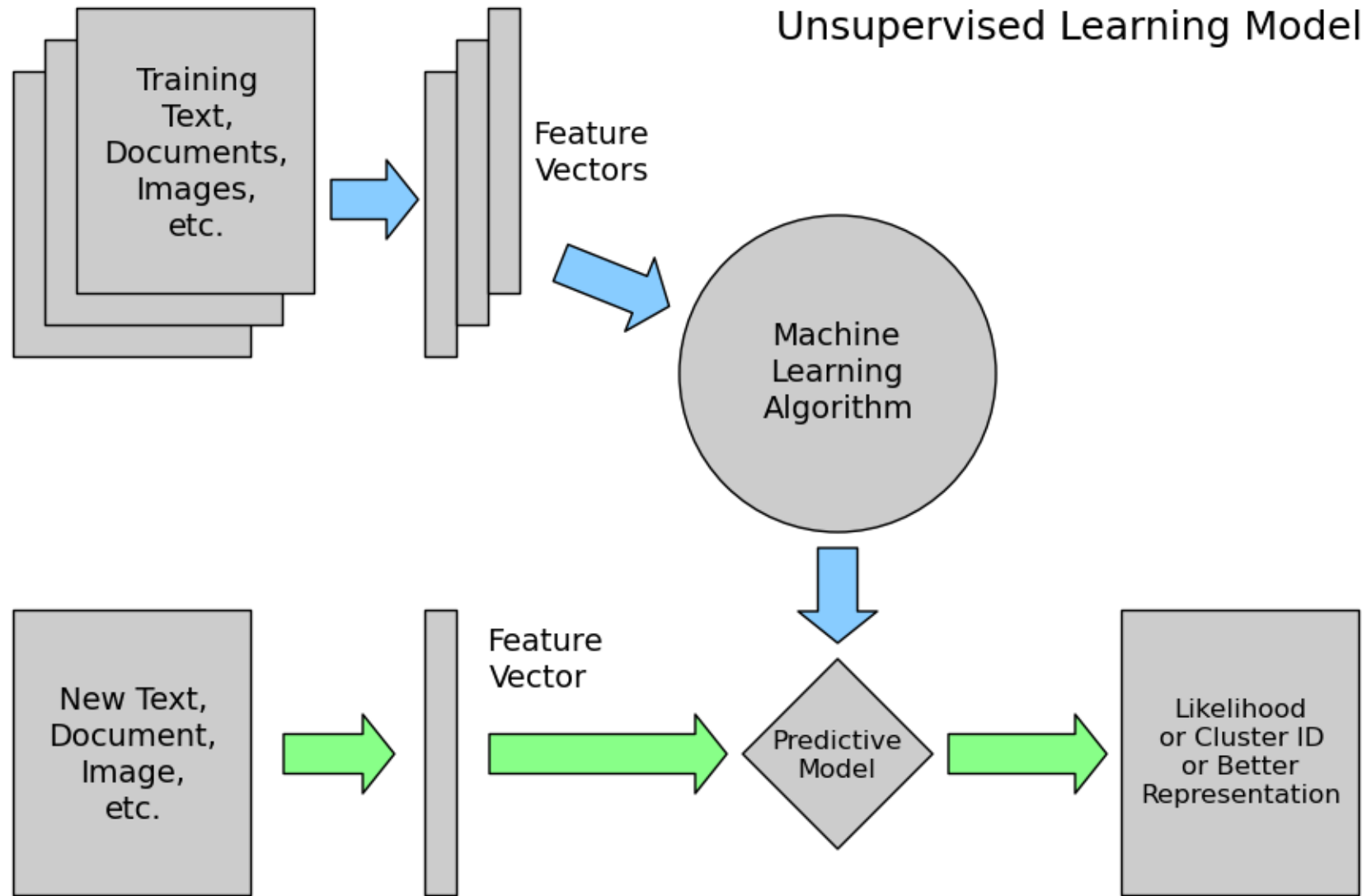
Unsupervised Learning
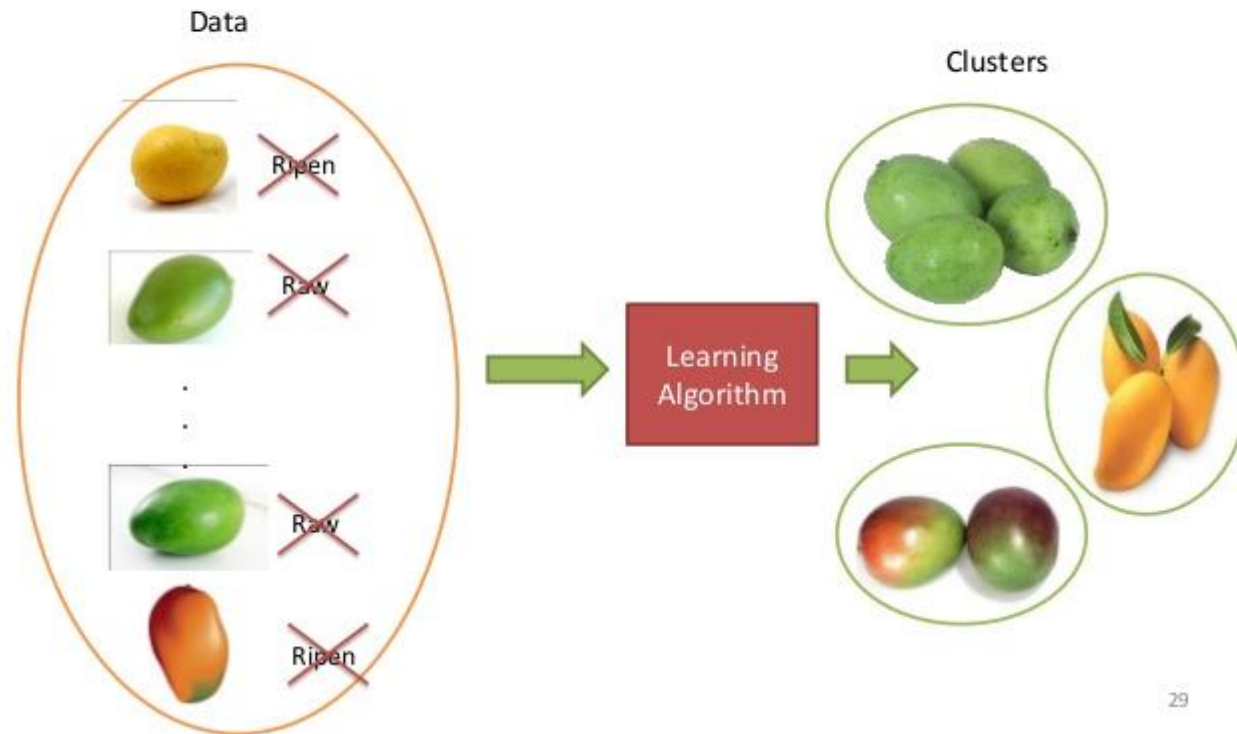
Unsupervised Learning

Objective is simply to divide above Images into N groups
Here ideally N = 2

The method can also assume other groups e.g. Group with one object or group with multiple objects

Outlier Detection: identify abnormal patterns

Example: identify engine anomalies
Features:
- Heat generated
- Vibration of engine

Hortonworks
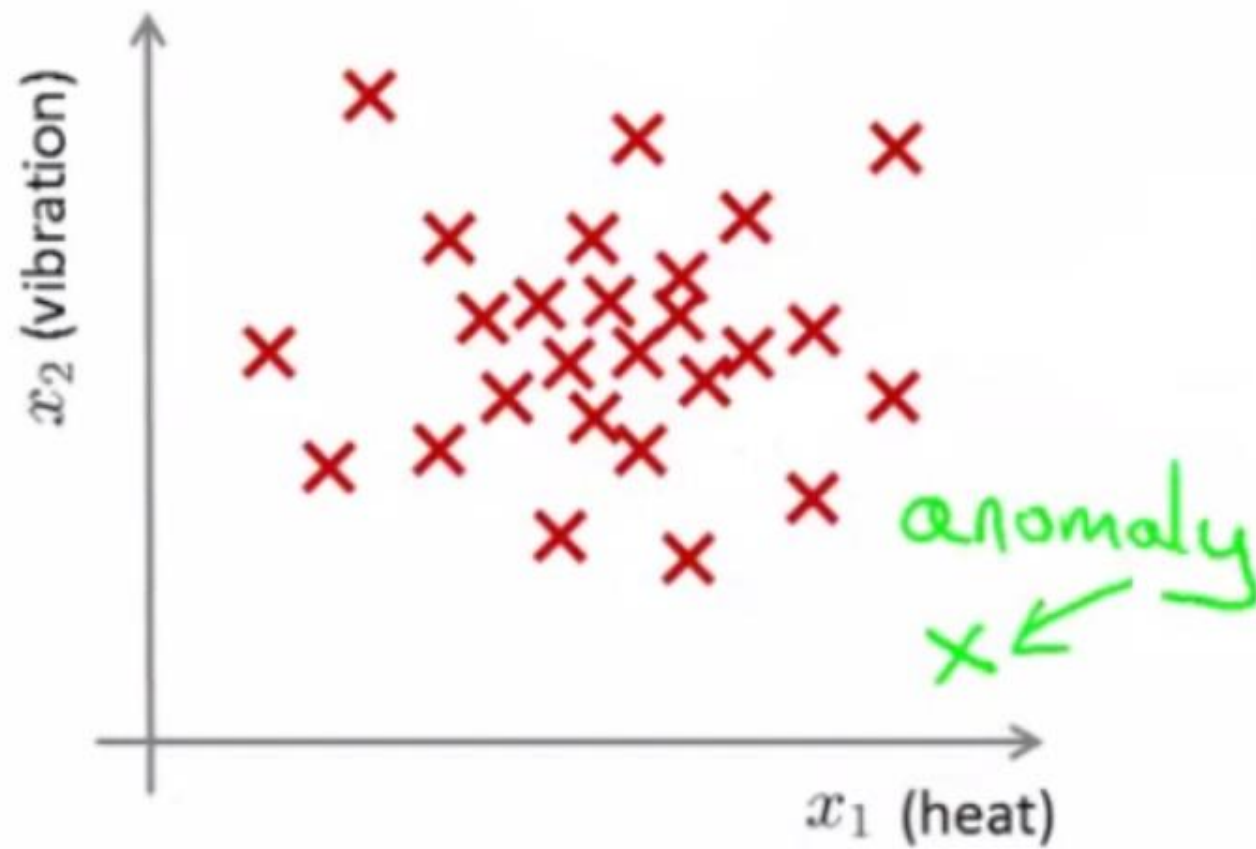
Affinity Analysis: identifying frequent item sets

Goal: identify frequent item set
Techniques: FP Growth, a priori

Product recommendation: predicting "preference"
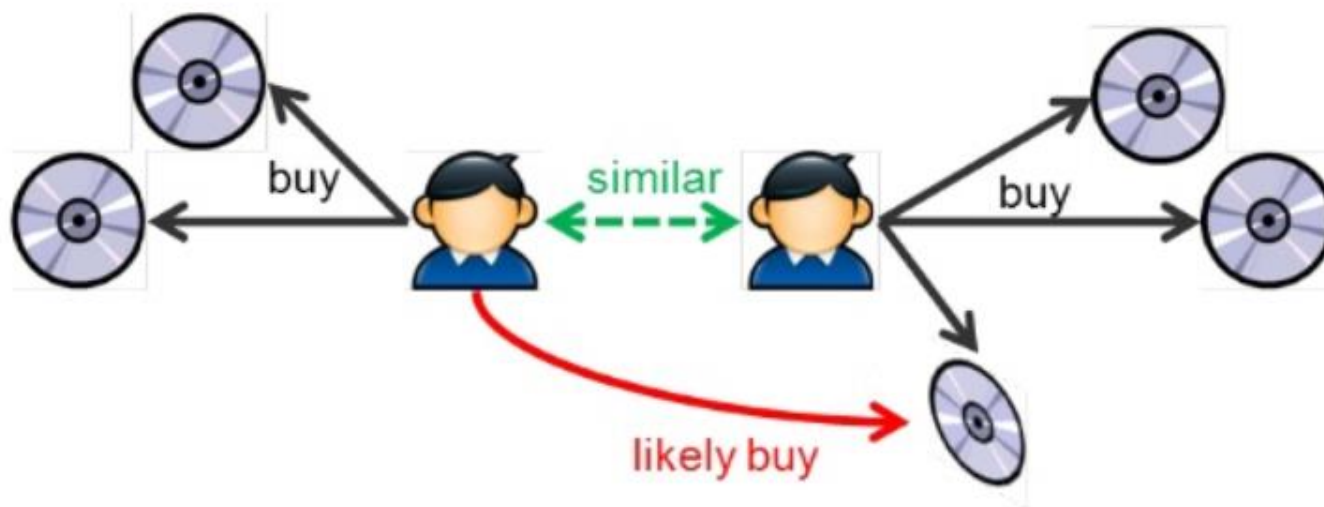
Collaborative Filtering
Identify users with similar "taste"

Hortonworks

Collaborative filtering -> matrix completion

	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	?	?
102	?	?	5	2	?
103	1	2	?	?	3
104					
105					
…					

	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	1	3
102	4	1	5	2	3
103	1	2	4	1	3
104					
105					
…					

Hortonworks

Example: Netflix
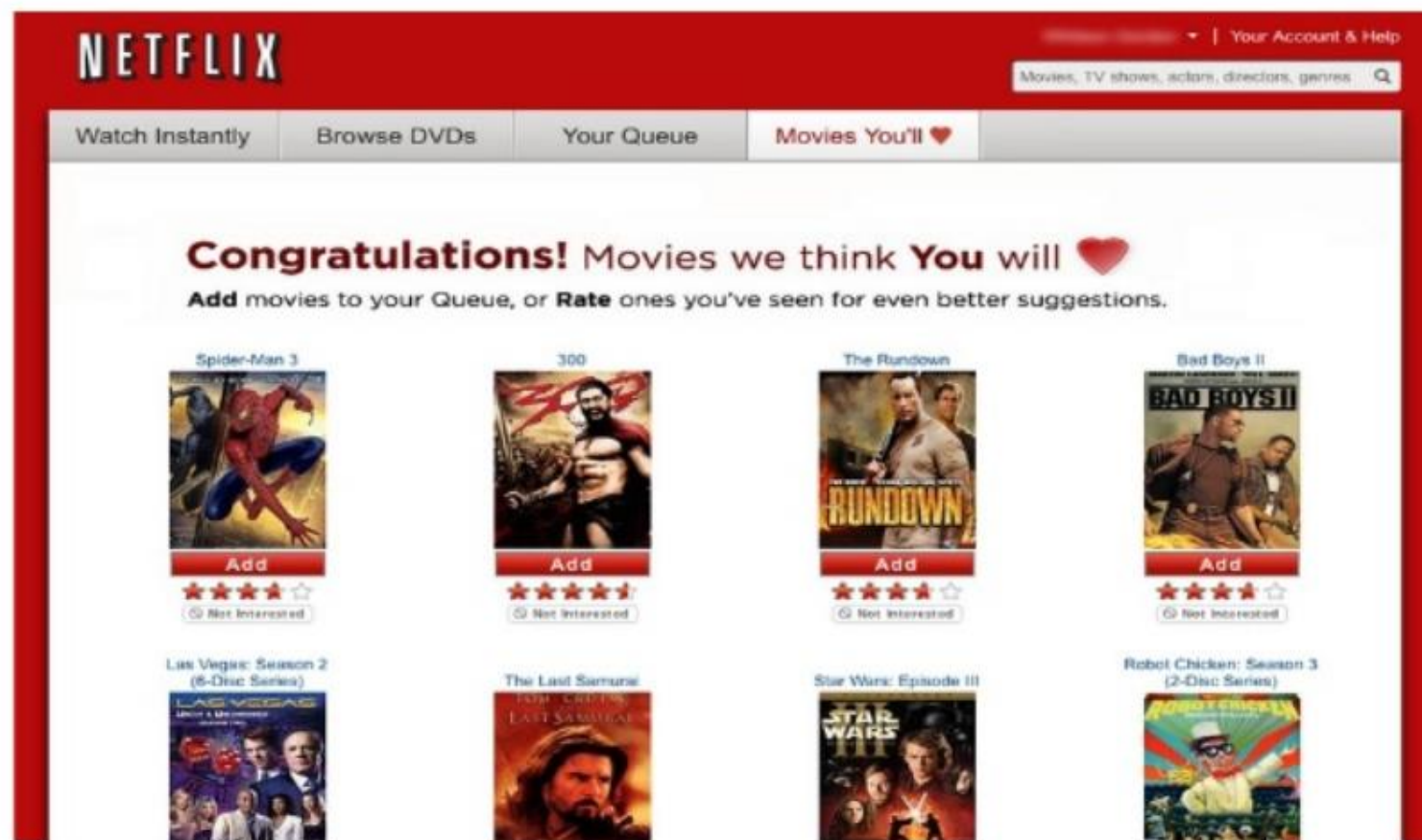
Example: market segmentation

Types of Unsupervised Learning

- In Clustering similar instances are grouped, based on their features or properties.

- **Association**: Association rules find associations amongst items within large commercial databases (e.g. Collaborative filtering)
  - Discover rules that describe large portions of data, such as people that buy X also tend to buy Y

  _____

- The similarity between two objects is measured by the **similarity function**
  - The distance between those two object is measured.
  - Shorter the distance higher the similarity, conversely longer the distance higher the dissimilarity.

Reinforcement learning

- Reinforcement learning is a type of dynamic programming that trains algorithms using a system of reward and punishment.

- A reinforcement learning algorithm, or agent, learns by interacting with its environment without intervention from a human by maximizing its reward and minimizing its penalty.

  - The agent receives rewards by performing correctly and penalties for performing incorrectly.

- Reinforcement learning contrasts with other machine learning approaches in that the algorithm is not explicitly told how to perform a task, but works through the problem on its own.

Reinforcement learning

- As an agent, which could be a self-driving car or a program playing chess, interacts with its environment, receives a reward state depending on how it performs, such as driving to destination safely or winning a game.

- Conversely, the agent receives a penalty for performing incorrectly, such as going off the road or being checkmated.
  - The agent over time makes decisions to maximize its reward and minimize its penalty using dynamic programming.

- The advantage of this approach to artificial intelligence is that it allows an AI program to learn without a programmer spelling out how an agent should perform the task.

forward pass

image | block of differentiable compute (e.g. neural net)

backward pass

log probabilities

| -1.2 | -0.36 |

gradients

| 0 | **-1.0** |

sample an action:
sampled action = 1

eventual reward -1.0

Reinforcement Learning
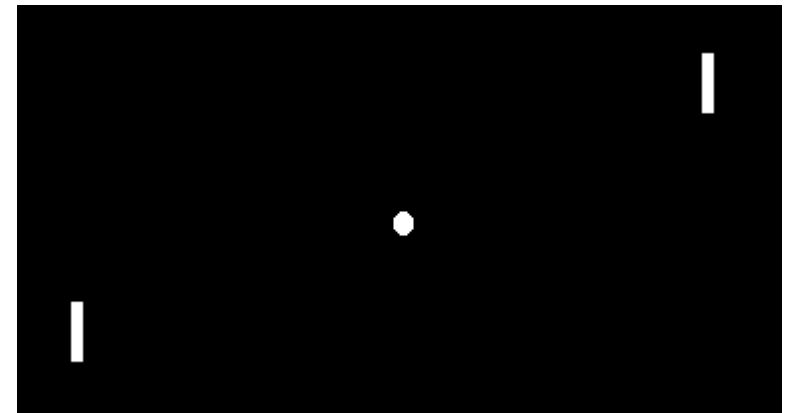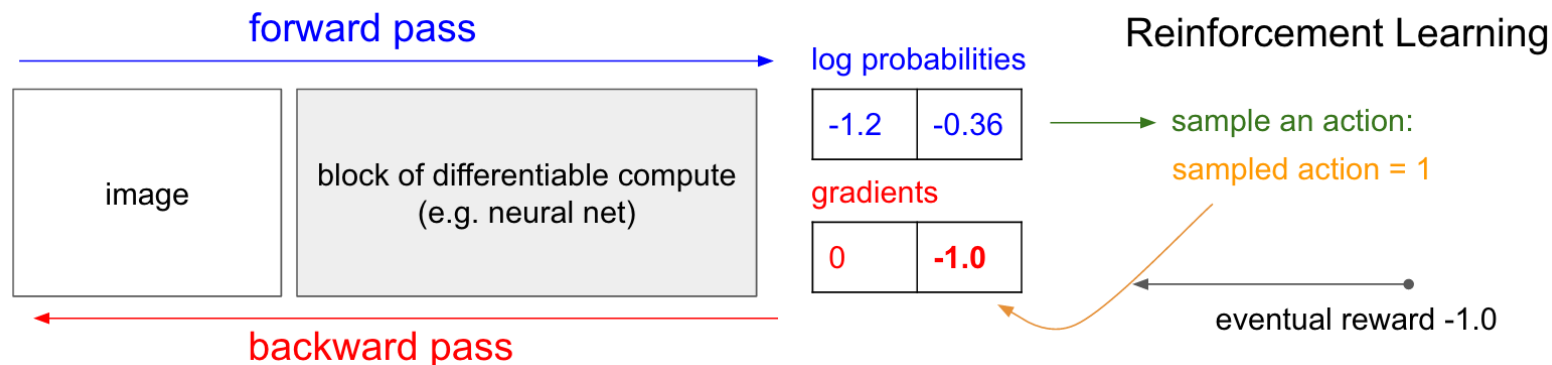
Reinforcement learning

- The agent is supposed to find the best possible path to reach the reward.
- The goal of the robot is to get the reward that is the diamond and avoid the hurdles that is fire.
- The robot learns by trying all the possible paths and then choosing the path which gives him the reward with the least hurdles.
- Each right step will give the robot a reward and each wrong step will subtract the reward of the robot.
- The total reward will be calculated when it reaches the final reward that is the diamond.

forward pass

| image | block of differentiable compute (e.g. neural net) |

log probabilities

| -1.2 | -0.36 |

gradients

| 0 | -1.0 |

backward pass

Reinforcement Learning

sample an action:

sampled action = 1

eventual reward -1.0
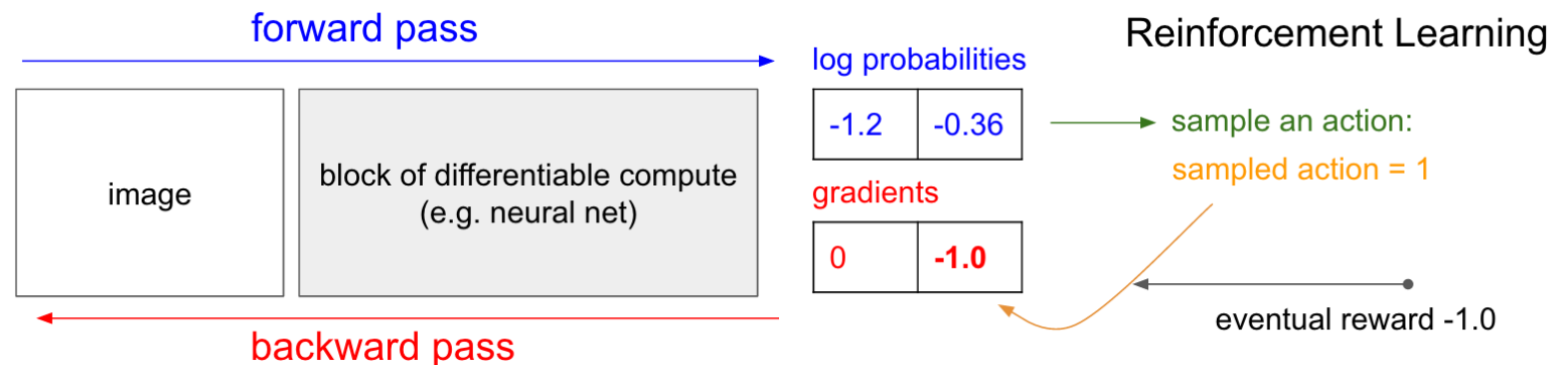
The clustering Problem:

- Given an integer k and a set of n data points in R^d, the goal is to choose k centers so as to minimize φ, the total squared distance between each point and its closest center.

- Solving this problem exactly is NP-hard, but Lloyd proposed a local search solution to this problem.

- k-means is the most popular clustering algorithm used in scientific and industrial applications" [3]

k-means Clustering

- K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups).

- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable **k**.

- The algorithm works iteratively to assign each data point to one of **k** groups based on the features that are provided.

- Data points are clustered based on feature similarity.

The results of the **k**-means clustering algorithm are:

- The centroids of the k clusters, which can be used to label new data
- Labels for the training data (each data point is assigned to a single cluster)

k-means Clustering

The k-means algorithm is a simple and fast algorithm for this problem, although it offers no approximation guarantees at all.

1. Arbitrarily choose an initial k centers C={c_1, c_2,$\cdots$, c_k}.
2. For each i $\in$ {1, . . . , k}, set the cluster C_i to be the set of points in X that are closer to c_i than they are to c_j for all j$\neq$i.
3. For each j $\in$ {1, . . . , k}, set c_i to be the center of mass of all points in C_i: $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.
4. Repeat Steps 2 and 3 until C no longer changes.

$$\text{e.g. } c_i = (170 + 168)/2 \mid = (60+56)/2$$

It is standard practice to choose the initial centers uniformly at random from X.

For Step 2, ties may be broken arbitrarily, as long as the method is consistent.

The idea here is that Steps 2 and 3 are both guaranteed to decrease φ, so the algorithm makes local improvements to an arbitrary clustering until it is no longer possible to do so.

Business Uses

- The *K*-means clustering algorithm is used to find groups which have not been explicitly labeled in the data.

- This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

Some examples of use cases are:

1. Behavioral segmentation:

    Segment by purchase history

    Segment by activities on application, website, or platform

    Define personas based on interests

    Create profiles based on activity monitoring

2. Inventory categorization:

- Group inventory by sales activity
- Group inventory by manufacturing metrics

3. Sorting sensor measurements:

- Detect activity types in motion sensors
- Group images
- Separate audio
- Identify groups in health monitoring

4. Detecting bots or anomalies:

- Separate valid activity groups from bots
- Group valid activity to clean up outlier detection

In addition, monitoring if a tracked data point switches between groups over time can be used to detect meaningful changes in the data.

K-means Clustering

- **D**etermines the centroid using the Euclidean method for distance calculation.

- **G**roups the objects based on minimum distance.
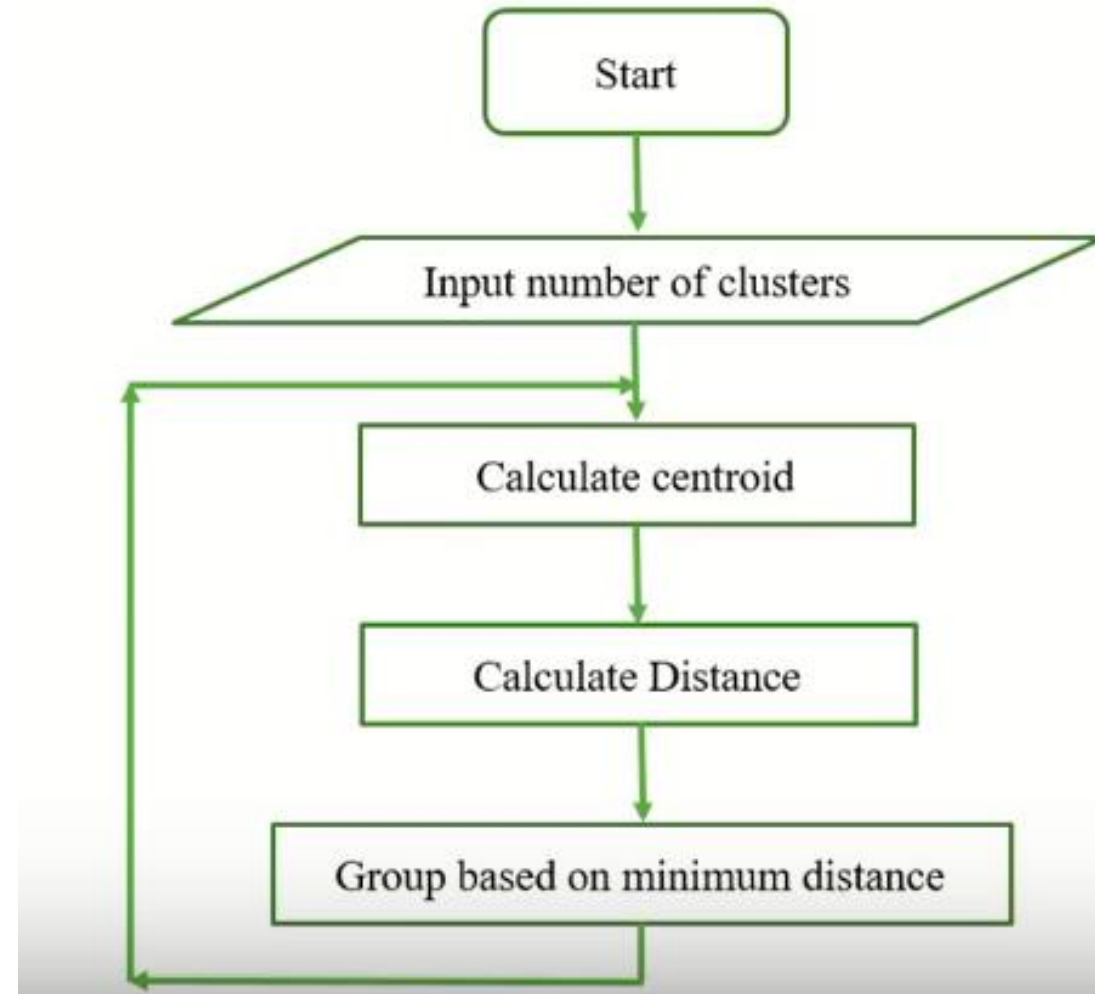
It analyze and explore the whole dataset.

K-means Clustering

Apply K-Mean Clustering for the following data sets for two clusters. Tabulate all the assignments.

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

k-means Clustering

- **G**iven $k = 2$

Initial Centroid

Cluster	X	Y
k1	185	72
k2	170	56

- **C**alculate Euclidean distance using the given equation.

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2 + (x-b)^2}$$

Cluster 1 $(185,72) = \sqrt{(185-185)^2 + (72-72)^2} = 0$

Distance from Cluster 2 $= \sqrt{(170-185)^2 + (56-72)^2}$

$$(170,56) = \sqrt{(-15)^2 + (-16)^2}$$

$$= \sqrt{255 + 256}$$

$$= \sqrt{481}$$

$$= 21.93$$

Distance from Cluster 1 $= \sqrt{(185-170)^2 + (72-56)^2}$

$(185,72)$ $\quad = \sqrt{(15)^2 + (16)^2}$

$$= \sqrt{255 + 256}$$

$$= \sqrt{481}$$

$$= 21.93$$

Cluster 2 $(170,56) = \sqrt{(170-170)^2 + (56-56)^2} = 0$

Cluster	Centroid		
	X	Y	ASSIGNMENT
k1	0	21.93	1
k2	21.93	0	2

Initial Centroid		
Cluster	X	Y
k1	185	72
k2	170	56

- **C**alculate Euclidean distance for the next dataset (168,60)

Distance $[(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$

Distance from Cluster 1 $= \sqrt{(168-185)^2+(60-72)^2}$

$\quad (185,72) \qquad = \sqrt{(-17)^2+(-12)^2}$

$\qquad\qquad\qquad = \sqrt{283+144}$

$\qquad\qquad\qquad = \sqrt{433}$

$\qquad\qquad\qquad = 20.808$

Distance from Cluster 2 $= \sqrt{(168-170)^2+(60-56)^2}$

$\quad (170,56) \qquad = \sqrt{(-2)^2+(-4)^2}$

$\qquad\qquad\qquad = \sqrt{4+16}$

$\qquad\qquad\qquad = \sqrt{20}$

$\qquad\qquad\qquad = 4.472$

k-means Clustering

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(168,60)	20.808	4.472	2

- **U**pdate the cluster centroid.

Cluster	X	Y
k1	185	72
k2	= (170 + 168)/ 2 = 169	= (60+56)/ 2 = 58

- Calculate Euclidean distance for the next dataset (179,68)

$$\text{Distance }[(x,y), (a,b)] = \sqrt{(x-a)^2 + (x-b)^2}$$

Distance from Cluster 1 =
(185,72)

$$= \sqrt{(179-185)^2 + (68-72)^2}$$
$$= \sqrt{(-6)^2 + (-4)^2}$$
$$= \sqrt{36 + 16}$$
$$= \sqrt{52}$$
$$= 7.211103$$

- Calculate Euclidean distance for the next dataset (179,68)

$$\text{Distance }[(x,y), (a,b)] = \sqrt{(x-a)^2 + (x-b)^2}$$

Distance from Cluster 2 =
(169,58)

$$= \sqrt{(179-169)^2 + (68-58)^2}$$
$$= \sqrt{(10)^2 + (10)^2}$$
$$= \sqrt{100 + 100}$$
$$= \sqrt{200}$$
$$= 14.14214$$

The k-means Clustering

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(179,68)	7.211103	14.14214	1

- Update the cluster centroid.

Cluster	X	Y
k1	= 185+179/2 =182	= 72+68/2 =70
k2	169	58

- Calculate Euclidean distance for the next dataset (182,72)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

Distance from Cluster 1 $= \sqrt{(182-182)^2+(72-70)^2}$

(182,70) $= \sqrt{(0)^2+(2)^2}$

$= \sqrt{0+4}$

$= \sqrt{4}$

$= 2$

- Calculate Euclidean distance for the next dataset (182,72)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

Distance from Cluster 2 $= \sqrt{(182-169)^2+(72-58)^2}$

(169,58) $= \sqrt{(13)^2+(14)^2}$

$= \sqrt{169+196}$

$= \sqrt{365}$

$= 19.10$

The k-means Clustering

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(182,72)	2	19.10	1

- Update the cluster centroid.

Cluster	X	Y
k1	= 182+182/2 =182	= 70+72/2 = 71
k2	169	58

The k-means Clustering

- **F**inal Assignment

Dataset No	X	Y	Assignment
1	185	72	1
2	170	56	2
3	168	60	2
4	179	68	1
5	182	72	1
6	188	77	1

The k-means++ algorithm

We propose a specific way of choosing centers for the k-means algorithm. In particular, let $D(x)$ denote the shortest distance from a data point to the closest center we have already chosen. Then, we define the following algorithm, which we call k-means++.

1a. Take one center c_1, chosen uniformly at random from $\mathcal{X}$.

1b. Take a new center c_i, choosing $x \in \mathcal{X}$ with probability $\frac{D(x)^2}{\sum_{x \in \mathcal{X}} D(x)^2}$.

1c. Repeat Step 1b. until we have taken k centers altogether.

2-4. Proceed as with the standard k-means algorithm.

We call the weighting used in Step 1b simply "D^2 weighting".

The k-means ++ Algorithm

- The first step is to choose a data point at random. Call this point s_1. Next, compute the squared distances

$$D_i^2 = ||Y_i - s_1||^2.$$

- Now choose a second point s_2 from the data. The probability of choosing Y_i is $D_i^2 / \sum_j D_j^2$

- Now recompute the distance as $D_i^2 = \min\left\{||Y_i - s_1||^2, ||Y_i - s_2||^2\right\}.$

- Now choose a third point s_3 from the data where the probability of choosing Y_i is $D_i^2 / \sum_j D_j^2$.

- We continue until we have k points $s_1, s_2, s_3, \ldots, s_k$.

- Finally, we run k-means clustering using $s_1, s_2, s_3, \ldots, s_k$ as starting values. Call the resulting centers $c_1, c_2, c_3, \ldots, c_k$.

- Arthur and Vassilvitskii proved that the expected value is over the randomness in the algorithm

$$\mathbb{E}[R(\hat{c}_1, \ldots, \hat{c}_k)] \leq 8(\log k + 2) \min_{c_1, \ldots, c_k} R(c_1, \ldots, c_k).$$

- .

Nine Data Items in Two-Dimension into Three Clusters

Least likely third mean (smallest distance^2)

first initial mean

second initial mean

Most likely third mean (largest distance^2)
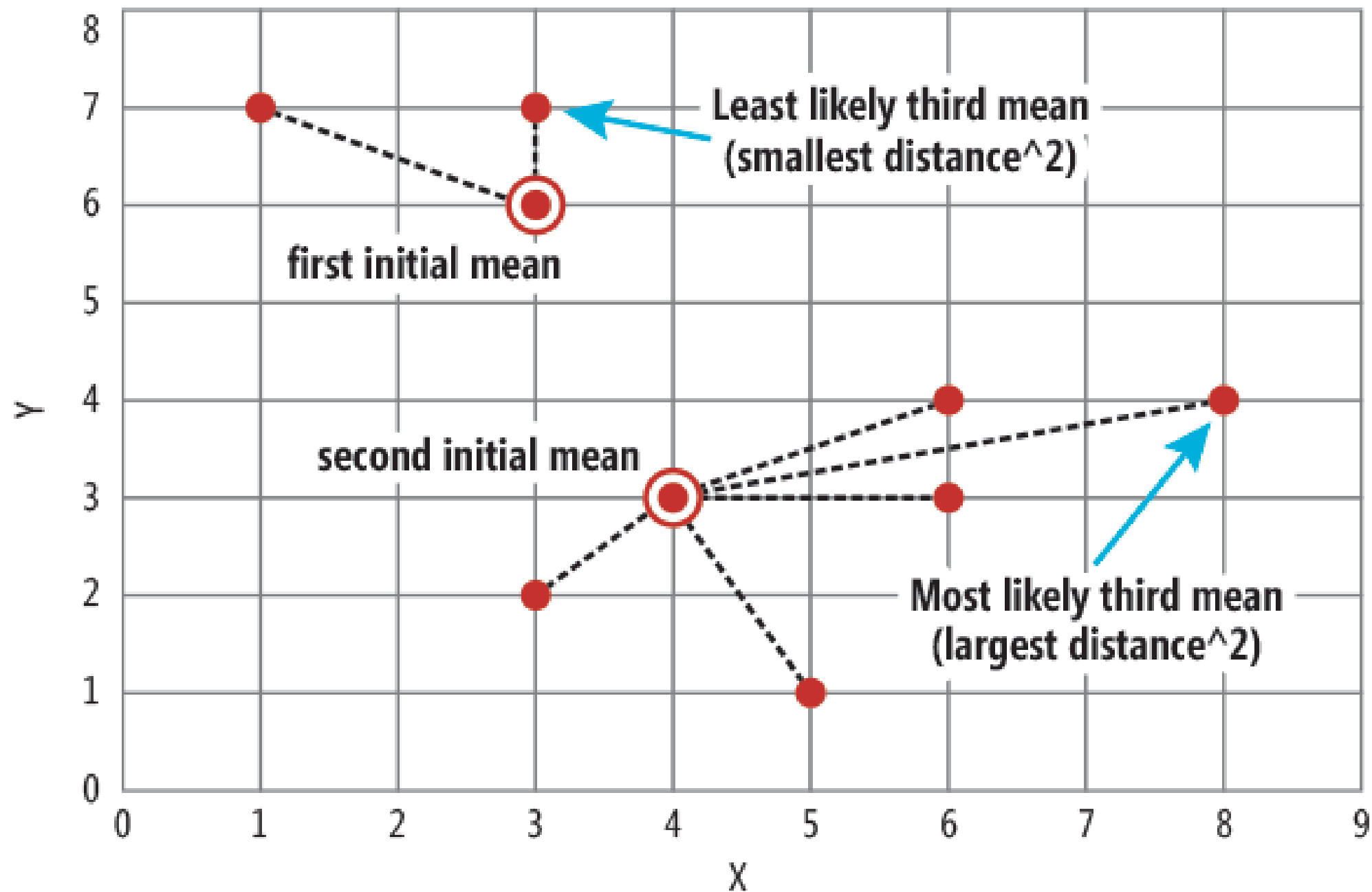
The k-means ++ Algorithm

- The first initial mean at (3, 6) was randomly selected.

- Then the distance-squared from each of the other 8 data items to the first mean was computed, and using that information, the second initial mean at (4, 3) was selected.

- To select a data item as the third initial mean, the squared distance from each data point to its closest mean is computed.

- The distances are shown as dashed lines.

- Using these squared distance values, the third mean will be selected so that data items with small squared distance values have a low probability of being selected, and data items with large squared distance values have a high probability of being selected.

- This technique is sometimes called proportional fitness selection.

The k-means ++ Algorithm: Proportional fitness selection using Roulette wheel selection

- Proportional fitness selection is the heart of the k-means++ initialization mechanism.
- There are several ways to implement proportional fitness selection.
- Here we use Roulette wheel selection for proportional fitness selection.
- Suppose there are four candidate items (0, 1, 2, 3) with associated values (20.0, 10.0, 40.0, 30.0).
- The sum of the values is 20.0 + 40.0 + 10.0 + 30.0 = 100.0.
- Proportional fitness selection will pick item 0 with probability 20.0/100.0 = 0.20; pick item 1 with probability 10.0/100.0 = 0.10; pick item 2 with probability 40.0/100.0 = 0.40; and pick item 3 with probability 30.0/100.0 = 0.30.

The k-means ++ Algorithm: Proportional fitness selection using Roulette wheel selection

- If the probabilities of selection are stored in an array as (0.20, 0.10, 0.40, 0.30), the cumulative probabilities can be stored in an array with values (0.20, 0.30, 0.70, 1.00).

- Now, suppose a random p is generated with value 0.83.

- If i is an array index into the cumulative probabilities array, when i = 0, cum[i] = 0.20, which isn't greater than p = 0.83, so i increments to 1.

- Now cum[i] = 0.30, which is still not greater than p, so i increments to 2.

- Now cum[i] = 0.70, which is still not greater than p, so i increments to 3.

- Now cum[i] = 1.00, which is greater than p, so i = 3 is returned as the selected item.

Classification Workflow

What is a classifier: A classifier is a machine learning model that is used to discriminate different objects based on certain features.

- Whenever you perform classification, the first step is to understand the problem and identify potential features and label.

- Features are those characteristics or attributes which affect the results of the label.

- For example, in the case of a loan distribution, bank manager's identify customer's occupation, income, age, location, previous loan history, transaction history, and credit score.

- These characteristics are known as features which help the model classify customers.

Classification Workflow

- The classification has two phases, a learning phase, and the evaluation phase
    - In the learning phase, classifier trains its model on a given dataset and
    - In the evaluation phase, it tests the classifier performance
- Performance is evaluated on the basis of various parameters such as accuracy, error, precision, and recall
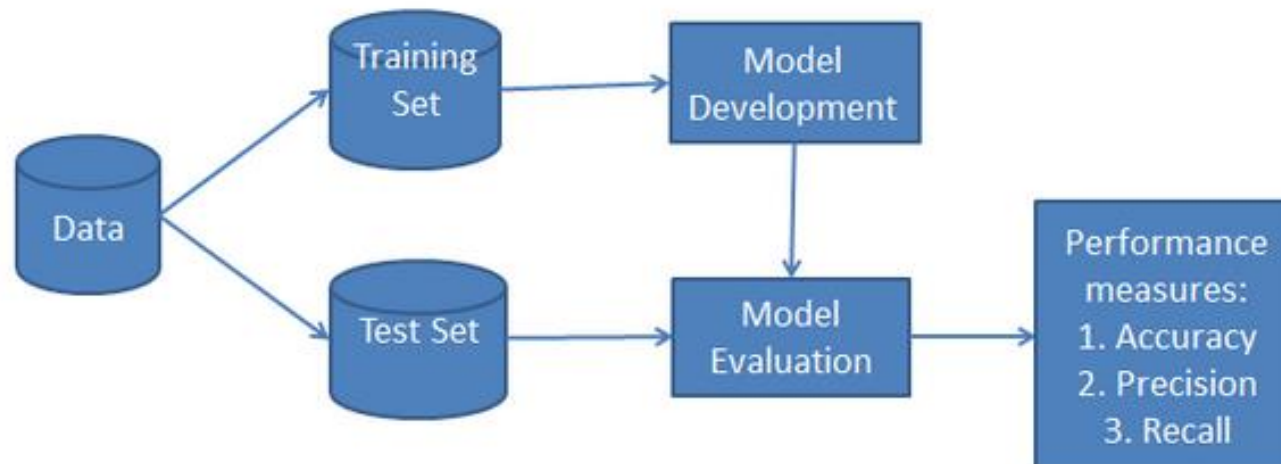
K Nearest Neighbors - Classification

- K nearest neighbors algorithm stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

- KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its k nearest neighbors measured by a distance function.

- If k = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance functions

Euclidean
$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan
$$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski
$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

All three distance measures are only valid for continuous variables

K Nearest Neighbors - Classification

- In the instance of categorical variables the Hamming distance must be used.
- It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$
$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

- Choosing the optimal value for K is best done by first inspecting the data.
- In general, a large K value is more precise as it reduces the overall noise but there is no guarantee.
- Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value.
- Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

k Nearest Neighbors – Classification

- k-NN is a non-parametric method used for classification

- Prediction for the test data is done on the basis of its neighbor

- k is an integer(small), if k=1, k is assigned to the class of single nearest neighbor

Name	Acid Durability	Strength	Class
Type-1	7	7	Bad
Type-2	7	4	Bad
Type-3	3	4	Good
Type-4	1	4	Good

Assume the Test Data is: Acid Durability=3, and Strength=7. What is the class?

k Nearest Neighbors – Classification

The similarity is calculated using distance measure like Euclidean

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Name	Acid Durability	Strength	Class	Distance
Type-1	7	7	Bad	Sqrt((7-3)^2 + (7-3)^2)=4
Type-2	7	4	Bad	5
Type-3	3	4	Good	3
Type-4	1	4	Good	3.6

k Nearest Neighbors – Classification

Rank these Attributes

Name	Acid Durability	Strength	Class	Distance	Rank
Type-1	7	7	Bad	4	3
Type-2	7	4	Bad	5	4
Type-3	3	4	Good	3	1
Type-4	1	4	Good	3.6	2

k Nearest Neighbors - Classification

k= 1

Name	Acid Durability	Strength	Class	Distance	Rank
Type-1	7	7	Bad	4	3
Type-2	7	4	Bad	5	4
Type-3	3	4	Good	3	1
Type-4	1	4	Good	3.6	2

Acceptance level is good in the two neighbors

Naive Bayes Classification

- Suppose you are a product manager, you want to classify:
  - Customer reviews in positive and negative classes.
  - As a loan manager, you want to identify which loan applicants are safe or risky?
  - As a healthcare analyst, you want to predict which patients can suffer from diabetes disease.
- All the examples have the same kind of problem to classify reviews, loan applicants, and patients.
- Naive Bayes is the most straightforward and fast classification algorithm, which is suitable for a large chunk of data.
- Naive Bayes classifier is successfully used in various applications such as:
  - spam filtering, text classification, sentiment analysis, and recommender systems.
- It uses Bayes theorem of probability for prediction of unknown class.

Naive Bayes Classifier

- A Naive Bayes classifier is a probabilistic machine learning model that is used for classification task. The classifier is based on the Bayes theorem.

- We can find the probability of **A** happening, given that **B** has occurred by the Bayes Theorem as:

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

- Here, **B** is the evidence and **A** is the hypothesis.

- P(A\B) : Probability of A when B is True
- P(A) : Probability of Hypothesis
- P(B): Probability of Evidence
- P(B\A) : Probability of B when A is True

Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.

- P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

- P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.

- P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

Naive Bayes Classifier

- The Naive Bayes classifier assumes that the predictors/features are independent.
    - The presence of one particular feature does not affect the other.
    - The presence of a feature in a class is unrelated to any other feature.
- Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that a particular fruit is an apple or an orange or a banana, and that is why it is known as **"Naive."**
- Naive Bayes classifiers have high accuracy and speed on large datasets.

Consider the problem of playing golf. The dataset is represented as below.

We classify whether the day is suitable for playing golf, given the features of the day. According to this example, Bayes theorem can be written as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

The variable **y** is the class variable(play golf), which represents if it is suitable to play golf or not given the conditions X. Variable **X** represent the parameters/features. **X** is given as,

$$X = (x_1, x_2, x_3, \ldots, x_n)$$

$$P(y|x_1, \ldots, x_n) =$$

$$\frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

HOW NAIVE BAYES CLASSIFIER WORKS?

Whether	Temperature	Play
Sunny	Hot	No
Sunny	Hot	No
Overcast	Hot	Yes
Rainy	Mild	Yes
Rainy	Cool	Yes
Rainy	Cool	No
Overcast	Cool	Yes
Sunny	Mild	No
Sunny	Cool	Yes
Rainy	Mild	Yes
Sunny	Mild	Yes
Overcast	Mild	Yes
Overcast	Hot	Yes
Rainy	Mild	No

01 CALCULATE PRIOR PROBABILITY FOR GIVEN CLASS LABELS

02 CALCULATE CONDITIONAL PROBABILITY WITH EACH ATTRIBUTE FOR EACH CLASS

03 MULTIPLY SAME CLASS CONDITIONAL PROBABILITY.

04 MULTIPLY PRIOR PROBABILITY WITH STEP 3 PROBABILITY.

05 SEE WHICH CLASS HAS HIGHER PROBABILITY, HIGHER PROBABILITY CLASS BELONGS TO GIVEN INPUT SET STEP.

Question on NAIVE BAYE's Algorithm:

Ques:) For the given dataset, Apply Naive-Baye's Algorithm and Predict the outcome for a Car = { Red, Domestic, SUV }

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	NO
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	NO
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	NO
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	NO
Red	SUV	Imported	NO
Red	Sports	Imported	Yes

Prob. of B when A is True (likelihood)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \longrightarrow \text{Proposition}$$

Posterior — $P(A|B)$

Prob (A) when B is true

evidence — $P(B)$

→ (NO) ANS..

$X = [\text{Red, Domestic, SUV}] = \frac{P(X|Yes)}{}$

$P(X|Yes) \in \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} = \frac{6}{125} = 0.024$ Yes

i) $P(Red|Yes) = \frac{P(Yes|Red) \cdot P(Red)}{P(Yes)} = \frac{\frac{3}{5} \cdot \frac{5}{10}}{\frac{5}{10}} \cdot \frac{3}{5}$

ii) $P(Domestic|Yes) = 2/5$ iii) $P(SUV|Yes) = 1/5$

$P(Red|No) = \frac{P(No|Red) \cdot P(Red)}{P(No)} = \frac{\frac{2}{5} \cdot \frac{5}{10}}{\frac{5}{10}} = 2/5$

$P(Domestic|No) = 3/5, \quad P(SUV|No) = 2/5 = \frac{2}{5} \times \frac{3}{5} \times \frac{2}{5} = 0.072$ NO

Bayes' Theorem Example

- Let's suppose we have a Deck of Cards and we wish to find out the probability of the card we picked at random to being a king, given that it is a face card. So, according to Bayes' Theorem, we can solve this problem. First, we need to find out the probability:

- **P(King)** which is **4/52** as there are 4 Kings in a Deck of Cards.

- **P(Face|King)** is equal to **1** as all the Kings are face Cards.

- **P(Face)** is equal to **12/52** as there are 3 Face Cards in a Suit of 13 cards and there are 4 Suits in total.

$$P(King|Face) = \frac{P(Face|King).P(King)}{P(Face)}$$

$$= \frac{1.(1/13)}{3/13} = 1/3$$

P(King) = 4/52 =1/13

P(Face|King) = 1

P(Face) =12/52 = 3/13