MACHINE LEARNING

SLIDES ADAPTED (AND EXTENDED) FROM ETHEM ALPAYDIN © THE MIT PRESS, 2004

MODULE I - CONTENTS

Introduction to Machine Learning, Examples of Machine Learning applications-

Learning associations, Classification, Regression, Unsupervised Learning, Reinforcement Learning. Supervised learning- Input representation, Hypothesis class, Version space, Vapnik-Chervonenkis (VC) Dimension

WHAT IS LEARNING?

- Many applications do not have an algorithm but do have example data
- Any process by which a system improves performance from **experience**
- Learning program:
  - A computer program which learns from experience
  - Also referred to as a learner.

LEARNING

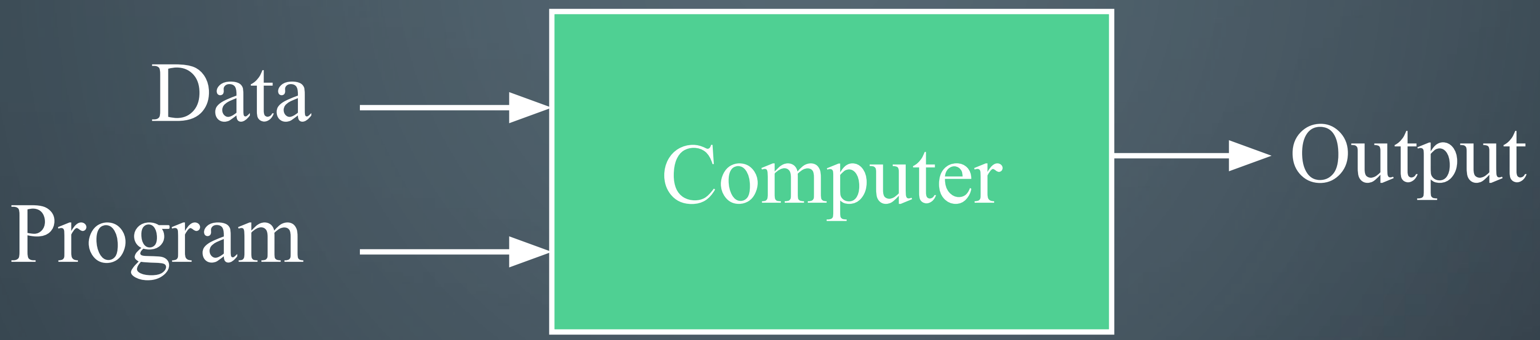
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)

- Data is cheap and abundant; Knowledge is expensive and scarce

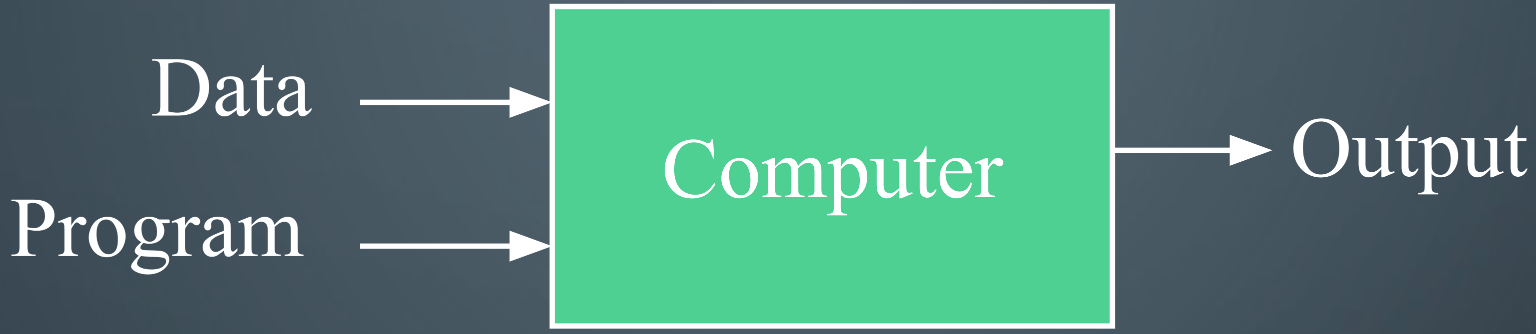
- Build a model that is good and useful approximation to the data

  Example in retail :Customer transactions to consumer behavior

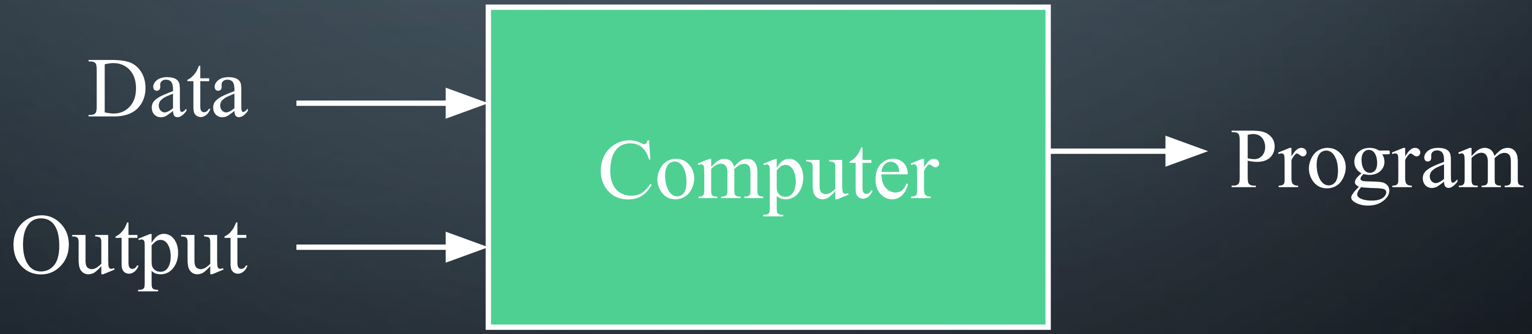
Traditional Programming



Traditional Programming



Machine Learning



MACHINE LEARNING

- Arthur Samuel coined the term ‘Machine Learning’
- Definition:

A field of artificial intelligence (AI) where the systems will be given the ability to learn things automatically and make decisions with very less human intervention

Other Definitions:

1. ML is programming computers to optimize a performance criterion using example data or past experience
    - Role of statistics : Inference from a sample
    - Role of CS: Efficient algorithms to
        - Solve the optimization problem

MACHINE LEARNING

2. Concerned with the question; ‘how to construct computer programs that automatically improve with experience?’

A computer program is said to

learn from experience E

w.r.t. some class of tasks T and

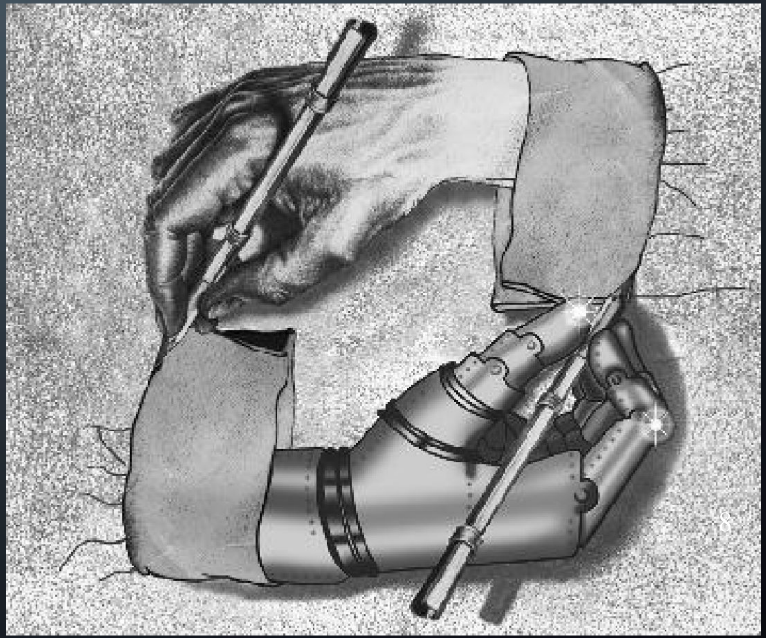
performance measure P,

if

its performance at tasks T

as measured by P,

improves with experience E



EXAMPLES

1. Handwriting recognition learning problem
   - ✔ Task T: Recognising and classifying handwritten words within images
   - ✔ Performance P: Percent of words correctly classified
   - ✔ Training experience E: A dataset of handwritten words with given classifications
2. A robot driving learning problem
   - ✔ Task T: Driving on highways using vision sensors
   - ✔ Performance measure P: Average distance travelled before an error
   - ✔ Training experience: A sequence of images and steering commands recorded while observing a human driver
3. A chess learning problem
   - ✔ Task T: Playing chess
   - ✔ Performance measure P: Percent of games won against opponents
   - ✔ Training experience E: Playing practice games against itself

ELEMENTS OF MACHINE LEARNING

- Data (Storage)
  - Storing and retrieving huge amounts of data
  - Training data : Specific examples to learn from
  - Test data : (new) Specific examples to access performance

- Abstraction (Models)
  - Extracting knowledge about stored data
  - Creating general concepts about the data as a whole
  - Theoretical assumptions about the task/domain
  - Parameters that can be inferred from data –
    - application of known models and creation of new models

- Generalization
  - Generalize from specific examples
  - Discover those properties of the data that will be most relevant to future tasks
  - Based on statistical inference: on tasks that are similar, but not identical
- Evaluation
  - Giving feedback to the user to measure the utility of the learned knowledge
  - Utilized to effect improvements in the whole learning process

BIG DATA

- Widespread use of personal computers and wireless communication leads to “big data”
- We are both producers and consumers of data
- Data is not random, it has structure, e.g., customer behavior
- We need “big theory” to extract that structure from data for
  - Understanding the process
  - Making predictions for the future
- Data Mining

APPLICATIONS OF MACHINE LEARNING

- Retail : Market based analysis, Customer relationship management(CRM)
- Finance: Credit scoring ,fraud detection
- Manufacturing: Control ,robotics(optimization),troubleshooting
- Medicine: Medical diagnosis
- Telecommn: QoS, spam filters, intrusion detection
- Bioinformatics : Motifs, alignment
- Webmining : Search engines
- AI : system can learn and adapt to changes
- Robotics, solution to speech recognition and vision problem
- Computer controlled vehicles

UNDERSTANDING DATA

- **Units of observation**: the smallest entity with measured properties of interest for a study
- **Examples**: an instance of the unit of observation for which properties have been recorded
- **Features (attribute/ variable):** a recorded property or a characteristic of examples

- Eg.: **Cancer detection-**
  - **Units of observation : the patients**
  - **Examples: members of a sample of cancer patients**
  - **Attributes of the patients (the following may be chosen as the features):**
    - **gender**
    - **age**
    - **blood pressure**

- Different forms of data
  - Numeric (year, price, mileage)
  - Categorical or nominal (model, color, transmission)
  - Ordinal (sizes- S, M, L, XL)

features

year	model	price	mileage	color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	10926	Gray	AUTO
2011	SEL	19995	7351	Silver	AUTO
2011	SEL	17809	11613	Gray	AUTO
2012	SE	17500	8367	White	MANUAL
2010	SEL	17495	25125	Silver	AUTO
2011	SEL	17000	27393	Blue	AUTO
2010	SEL	16995	21026	Silver	AUTO
2011	SES	16995	32655	Silver	AUTO

examples

GENERAL CLASSES OF MACHINE LEARNING PROBLEM

1. <u>LEARNING ASSOCIATIONS</u>

- **Association Rule Learning-** method for discovering interesting relations, called "association rules" between variables
- Example : Shopping and basket analysis
  - P(Y|X) probability that somebody who buys X also buys Y where X and Y are products/services
  - We learn Association Rule : P(chips | juice) =0.7
  - Use this association rule like this:

    Target customers who bought X, but not Y. Try to convince them to but Y

GENERAL CLASSES OF MACHINE LEARNING PROBLEM (CONTD.)

2. <u>CLASSIFICATION</u>

- Identifying to which of a set of categories a new observation belongs
- Based on - a training set containing observations whose category membership is known
- **Discriminant**: a rule or a function that is used to assign labels to new observations

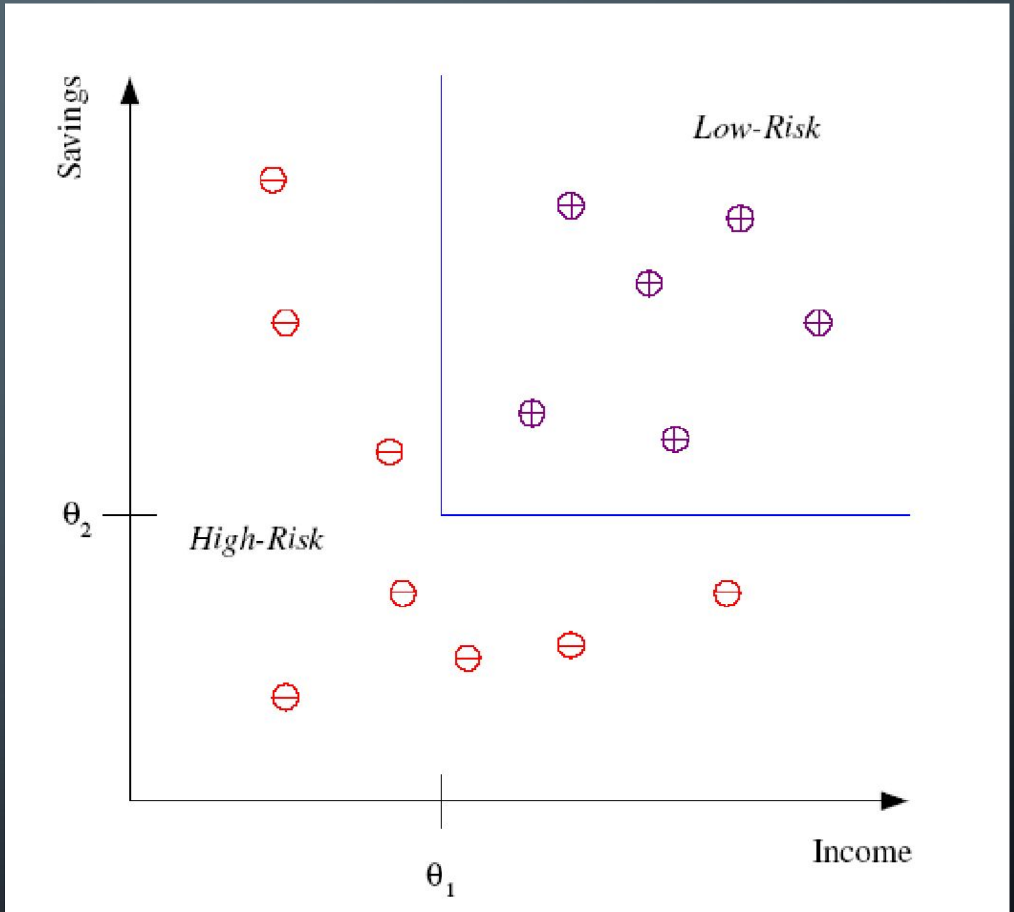
CLASSIFICATION

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

  Discriminant: If income $> \theta_1$ and

  $savings > \theta_2$ $then\ Low\ risk$

  $else\ High\ risk$
- Main application : Prediction

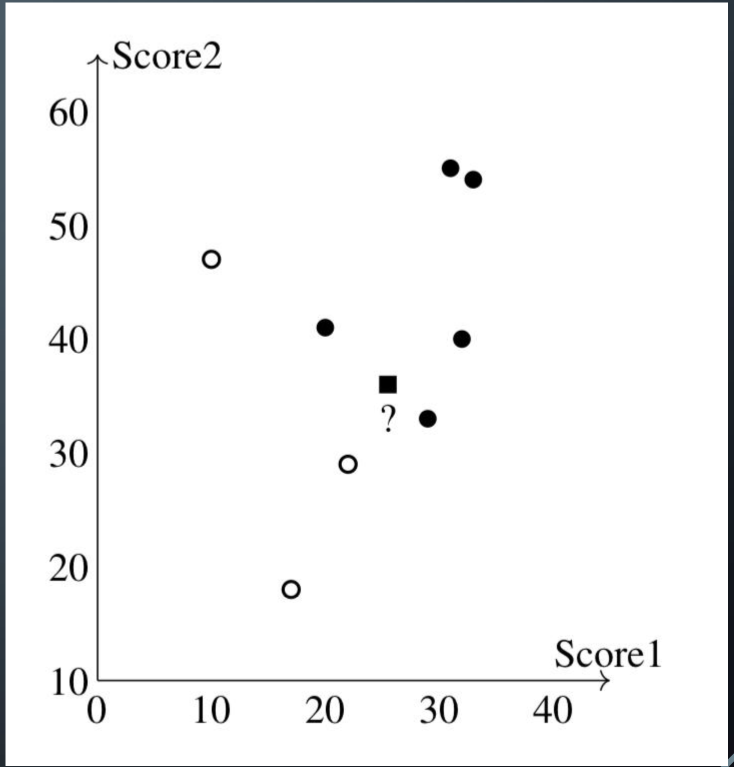


CLASSIFICATION

Example: Grading Result (Pass/ Fail)

Score1	29	22	10	31	17	33	32	20
Score2	43	29	47	55	18	54	40	41
Result	Pass	Fail	Fail	Pass	Fail	Pass	Pass	Pass

If we have some new data, say "Score1 = 25" and "Score2 = 36", what value should be assigned to "Result"?



Discriminant:

IF Score1 + Score2 >= 60, THEN "Pass" ELSE "Fail".

or

IF Score1 >= 20 AND Score2 >= 40 THEN "Pass" ELSE "Fail"

Or consider the following rules with unspecified values forM; m1; m2

(and then by some method estimate their values)

IF Score1 + Score2 >= M, THEN "Pass" ELSE "Fail".

or

IF Score1 >= m1 AND Score2 >= m2 THEN "Pass" ELSE "Fail"

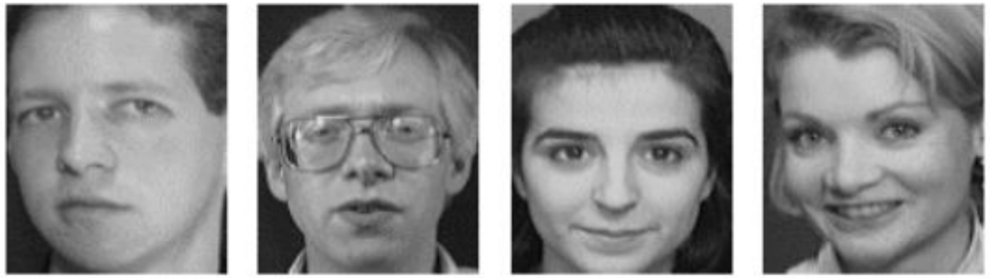
CLASSIFICATION-OTHER APPLICATIONS

- Face recognition
- Character recognition
- Speech recognition : Temporal dependency
  - Use of a dictionary or syntax of a language
  - Sensor fusion: Combine multiple modalities
- Gesture recognition
- Medical diagnosis
- Brainwave understanding

FACE RECOGNITION

Training examples of a person

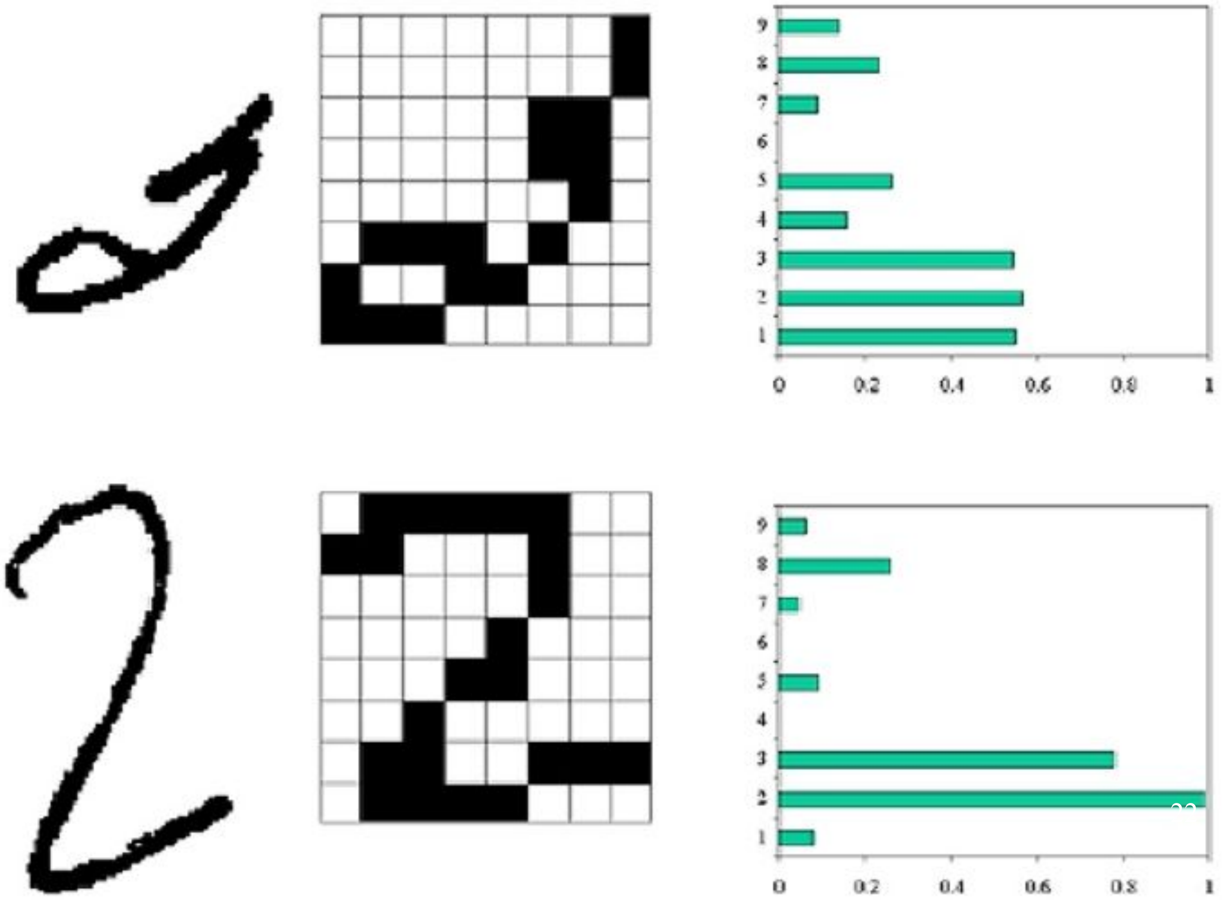
Test images



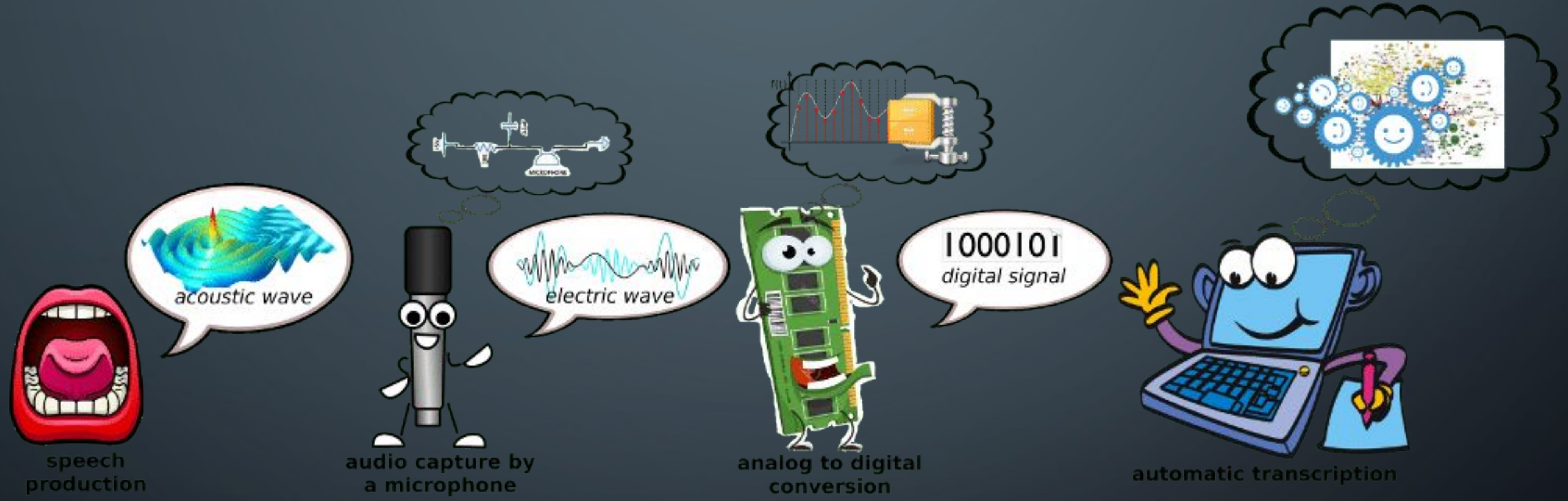
AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

CHARACTER RECOGNITION E.G.

Want to learn how to recognize characters, even if written in different ways by different people



SPEECH RECOGNITION

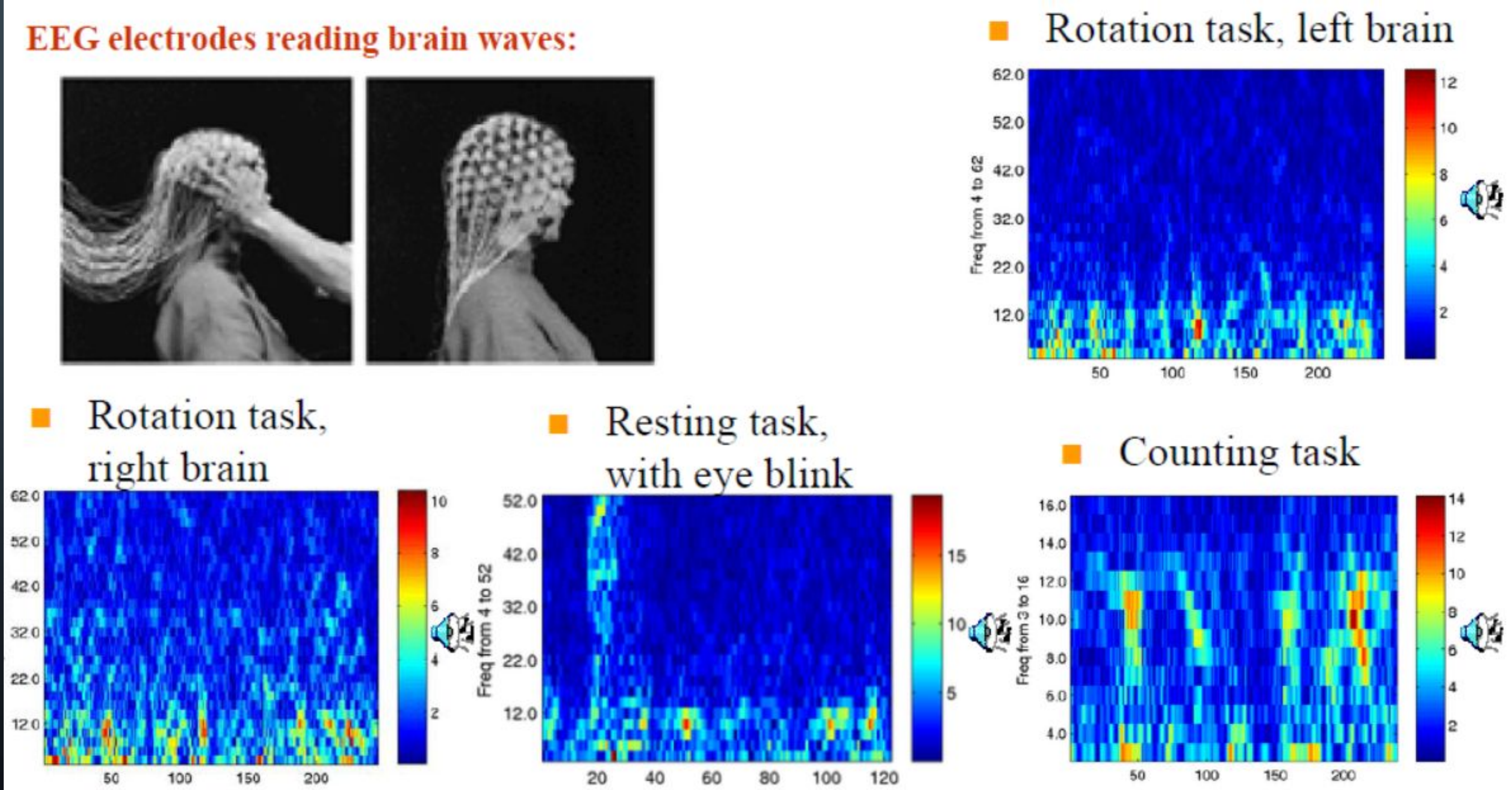


Img source:http://www.authot.com/

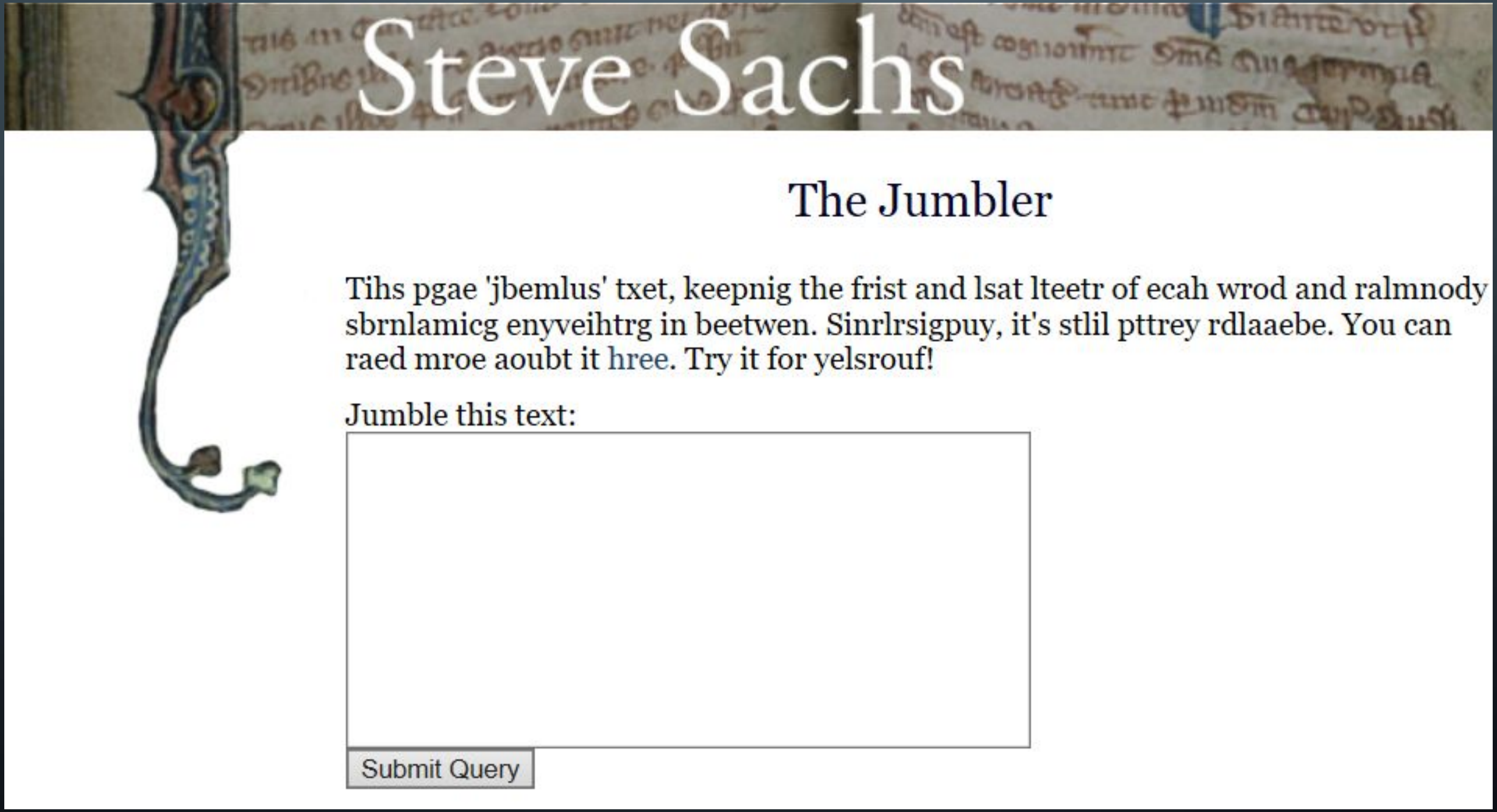
GESTURE RECOGNITION

MEDICAL DIAGNOSIS

BRAINWAVE UNDERSTANDING



READING TEXT



REGRESSION

- Predicting the value of a numeric variable based on observed values of the variable
  - Output variable(x): may be a number, such as an integer or a floating point value
  - Input variables(y): may be discrete or real-valued
  - Suppose we are required to estimate the price of a car -
    - aged 25 years
    - distance 53240 KM
    - weight 1200 pounds
  - General Approach: Find some mathematical relation between x and y

    $y = f(x,\theta)$

  - Optimize the parameters in the set such that the approximation error is minimized
  - The estimates of the values of the dependent variable y are as close as possible to the

Price	Age	Distance	Weight
(US$)	(years)	(KM)	(pounds)
13500	23	46986	1165
13750	23	72937	1165
13950	24	41711	1165
14950	26	48000	1165
13750	30	38500	1170
12950	32	61000	1170
16900	27	94612	1245
18600	30	75889	1245
21500	27	19700	1185
12950	23	71138	1105

REGRESSION

- The model may be $y = f(x, \theta)$

  Price = a0 + a1 × (Age) + a2 × (Distance) + a3 × (Weight)

  where x = (Age, Distance, Weight) denotes the set of input variables

  θ = (a0, a1, a2, a3) denotes the set of parameters of the model

Different regression models:

- Simple linear regression: There is only one continuous independent variable x

$$y = a + bx$$

- Multivariate linear regression: There are more than one independent variable, say $x_1 .. x_n$

$$y = a_0 + a_1x_1 + \ldots + a_nx_n$$

- Polynomial regression: There is only one continuous independent variable x and the assumed model is

$$y = a_0 + a_1x + \ldots + a_nx^n$$

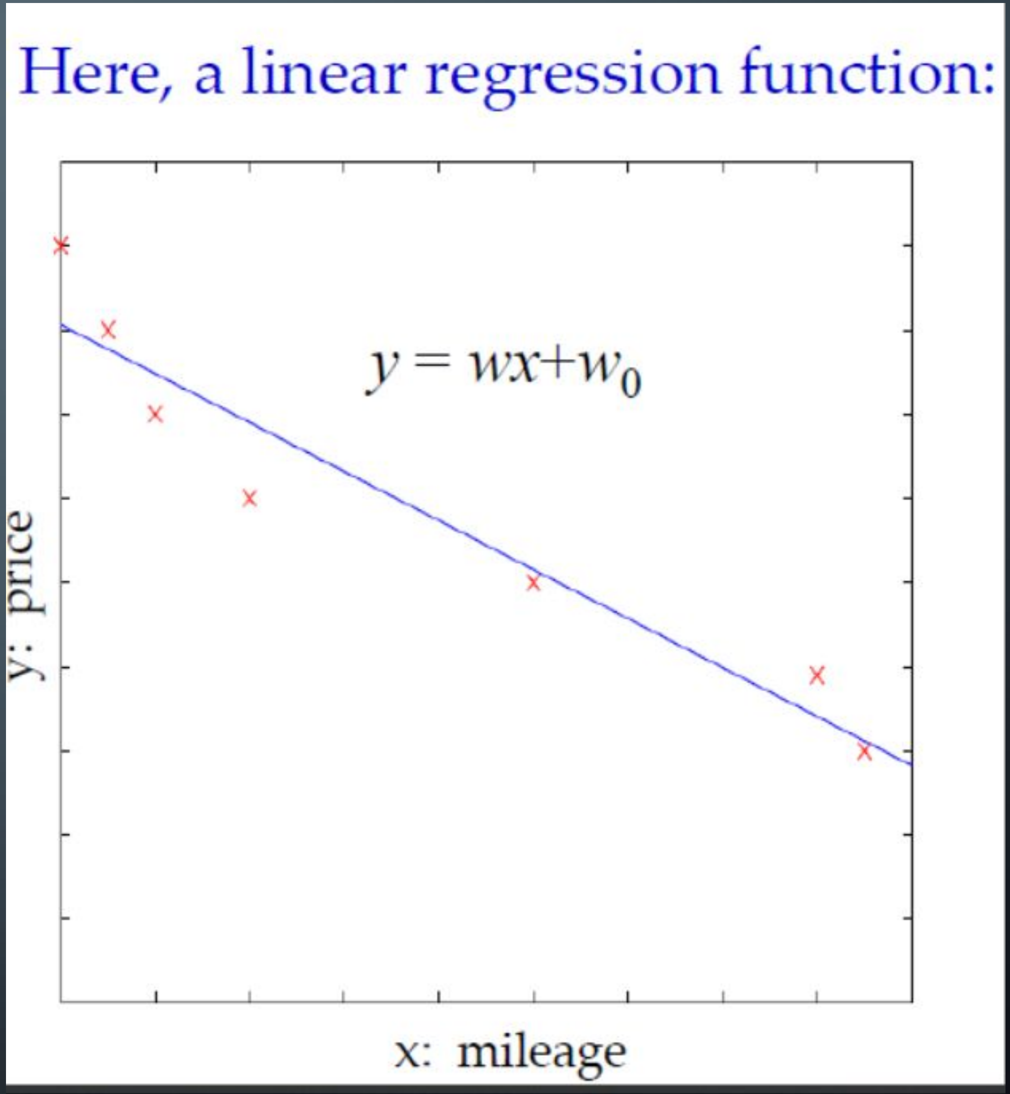
- Logistic regression: The dependent variable is binary

REGRESSION

- E.g. Predict price of a used car
- (Input) x : car attributes

  (output) y : Price
- Task: Lean the mapping from input to output
  - G() model
  - θ parameters that minimize the error in the approximation

$$y=g(x|\theta)$$



TYPES OF MACHINE LEARNING

- Learning associations
- Supervised learning
- Unsupervised learning
- Reinforcement learning

LEARNING ASSOCIATIONS

- Example : Shopping and basket analysis
- P(Y|X) probability that somebody who buys X also buys Y where X and Y are products/services
  - We learn Association Rule : P(chips | juice) =0.7
- Use this association rule like this:
  - Target customers who bought X, but not Y
    - Try to convince them to but Y

SUPERVISED LEARNING

- The machine learn from the training data that is labeled
- Prediction of future cases : Use the rule to predict the output for future inputs
- Knowledge extraction : The rule is easy to understand
- Compression : The rule is simpler than the data it explains
- Outlier detection : Exceptions that are not covered by the rule(e.g. To detect fraud)

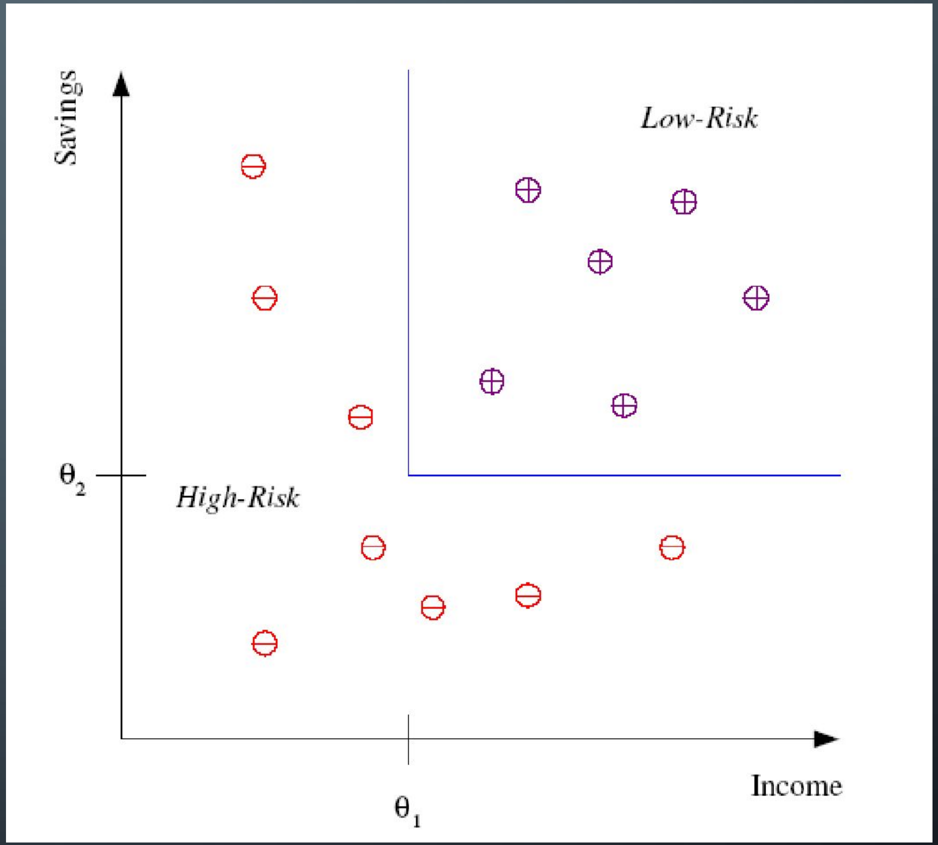
CLASSIFICATION

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

  Discriminant: If income $>\theta_1$ and

  $savings > \theta_2 \;\; then\ Low\ risk$

  $else\ High\ risk$
- Main application : Prediction

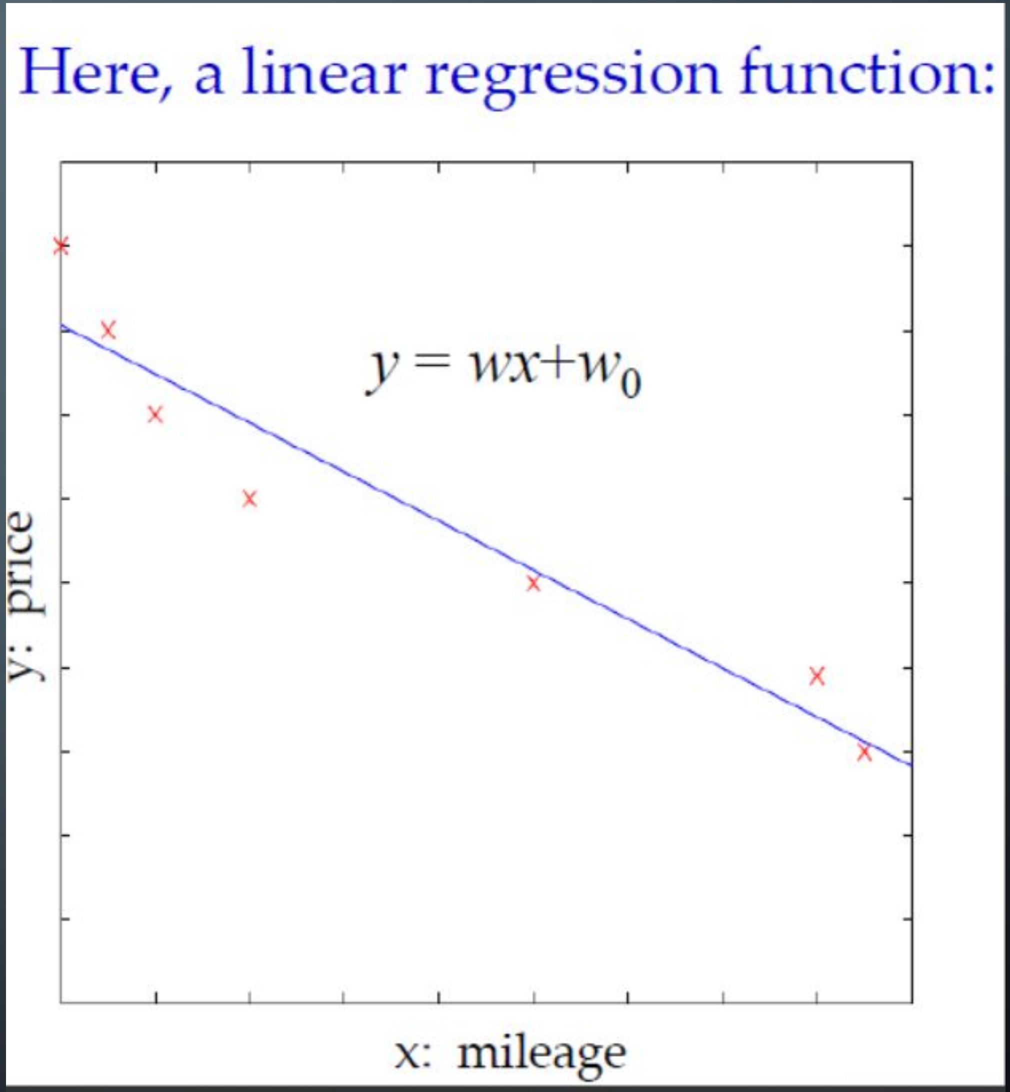


REGRESSION

- E.g. Predict price of a used car
- (Input) x : car attributes

  (output) y : Price
- Task: Lean the mapping from input to output
  - G() model
  - θ parameters that minimize the error in the approximation

$$y=g(x\,|\,\theta)$$



UNSUPERVISED LEARNING

- Finding regularities in data (Learning “what normally happens”)
- No mapping to outputs (i.e., we don’t know the “right” answer)
- Clustering : Grouping similar instances
- Example applications
  - E-commerce
  - Fraud detection in banking

REINFORCEMENT LEARNING

- Learning policy : A sequence of outputs/actions
- No supervised output, but feedback is used
- Credit assignment problem
- Example applications
  - Game playing

ML TYPES

SUPERVISED

- Classification
- Regression
- Ranking

UNSUPERVISED

- Clustering
- Association Mining
- Segmentation
- Dimension reduction

REINFORCEMENT

- Decision Process
- Reward System
- Recommendation system

SUPERVISED LEARNING

SOME GENERAL CONCEPTS

- **Input Representation**
- **Hypothesis**
- **Hypothesis Space**
- **Ordering of Hypotheses**
- **Version Space**

INPUT REPRESENTATION

EXAMPLE:

- Consider the **problem of assigning the label "family car" or "not family car" to cars.**
- Assumption: features –
  - **PRICE**
  - **ENGINE POWER**
- These **attributes or features constitute the input representation for the problem.**
- **Ignoring various other attributes like seating capacity or colour as irrelevant.**

SUPERVISED CLASSIFICATION

- Learning the Class C of a "family car" from examples
  - Prediction: Is car x a family car?
  - Knowledge extraction: What do people expect from a family car?
- Output: (labels)

  Positive (+) and negative (–) examples
- **Input representation**: (features)

SUPERVISED CLASSIFICATION

- Learning the Class C of a “family car” from examples
  - Prediction: Is car x a family car?
  - Knowledge extraction: What do people expect from a family car?
- Output: (labels)

  Positive (+) and negative (–) examples
- **Input representation**: (features)

  $x1$: price, $x2$: engine power

HYPOTHESIS SPACE

- We consider only "**binary classification**" problems:

  classification problems **with only two class labels**.
- The class labels are usually taken **as "1" and "0"**
- The label "**1**" may indicate "**True", or "Yes", or "Pass**", or any such label
- The label "**0**" may indicate "**False", or "No" or "Fail**", or any such label
- examples with class labels 1 are called "**positive examples**"
- examples with labels "0" are called "**negative examples**".

HYPOTHESIS SPACE

Definitions:

- **Hypothesis:**
  - ✔ In a binary classification problem, a hypothesis is a statement or a proposition purporting to explain a given set of facts or observations.
- **Hypothesis space:**
  - ✔ The hypothesis space for a binary classification problem is a set of hypotheses for the problem that might possibly be returned by it.

EXAMPLES - 1

- Consider the set of observations of a variable x with the associated class labels given in Table 2.1:

x	27	15	23	20	25	17	12	30	6	10
Class	1	0	1	1	1	0	0	1	0	0

Table 2.1: Sample data to illustrate the concept of hypotheses

Figure 2.1 shows the data plotted on the x-axis.

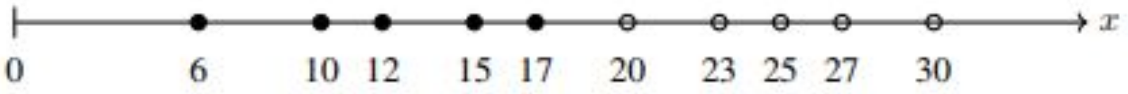


Figure 2.1: Data in Table 2.1 with hollow dots representing positive examples and solid dots representing negative examples

- hypothesis h’ : IF x >= 20 THEN “1” ELSE “0”
- h’ is consistent with the training examples
- another hypothesis h’’ : IF x <= 19 THEN “0” ELSE “1”

Figure 2.1 shows the data plotted on the x-axis.

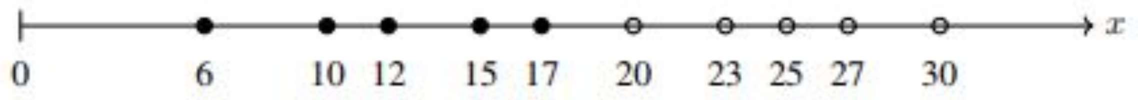


Figure 2.1: Data in Table 2.1 with hollow dots representing positive examples and solid dots representing negative examples

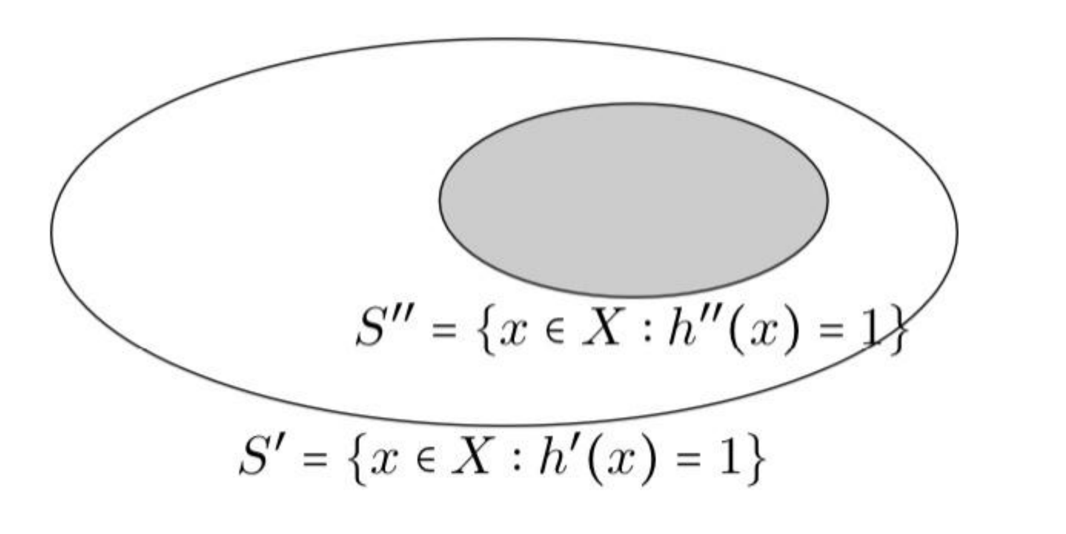
- The set of **hypotheses** can be defined using a parameter, say m, as given below:

  h_m : IF x >= m THEN “1” ELSE ”0”

- The set of all hypotheses obtained by assigning different values to m constitutes the **hypothesis space** H; that is,

  $H = \{h_m : m \text{ is a real number}\}$

ORDERING OF HYPOTHESES



- Let X be the set of all possible examples for a binary classification problem and let h’ and h’’ be two hypotheses for the problem.

- h’ is more general than h’’ if and only if for every x ε X, if x satisfies h” then x satisfies h’ also;
  - that is, if h”(x) = 1 then h’(x) = 1 also

- The relation “is more general than” defines a partial ordering relation in hypothesis space.

VERSION SPACE

Consider a binary classification problem:

- D be a set of training examples
- H a hypothesis space for the problem
- The version space for the problem with respect to the set D and the space H is:

  the set of hypotheses from H consistent with D; that is, it is the set

$$VS_{D,H} = \{h \varepsilon H : h(x) = c(x) \text{ for all } x \varepsilon D\}$$

Figure 2.1 shows the data plotted on the x-axis.

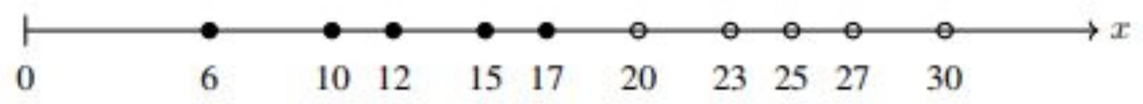


Figure 2.1: Data in Table 2.1 with hollow dots representing positive examples and solid dots representing negative examples

- The set of **hypotheses** can be defined using a parameter, say m, as given below:

  h_m : IF x >= m THEN "1" ELSE "0"

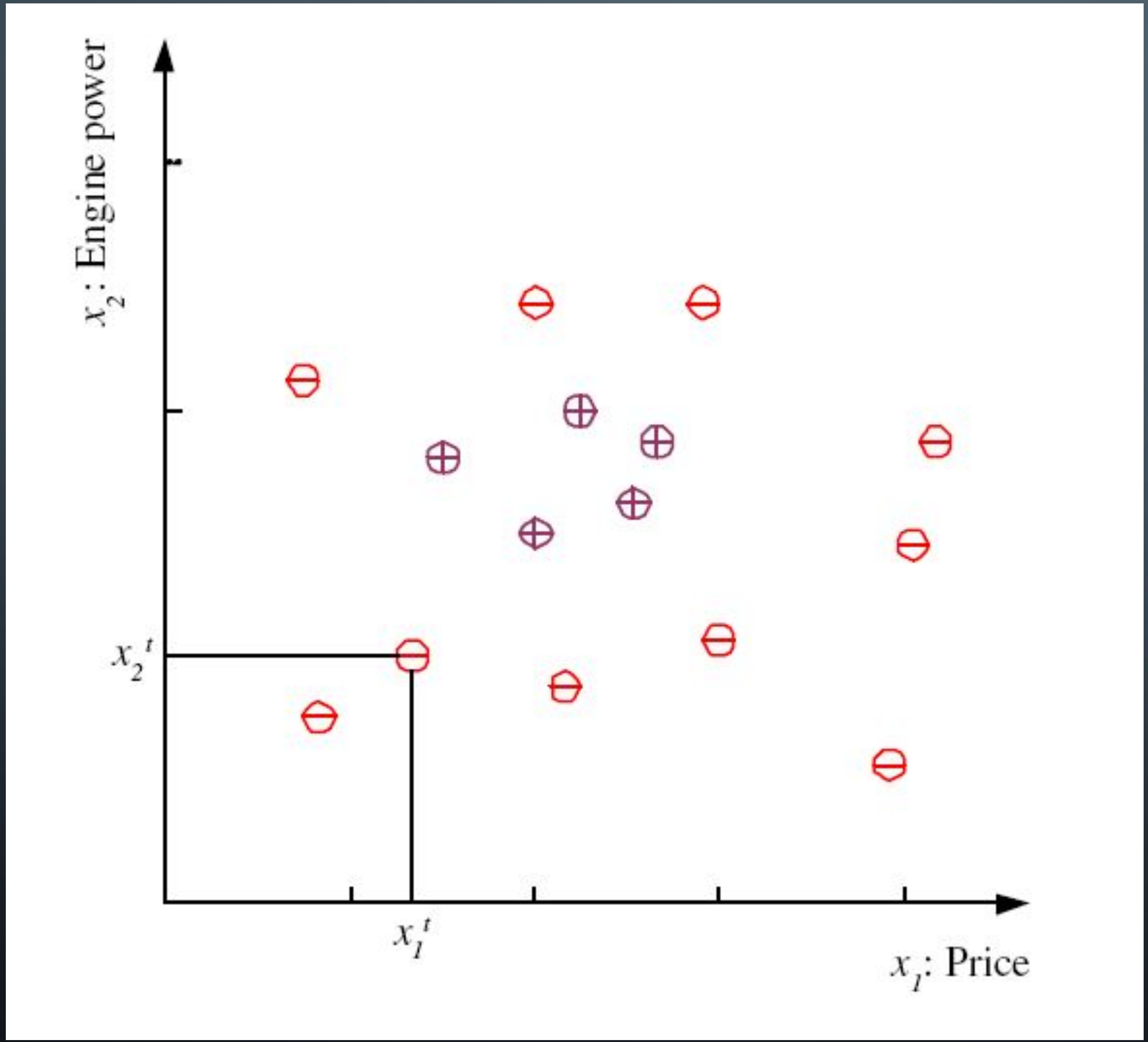
- The set of all hypotheses obtained by assigning different values to m constitutes the **hypothesis space** H; that is,

  $H = \{h_m : m \text{ is a real number}\}$

- The **version space** defined by the given data set and hypothesis space is

  $VS_{D,H} = \{h_m : 17 < m <= 20\}$

EXAMPLE 2: TRAINING SET X



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$r = \begin{cases} 1 \text{ if } \mathbf{x} \text{ is positive} \\ 0 \text{ if } \mathbf{x} \text{ is negative} \end{cases}$$

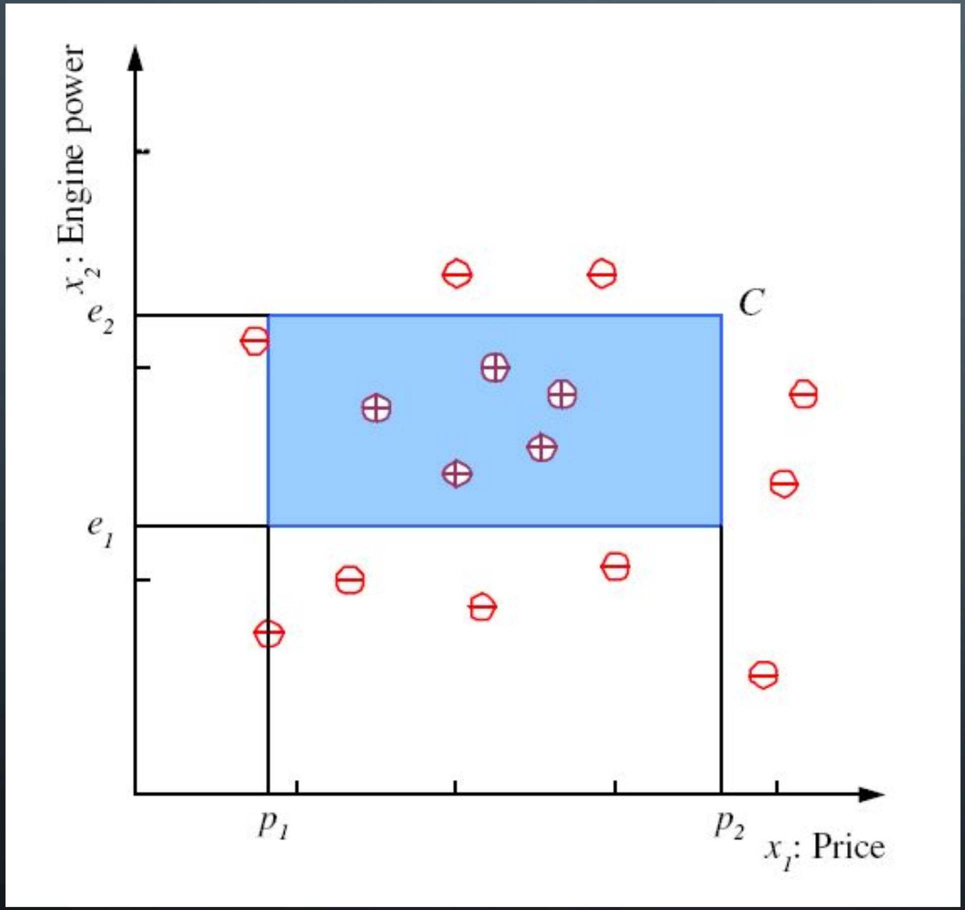
$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N$$

HYPOTHESIS CLASS H

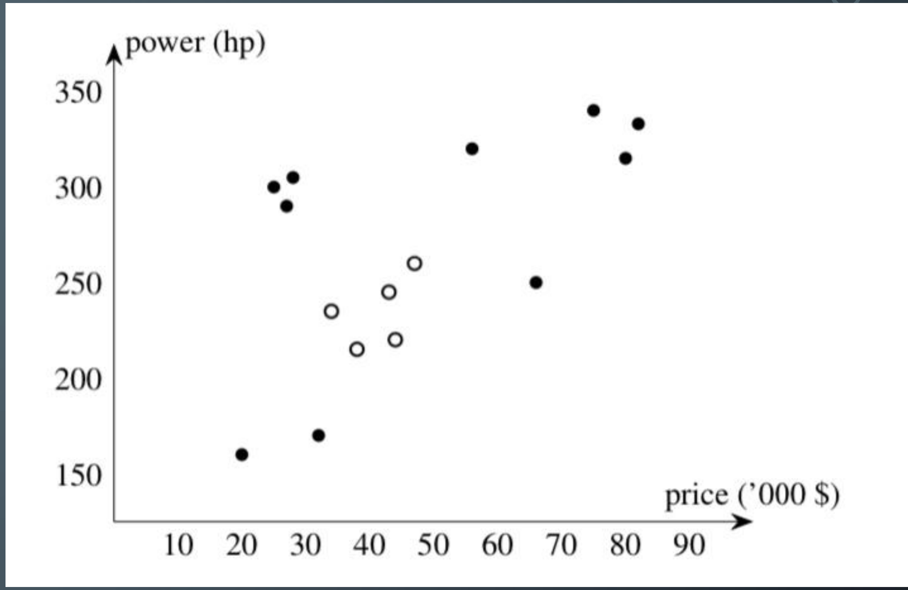
$$(p_1 \leq \text{price} \leq p_2) \text{ AND } (e_1 \leq \text{engine power} \leq e_2)$$

HYPOTHESIS CLASS H

$$(p_1 \leq \text{price} \leq p_2) \text{ AND } (e_1 \leq \text{engine power} \leq e_2)$$



- Scatter plot of price-power data (hollow circles indicate positive examples and solid dots indicate negative examples)

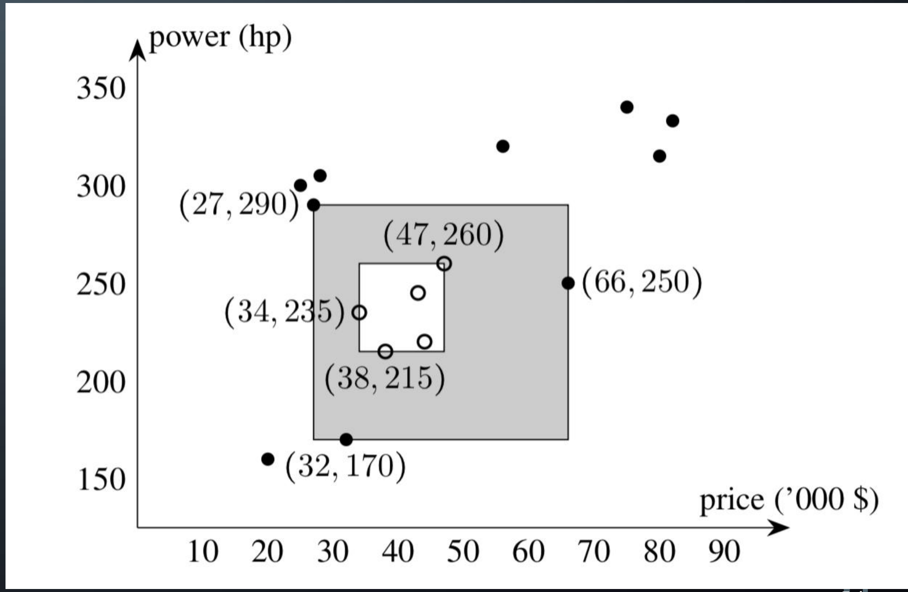


- Hypothesis space: the set of all axis-aligned rectangles in the price-power plane
- Version space consists of hypotheses corresponding to axis-aligned rectangles contained in the shaded region
- The inner rectangle is defined by

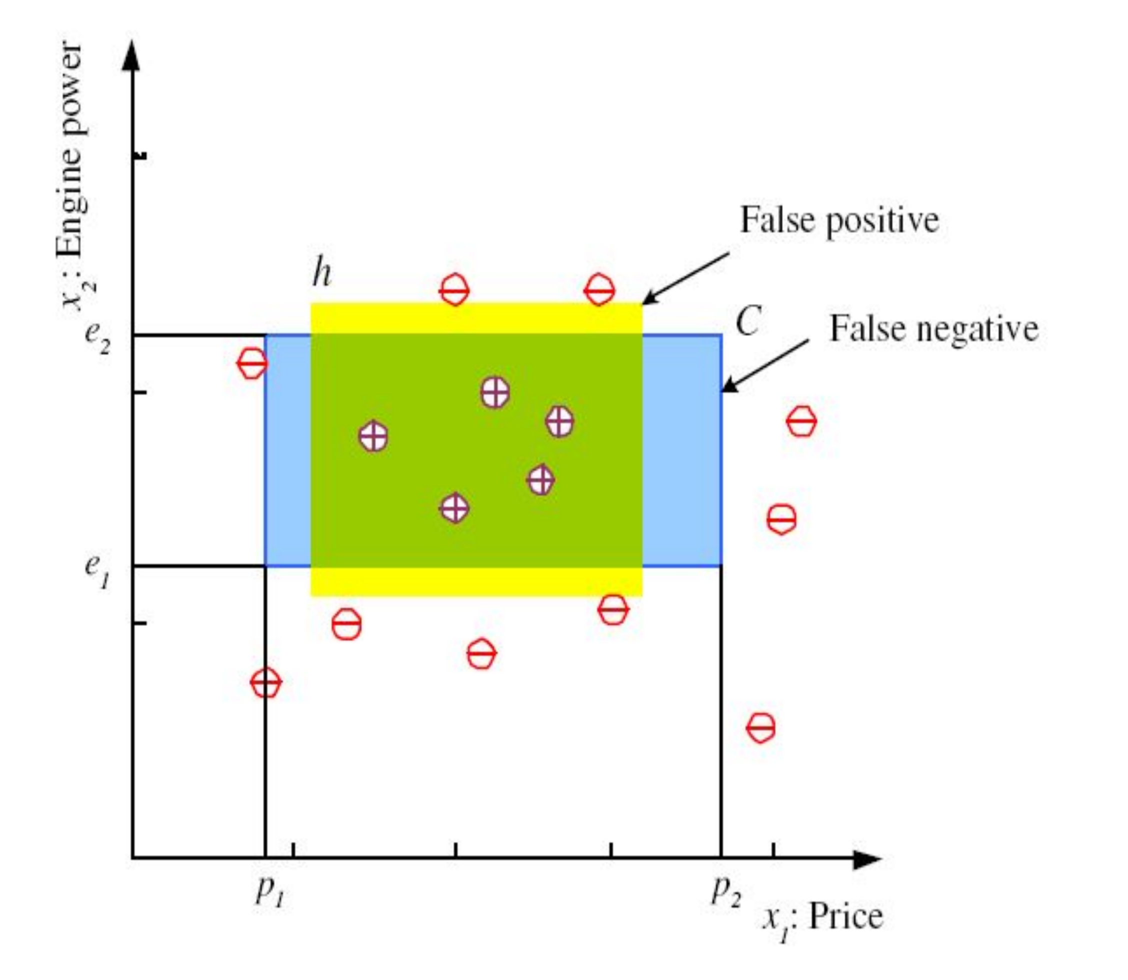
$$(34 < \text{price} < 47) \text{ AND } (215 < \text{power} < 260)$$

- The outer rectangle is defined by

$$(27 < \text{price} < 66) \text{ AND } (170 < \text{power} < 290)$$



EMPERICAL (TRAINING) ERROR



$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ says } \mathbf{x} \text{ is positive} \\ 0 \text{ if } h \text{ says } \mathbf{x} \text{ is negative} \end{cases}$$

Error of h on H

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

REFERENCES

- Introduction to Machine Learning ,MIT Press 2004-Ethem Alpaidin
- Machine Learning- Mitchell T
- Data mining concepts and techniques-Jiawei Han-Micheline Kamber - Jian Pei –Elsevier
- NPTEL course by IITK
- An Introduction to Statistical Learning-Gareth James , *Daniela Witten, Trevor Hastie* Robert Tibshirani –Springer
- ML lecture notes-Dr V N Krishnachandran