Machine Learning

BITS Pilani
Pilani Campus

Dr. Monali Mavani

Machine Learning

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content flow to suit the requirements of the course and for ease of class presentation
- Students are requested to refer to the textbook and detailed content of this presentation deck over canvas

Course Introduction

- **Objective of course**
  - Introduction to the basic concepts and techniques of Machine Learning
  - Gain experience in basics of doing independent study and research in the field of Machine Learning
  - Develop skills of using recent machine learning software tools to evaluate learning algorithms and model selection for solving practical problems
- **Focus of this course**
  - Strong Mathematical Foundations of ML algorithms
  - Structured Data Analytics
  - IDD (Independent & Identically Distributed Data)
- **Topics not expected of this course**
  - Unstructured Data Analytics
  - Time Series/Sequence Data Analytics
  - Deep Learning

Course Plan

M1	Introduction
M2	Machine learning Workflow
M3	Linear Models for Regression
M4	Linear Models for Classification
M5	Decision Tree
M6	Instance Based Learning
M7	Support Vector Machine
M8	Bayesian Learning
M9	Ensemble Learning
M10	Unsupervised Learning
M11	Machine Learning Model Evaluation/Comparison

Pre-requisites

- Linear algebra: vector/matrix manipulations, properties
- Calculus: partial derivatives
- Probability: common distributions; Bayes Rule
- Statistics: mean/median/mode; maximum likelihood

Text books and Reference book(s)

T1 Tom M. Mitchell: Machine Learning, The McGraw-Hill Companies

R1 Christopher M. Bishop: Pattern Recognition & Machine Learning, Springer

P. Tan, et al. Introduction to Data Mining, Pearson

R2 C.J.C. BURGES: A Tutorial on Support Vector Machines for Pattern Recognition,

R3 Kluwer Academic Publishers, Boston.

Evaluation scheme

- **Quiz (10% - Best 2 of 3 quizzes)**
- **Assignment (20% - 1 Progressive)**
- **Mid-semester exam (30%)**
- **Comprehensive exam (40%)**

Lab Plan

Lab No.	Lab Objective
1	End to End Machine Learning
2	Linear Regression and Gradient Descent Algorithm
3	Logistic Regression Classifier
4	Decision Tree
5	Naïve Bayes Classifier
6	Random Forest

- **Labs not graded**
- **Most of the Lab recordings available at CSIS virtual labs**
- **Webinars will be conducted for lab sessions**
- **Labs will be conducted in Python**

Agenda

- What is Machine Learning?
- Why Machine Learning is important?
- Types of Machine Learning
- ML workflow
- Few Terminologies
- Data types
- Demo

What is Machine Learning

How can we solve a specific problem?

- we write a program that encodes a set of rules that are useful to solve the problem

- Write a program : given a picture determine whether there is a cat in the image

- Learning systems are not directly programmed to solve a problem, instead develop own program based on:
  - Examples of how they should behave

- Learning simply means incorporating information from the training examples into the system

What is Machine Learning?

- The science (and art) of programming computers so they can *learn from data*
- More general definition
  - Field of study that gives computers the ability to learn without being explicitly programmed
  -
- Engineering-oriented definition
  - Algorithms that improve their performance P at some task T with experience E
  - A well-defined learning task is given by <*P*, *T*, *E*>

Defining the Learning Tasks

Improve on task T, with respect to performance metric P, based on experience E

Example 1

T: Recognizing hand-written words

P: Percentage of words correctly classified

E: Database of human labelled images of handwritten words

Example 2

T: Categorize email messages as spam or legitimate.

P: Percentage of email messages correctly classified.

E: Database of emails, some with human-given labels

Traditional Approach - Spam Filtering

Spam typically uses words or phrases such as “4U,” “credit card,” “free,” and “amazing”

Solution

Write a detection algorithm for frequently appearing patterns in spams

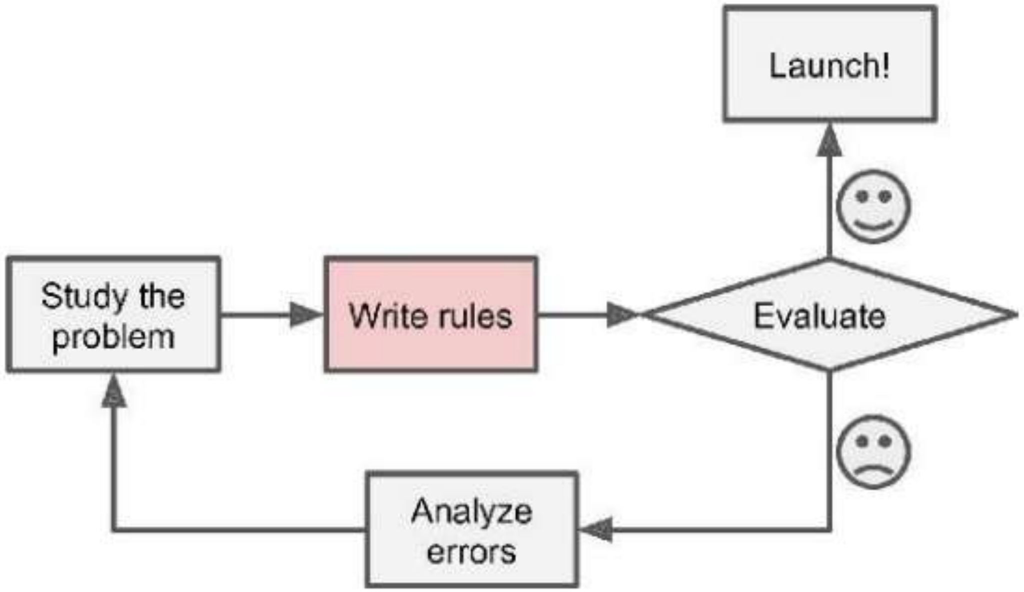
Test and update the detection rules until it is good enough.

Challenge

Detection algorithm likely

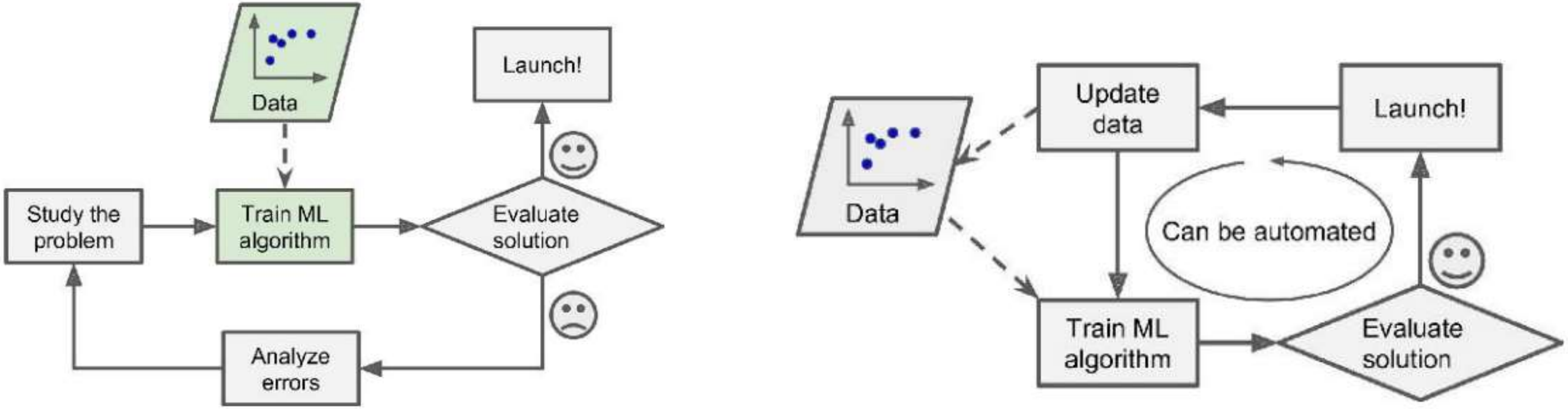
to be a long list of complex rules

hard to maintain.



ML Approach - Spam Filtering

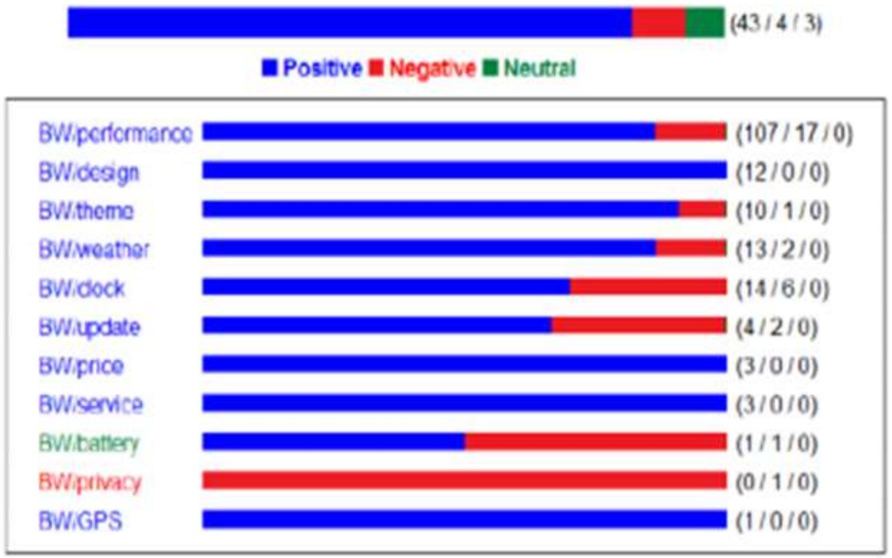
Automatically learns phrases that are good predictors of spam by detecting unusually frequent patterns of words in spams compared to “ham”

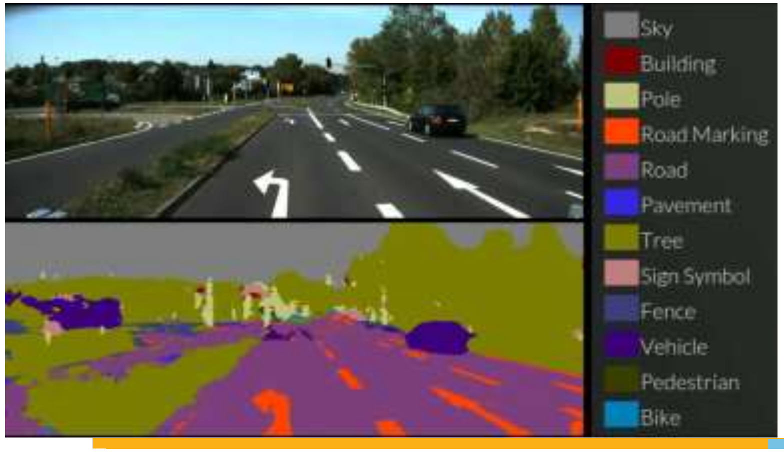


The program is much shorter, easier to maintain, and most likely more accurate.

Common Use cases - Security & Transaction Domain

Sentiment analysis on Product review of Mobile phone





- Self Driving Cars
- Fraud Detection in Banking
- Email Filtering
- Dynamic Pricing in Travel

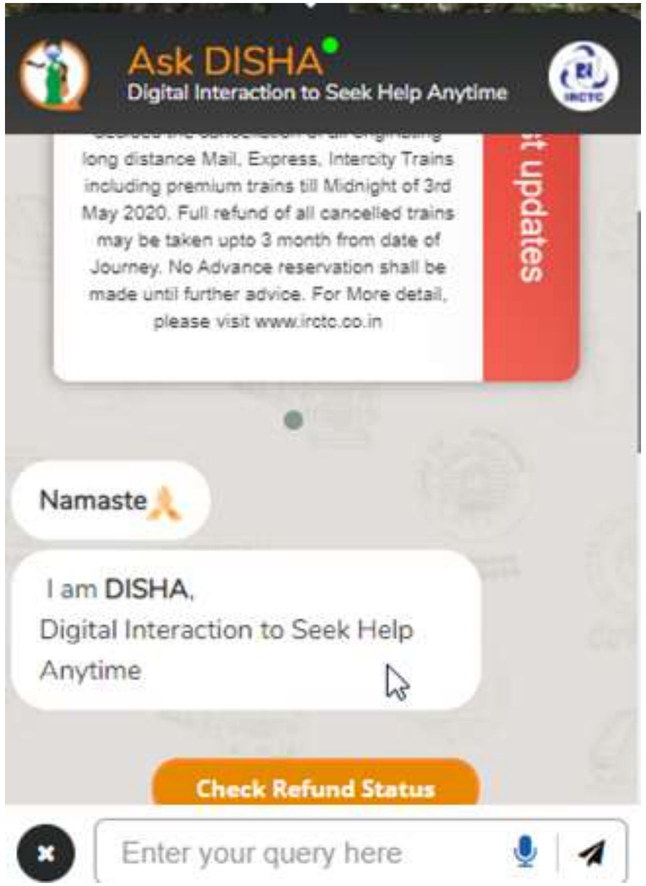
Derived Applications:

> Cyber Security

> Video Surveillance

> Object Detection

Common Use cases - Customer Support Systems



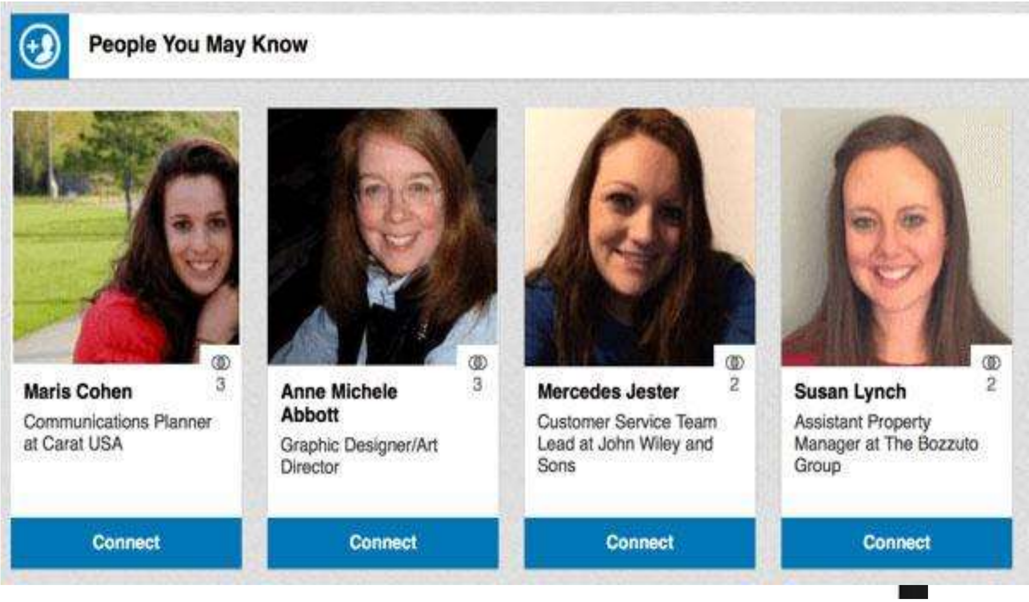
- Apple’s Siri
- Google Assistant
- Amazon’s Alexa
- Google Duplex
- Microsoft’s Cortana
- Samsung’s Bixby

Derived Applications:

> Customer Support Query (Voice vs Text)

> Chatbots

Common Use cases - Recommendation Engines



- E-commerce sites like Amazon and Flipkart
- Book sites like Goodreads
- Movie services like IMDb and Netflix
- Hospitality sites like MakeMyTrip, Booking.com, etc.
- Retail services like StitchFix
- Food aggregators like Zomato and Uber Eats

Derived Applications:

> Personalized Marketing

> Personalized Banking

Why ML

When Do We Use Machine Learning?

ML is used when:

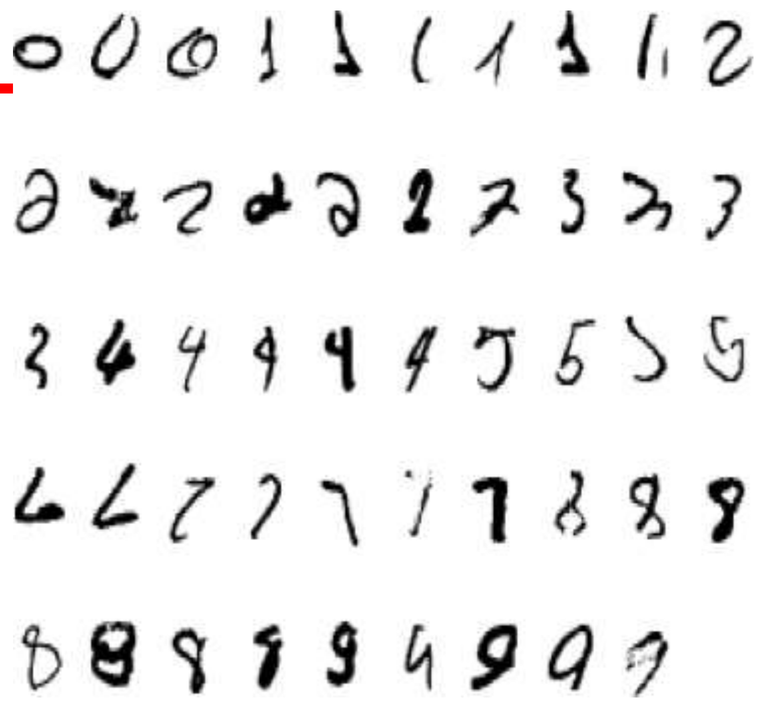
- Human expertise does not exist (navigating on Mars)

- Humans can’t explain their expertise (Biometrics)

- Models must be customized (personalized medicine)

Why only ML?

- Some tasks cannot be defined well, except by examples.
  - It is very hard to write programs that solve problems like recognizing a handwritten digit
  - What distinguishes a 2 from a 7?
  - How does our brain do it
- Hidden relationships and correlations in data
- large data makes it difficult for explicit encoding by humans (e.g., medical diagnostic)
- Continuous availability of new knowledge



Pattern recognition

Problems not to be solved using ML

- Learning isn't always useful:
  - Tasks in which humans are very effective
  - Tasks in which frequent human intervention is needed
  - Simple tasks which can be implemented using traditional programming paradigms
  - Situations where training data is not sufficient

Types of ML

Types of Learning Inputs: Based on level of supervision

- **Supervised (inductive) learning** Given: training data, desired outputs (labels)
- **Unsupervised learning**
  - Given: training data (without desired outputs)
- **Semi-supervised learning**
  - Given: training data + a few desired outputs
- **Reinforcement learning**
  - Given: rewards/penalty from sequence of actions

Slide Credit: Eric Eaton

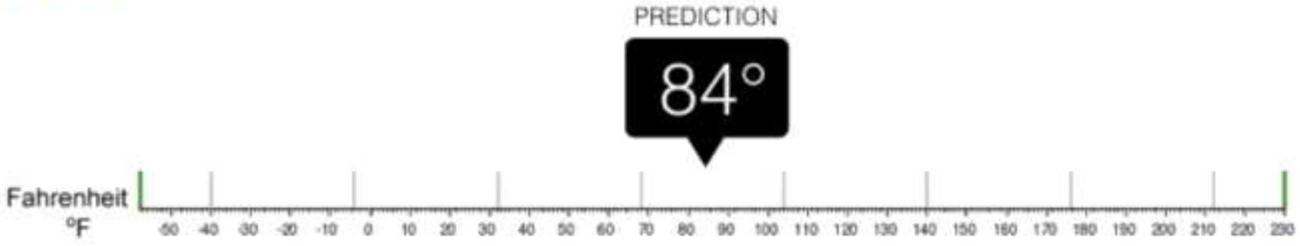
Supervised (inductive) learning

Supervised Learning Techniques / Algorithms

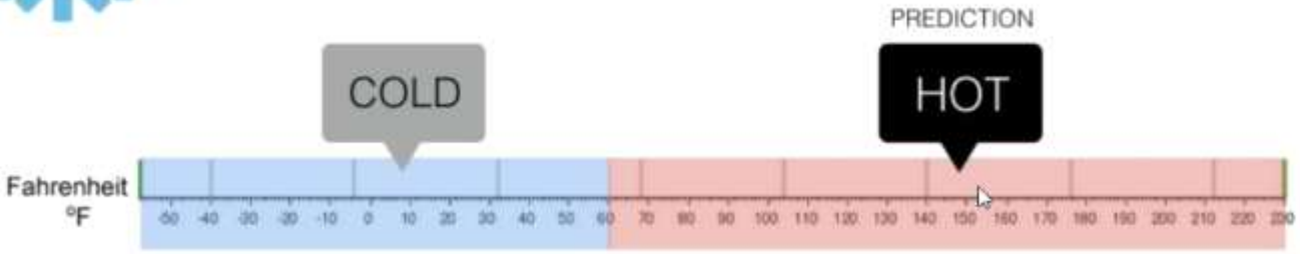
- Linear Regression
- Logistic Regression
- Naïve Bayes Classifiers
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural networks

Supervised Learning – Regression & Classification

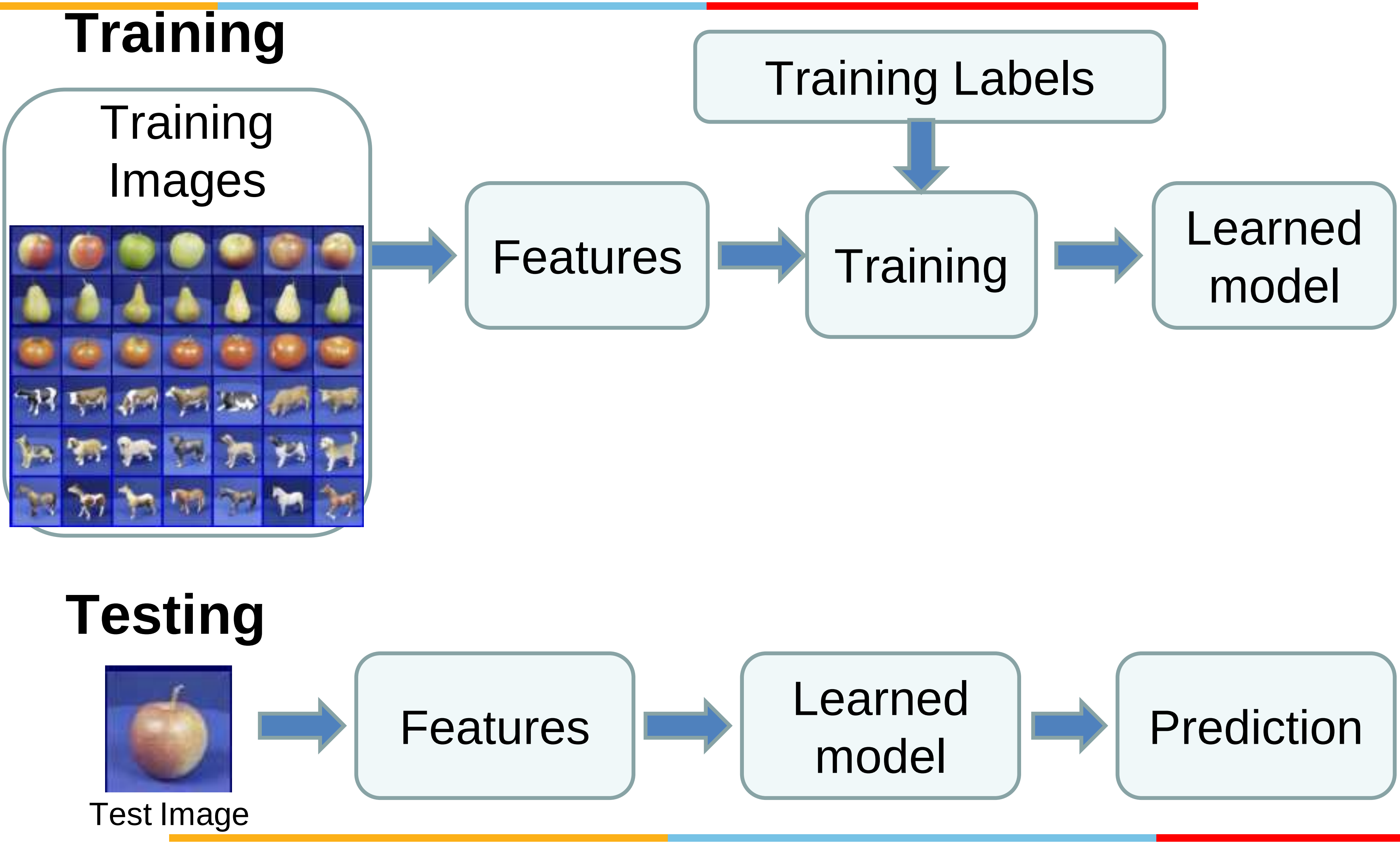
What is the temperature going to be tomorrow?



Will it be Cold or Hot tomorrow?



Classification



Slide credit: D. Hoiem and L. Lazebnik

Classification - Examples

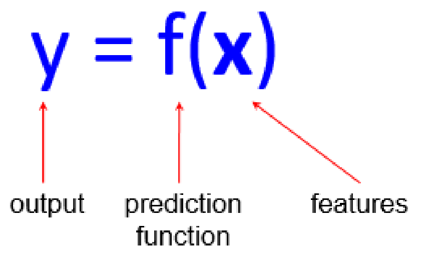
Objective: Employability Prediction

Features / Attributes / Predictors

- ✓ CGPA
- ✓ Communication Skills
- ✓ Aptitude
- ✓ Programming Skills

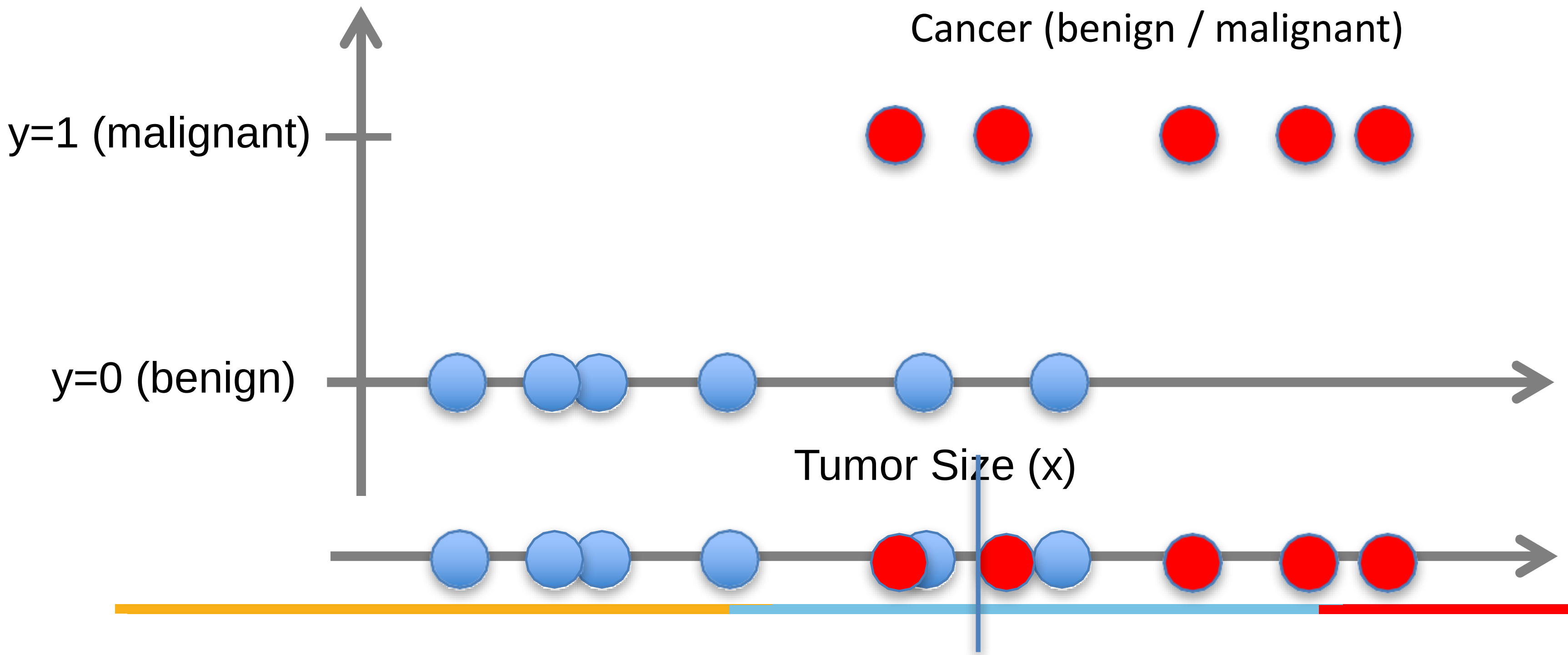
S.No.	CGPA	Communication Skills	Aptitude	Programming Skills	Job Offered?
1	9.1	Average	Good	Excellent	Yes
2	8.4	Good	Good	Good	Yes
3	8.3	Poor	Average	Average	No
4	7.1	Average	Good	Average	No
5	8.2	Good	Excellent	Excellent	No

Classification



- Given $(x^{[1]}, y^{[1]}), (x^{[2]}, y^{[2]}), \ldots, (x^{[n]}, y^{[n]})$
- Learn a function $f(x)$ to predict y given x
  - y is categorical

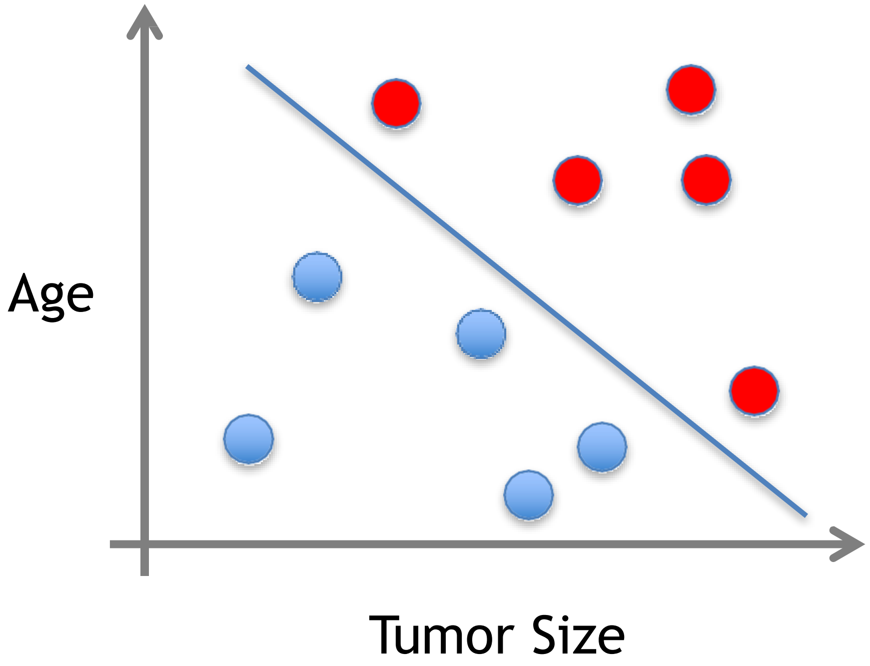
Learnt classifier
If x>T, malignant else benign



Classification

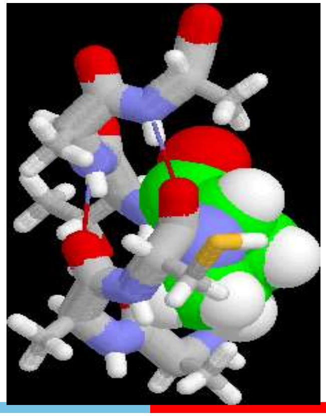
- x can be multi-dimensional
  - Each dimension corresponds to an attribute

Increasing Feature Dimension



Examples of Classification Task

- Fraud Detection
- Direct Marketing
- Churn prediction for telephone customers
- Email Spam detection
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc.
- Predicting tumour cells as benign or malignant



Regression- Example

Objective : Predicting price of a used car

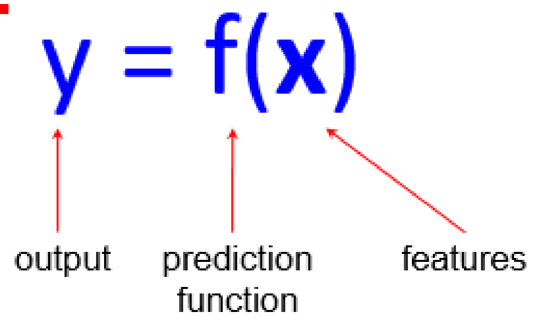
Features / Attributes / Predictors

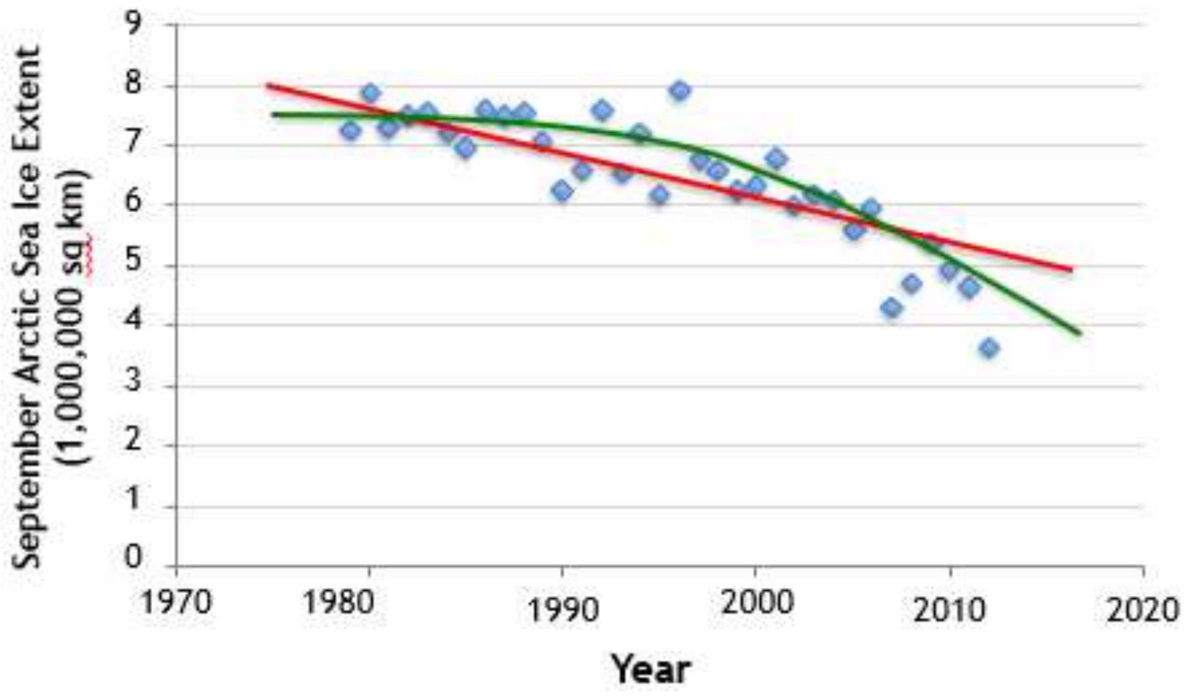
- ✓ Brand
- ✓ Year (Mfg)
- ✓ Engine Capacity
- ✓ Mileage
- ✓ Distance travelled
- ✓ Cab?

S.No	Brand	Year (Mfg)	Engine Capacity	Mileage	Distance travelled	Cab?	Price (in Rs.)
1.	Honda City ZX	2008	1100	10.5	45000	N	3,50,000
2							
3							
4							

Regression

- Given $(x^{[1]}, y^{[1]}), (x^{[2]}, y^{[2]}), \ldots, (x^{[n]}, y^{[n]})$
- Learn a function $f(x)$ to predict y given x





Examples of Regression Task

- Predicting house prices
- Forecasting sales figures
- Estimating patient recovery times
- Predicting tomorrow’s weather

Unsupervised learning

Unsupervised Learning – Clustering & Association

Clustering- Examples

Objective: Market Segmentation Study

Features / Attributes / Predictors

- ✓ Family income
- ✓ # of visits in a month
- ✓ Average money spent in a month
- ✓ Zip code

Customers for a retailer may fall into

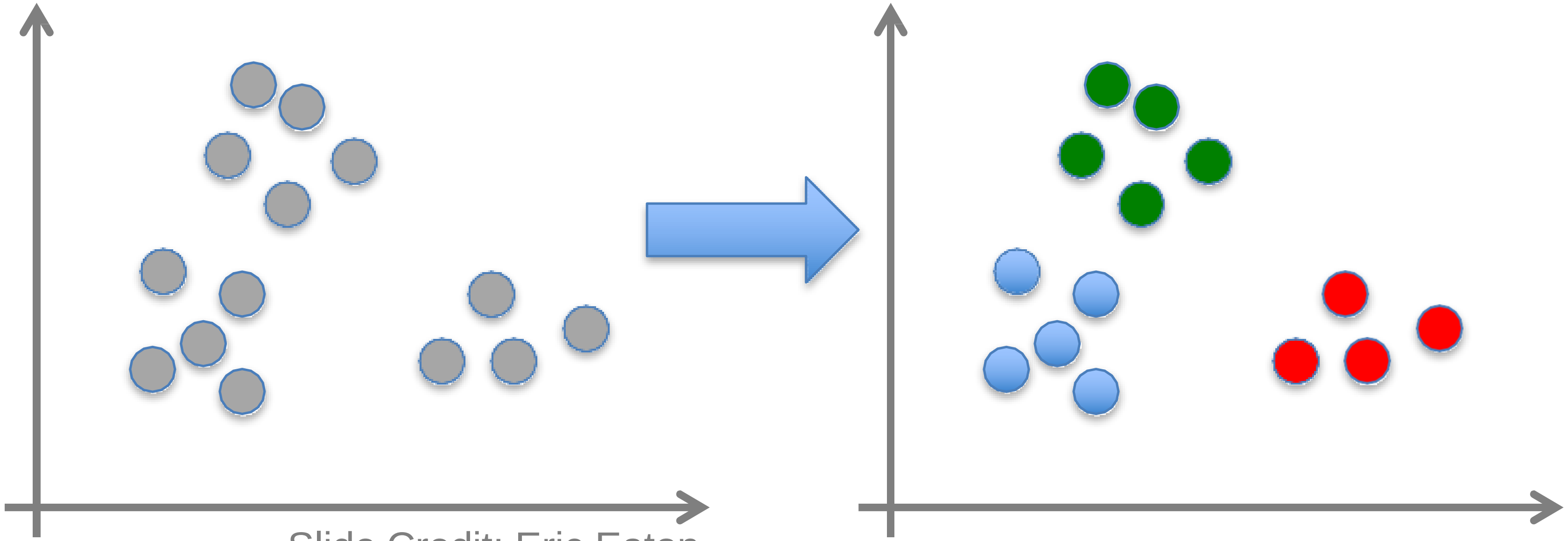
- ✓ two groups say big spenders and low spenders
- ✓ three groups say big spenders, medium spenders and low spenders
- ✓ Four groups, ...

S.No.	Zip Code	Family Income	# of visits in a month	Average Money Spent in a month
1	500078	11,50,000	4	8,000

Unsupervised Learning

GOAL : Intra cluster distances are minimized and inter cluster distances are maximized

- Given $x^{[1]}, x^{[2]}, ..., x^{[n]}$ (without labels)
- Output hidden structure behind the x's
  - e.g., clustering



Slide Credit: Eric Eaton

Unsupervised Learning

Applications

- Personalized recommendation system
- Targeted marketing
- Spam Filters
- Content Management – News hosted in Web
- Campaigning

Techniques

Clustering

- k-Means
- Hierarchical Cluster Analysis
- Expectation Maximization

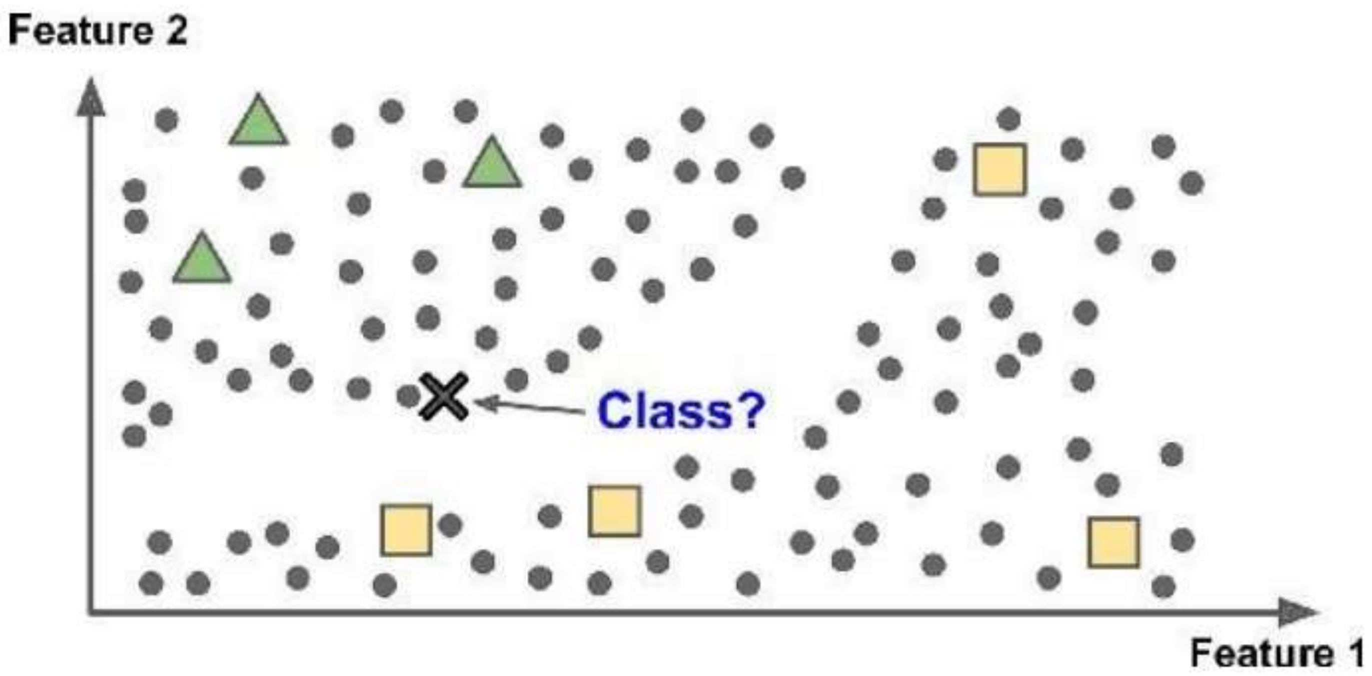
Visualization and dimensionality reduction

- Principal Component Analysis (PCA)
- Kernel PCA
- Locally-Linear Embedding (LLE)
- t-distributed Stochastic Neighbor Embedding (t-SNE)

Semi supervised Learning

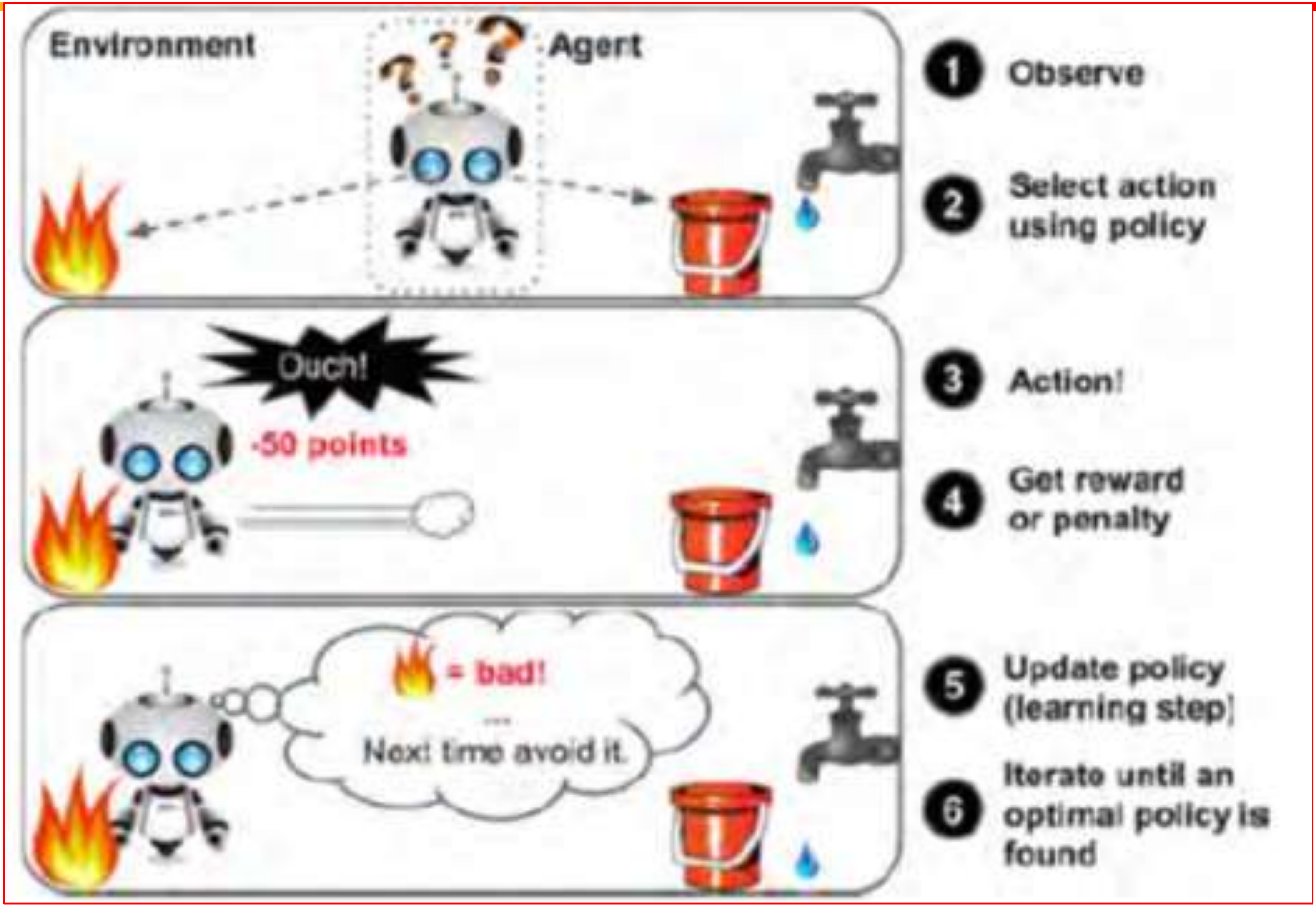
Semi supervised Learning

- Partially labelled data – some labelled data and a lot of unlabelled data
- Combines unsupervised and supervised learning algorithms
- Photo hosting service, e.g., google photos



Reinforcement Learning

Reinforcement Learning



Types of Learning

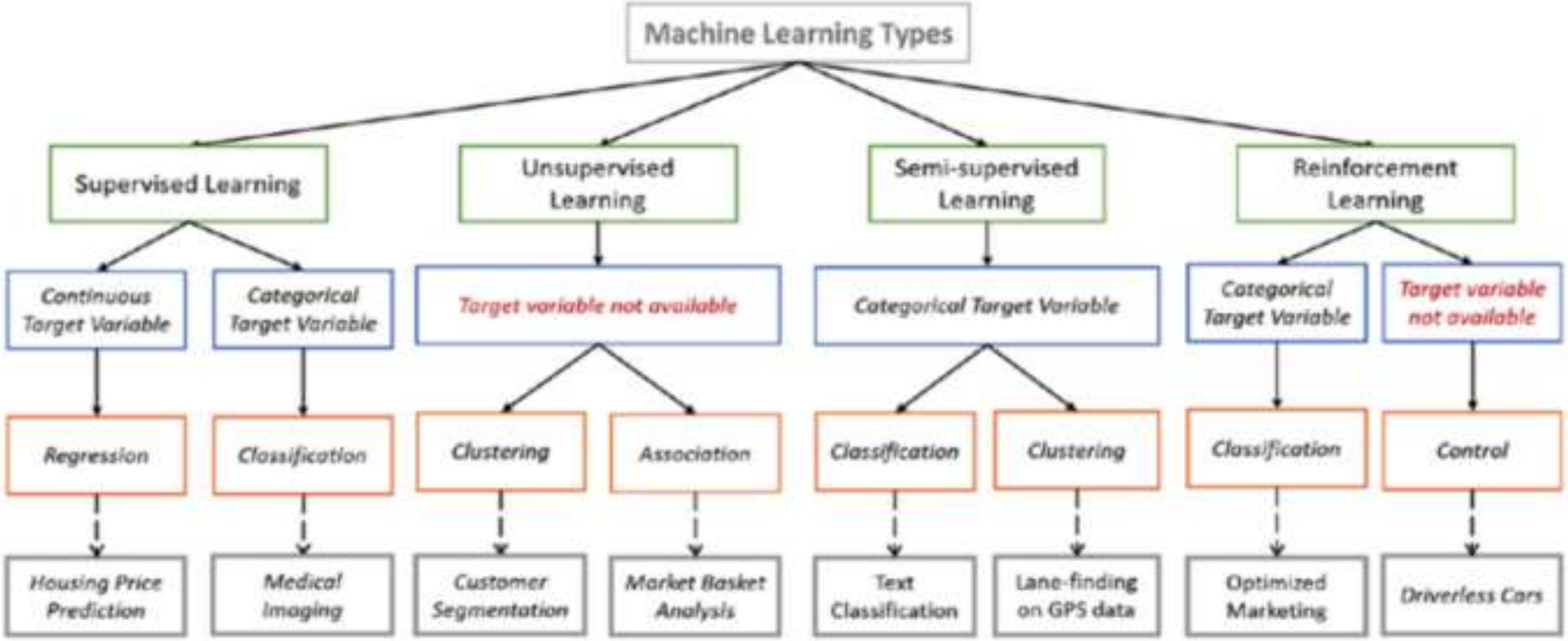
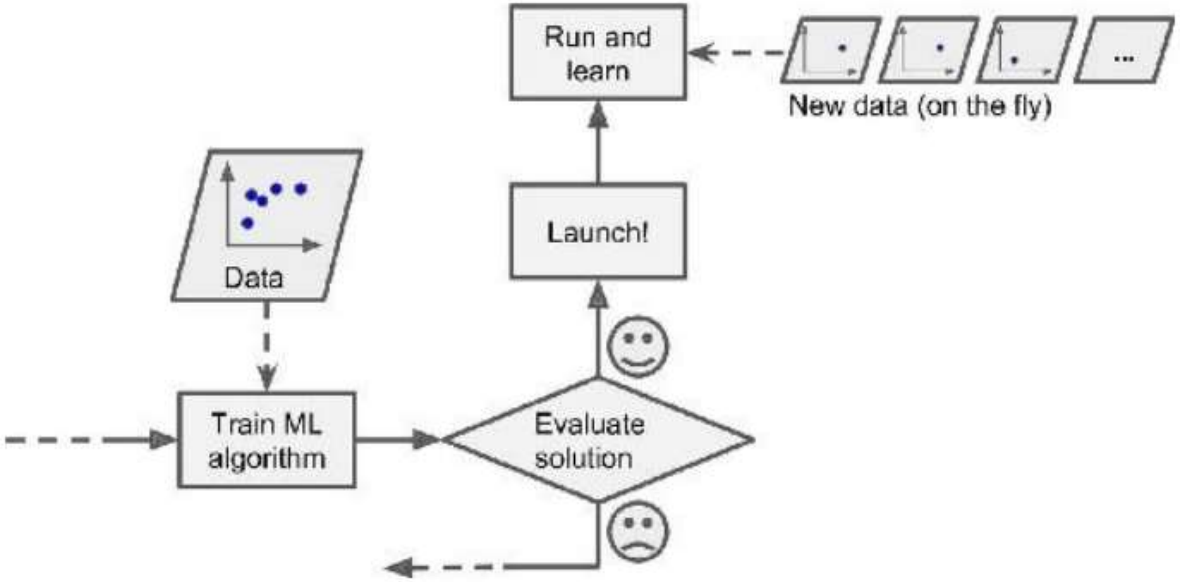
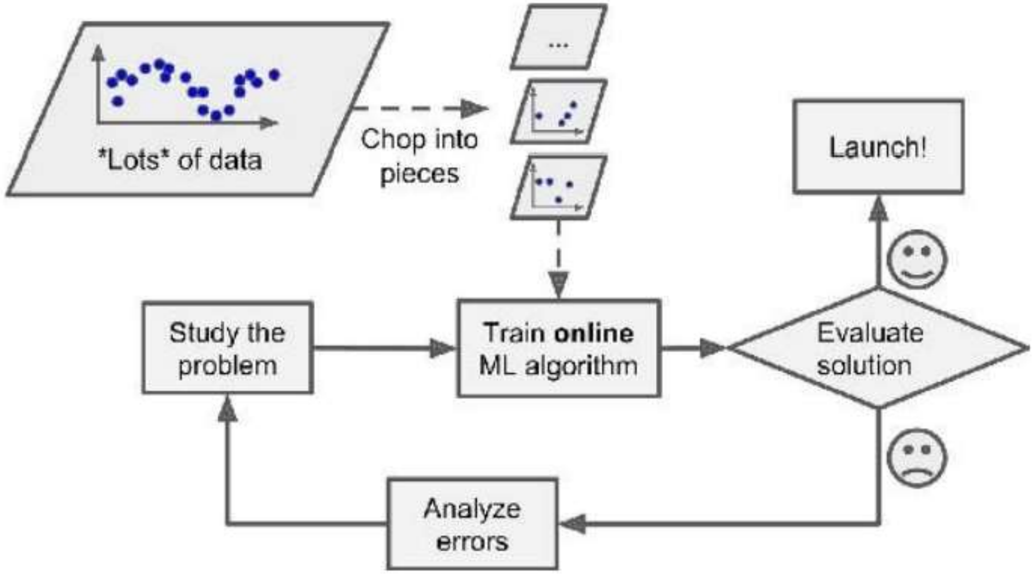


Image courtesy: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

Types: Based on how training data is used

- Batch learning: Uses all available data at a time during training
- Mini Batch learning: Uses a subset of available at a time during training
- Online (incremental) learning: a model is trained and launched into production, and then it keeps learning as new data comes in

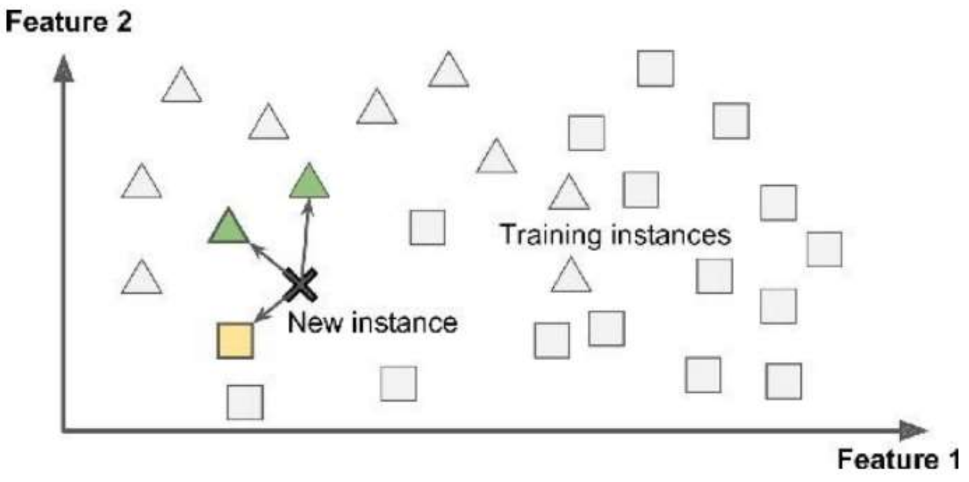




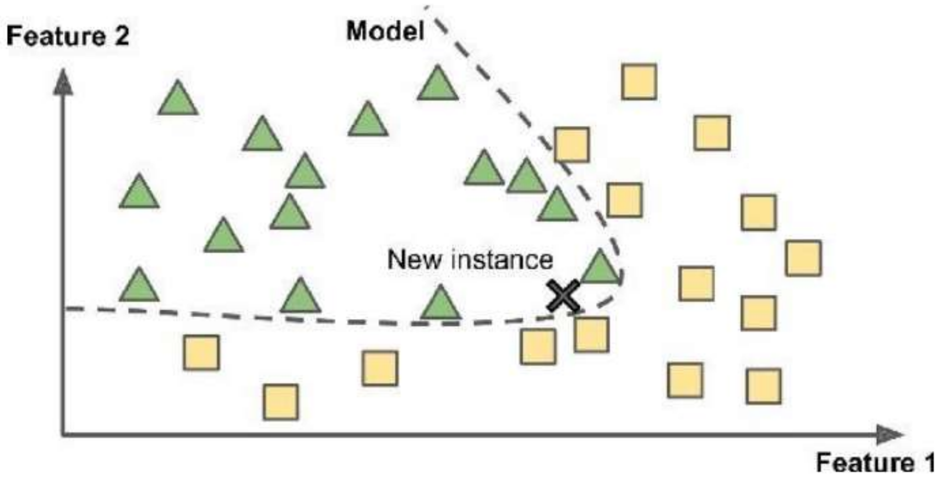
Types: Based on how training data is used

- Instance Based Learning: Compare new data points to known data points
- Model Based learning : Detect patterns in the training data and build a predictive model

Instance Based



Model Based



ML workflow

ML workflow

1. Should I use ML on this problem?
   – Is there a pattern to detect?
   – Can I solve it analytically?
   – Do I have data?
2. Gather and organize data.
3. Preprocessing, cleaning, visualizing.
4. Choosing a model, loss, regularization, ...
5. Optimization
6. Hyperparameter search
7. Analyze performance and mistakes, and iterate back to step 5 (or 3)

Example: Marks prediction

- Data: survey, marks from previous years
- Process the data
  - training set; test set
  - representation of input features; output
- Choose form of model: linear regression
- system’s performance evaluation: objective function
- optimize performance by setting appropriate parameters: Optimization
- Evaluate on test set: generalization

Few Terminologies

(To interpret the jargons in the prescribed text book)

Terminologies

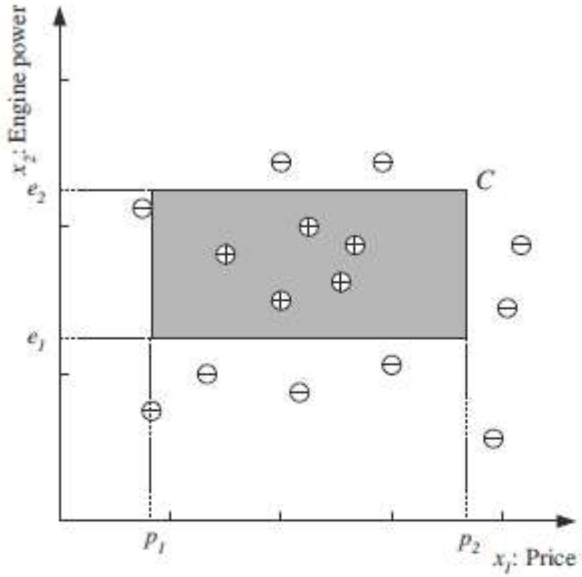
Amount taken	Period	Credit Score	Defaulter
40 lakhs	5 years	1000	No
10 Lakhs	5 months	550	YES
80 Lakhs	3 years	950	No
20 Lakhs	4 years	1500	No

- **Training example.** An example of the form $\langle \mathbf{x}, f(\mathbf{x}) \rangle$.
- **Target function (target concept).** The true function f.
- **Hypothesis.** A proposed function h believed to be similar to f.
- **Concept.** A boolean function. Examples for which $f(\mathbf{x}) = 1$ are called **positive examples** or **positive instances** of the concept. Examples for which $f(\mathbf{x}) = 0$ are called **negative examples** or **negative instances.**
- **Classifier.** A discrete-valued function. The possible values $f(\mathbf{x}) \in \{1, \ldots, K\}$ are called the **classes** or **class labels**.
- **Hypothesis Space.** The space of all hypotheses that can, in principle, be output by a learning algorithm.
- **Version Space.** The space of all hypotheses in the hypothesis space that have not yet been ruled out by a training example.

Hypothesis space

C=consistent hypothesis, S=specific hypothesis, G=Most general Hypothesis

Version space = Any h between S and G is a version space



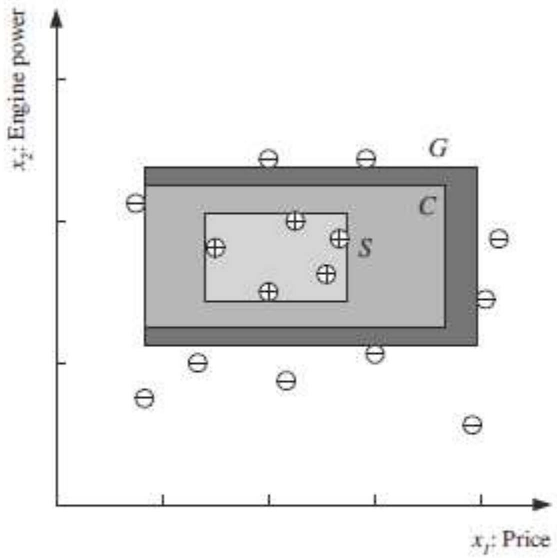


Figure 2.4 *S* is the most specific and *G* is the most general hypothesis.

Example : Credit card Processing by bank

- **Previous customer data**: salary, years in residence, outstanding loans, did bank make money on that customer etc.
- **Input** : X (customer data), **Output** : Y (yes/no decision)
- **Data set D** of input-output examples: $(x^{[1]}, y^{[1]})$, $(x^{[2]}, y^{[2]})$, ..., $(x^{[n]}, y^{[n]})$
- **Target function** $f: X \rightarrow Y$ (ideal formula for credit approval **but unknown**) → $y^{[n]} = f(x^{[n]})$
- **Learning algorithm** that uses the data set D to pick a formula g
- **Hypothesis** $g: X \rightarrow Y$ that approximates f.
- **Hypothesis set (hypothesis space)** : LA chooses g from a set of candidate formulas under consideration, which is called hypothesis set H. e.g. H could be the set of all linear formulas from which the algorithm would choose the best linear fit to the data
- **Generalization:** When a new customer applies for credit, bank will base its decision on g (the hypothesis that the learning algorithm produced), not on f (ideal target function which remains unknown).
  - **Correctness of the decision depends on how good *g* faithfully replicates *f***

Inductive Learning Hypothesis

Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples

Inductive learning or "**Prediction**":

- **Given** examples of a function *(X, F(X))*
- **Predict** function *F(X)* for new examples *X*

- Classification
  F(X) = Discrete

- Regression
  F(X) = Continuous

- Probability estimation
  F(X) = Probability*(X):*

Inductive Learning Hypothesis

- Target Concept
- Discrete : $f(x) \in \{Yes, No, Maybe\}$ Classification
- Continuous : $f(x) \in [20\text{-}100]$ Regression
- Probability Estimation : $f(x) \in [0\text{-}1]$

Sky	AirTemp	Humidity	Wind	Water	Forecast	*EnjoySport?*
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Inductive Learning Hypothesis

- Target Concept
- Discrete : $f(x) \in \{Yes, No, Maybe\}$ Classification
- Continuous : $f(x) \in [20-100]$ Regression
- Probability Estimation : $f(x) \in [0-1]$

Sky	AirTemp	Altitude	Wind	Water	Forecast	Humidity
Sunny	Warm	Normal	Strong	Warm	Same	60
Sunny	Warm	High	Strong	Warm	Same	75
Rainy	Cold	High	Strong	Warm	Change	70
Sunny	Warm	High	Strong	Cool	Change	45

Inductive Learning Hypothesis

- Target Concept
- Discrete : $f(x) \in \{Yes, No, Maybe\}$ Classification
- Continuous : $f(x) \in [20\text{-}100]$ Regression
- Probability Estimation : $f(x) \in [0\text{-}1]$

Sky	AirTemp	Humidity	Wind	Water	Forecast	*P(EnjoySport =Yes)*
Sunny	Warm	Normal	Strong	Warm	Same	0.95
Sunny	Warm	High	Strong	Warm	Same	0.7
Rainy	Cold	High	Strong	Warm	Change	0.5
Sunny	Warm	High	Strong	Cool	Change	0.6

Hypothesis

Example	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- **One possible hypothesis** (?, ***Cold,*** High, ?, ?, ?)
- The most **general hypothesis-**that every day is a positive example-(?, ?, ?, ?, ?, ?)
- The most **specific possible hypothesis-**that no day is a positive example **(ϕ, ϕ, ϕ, ϕ, ϕ, ϕ)**

References

- Chapter 1,2 – Machine Learning, Tom Mitchell
- Chapter 1, 2 – Introduction to Machine Learning, 2nd edition, Ethem Alpaydin
- Chapter 1 - Pattern Recognition & Machine Learning Christopher M. Bhishop
- http://www.cs.princeton.edu/courses/archive/spr08/cos511/ [Web]
- https://www.softwaretestinghelp.com/machine-learning-tools/

Thank you !