innovate achieve lead

BITS Pilani

Machine Learning (IS ZC464) Session 8: Decision Trees and Review Session

Decision Tree

- A decision tree takes as input an object or situation described by a set of attributes and returns a decision.
- This decision is the predicted output value for the input.
- The input attributes can be discrete or continuous.
- Classification Learning:
  - Learning a discrete valued function is called classification learning
- Regression :
  - Learning a continuous function is called Regression.

Decision Tree

- A decision tree reaches its decision by performing a sequence of tests.
- All non leaf nodes lead to partial decisions and assist in moving towards the leaf node.
- Leaf nodes are the decisions based on properties satisfied at non leaf nodes on the path from the root node.

Decision tree

- Leaf nodes depict the decision about a character having attributes falling on the path from the root node
- Each example that participate in the construction of the decision tree is called a training data and the complete set of the training data is called as training set.

Limitations of Decision Tree Learning

- The tree memorizes the observations but does not extract any pattern from the examples.
- This limits the capability of the learning algorithm in that the observations do not extrapolate to examples it has not seen.

How can we construct a decision tree for face recognition problem

- Define attributes
- Collect the attributes data from training samples
- Associate the output (to be used as leaf)

Imagine the size of decision tree with 1000 attributes capable of discriminating between persons!!!

Decision trees

- The **attributes** aid in taking decisions.
- The most appropriate attribute is selected for testing in the beginning else the **size of the tree** becomes large resulting in large computational time.
- Leaf nodes represent the **decisions**.
- The attributes falling in the path from represent the attributes fully able to define the decision at leaf.

Goal Predicate: WillWait()

Problem: decide whether to wait for a table at a restaurant, based on the following attributes:

1. Alternate: is there an alternative restaurant nearby?
2. Bar: is there a comfortable bar area to wait in?
3. Fri/Sat: is today Friday or Saturday?
4. Hungry: are we hungry?
5. Patrons: number of people in the restaurant (None, Some, Full)
6. Price: price range ($, $$, $$$)
7. Raining: is it raining outside?
8. Reservation: have we made a reservation?
9. Type: kind of restaurant (French, Italian, Thai, Burger)
10. WaitEstimate: estimated waiting time (0-10, 10-30, 30-60, >60)

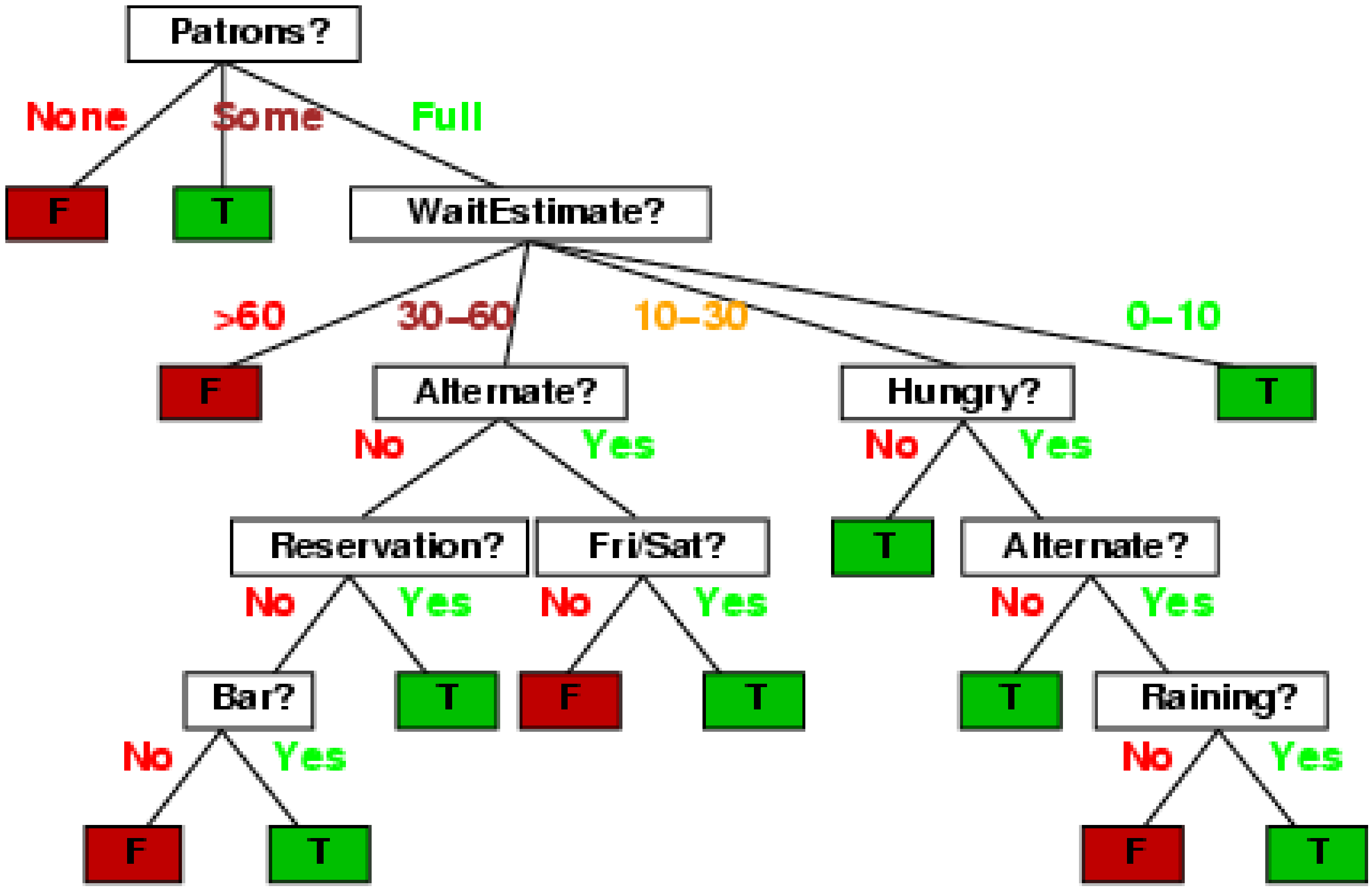
This slide is adapted from the text book and from the set of slides available at *aima.eecs.berkeley.edu/slides-**ppt**/m18-**learning**.ppt*

Attributes

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

This slide is adapted from the text book and from the set of slides available at *aima.eecs.berkeley.edu/slides-**ppt**/m18-**learning**.ppt*

Decision Tree



This slide is adapted from the text book and from the set of slides available at *aima.eecs.berkeley.edu/slides-**ppt**/m18-**learning.ppt***

Size of the decision tree

- The **size of the Decision** tree depends on the **choice of the attributes** and the **order** in which they are used to test the examples.
- Selection of attributes must be "fairly good" and "really useless" attributes (such as type) should be avoided
- The **quality of the attribute** can be measured.
- One measure can be the **amount of information** the attribute carries.

Information content

- If v_i are different possible answers and $P(v_i)$ are the probabilities that answer could be vi. Then the information content I of the actual answer is given by
  - $I(P(v_1), P(v_2), \ldots P(v_n)) = - \Sigma\, P(v_i)\log_2 P(v_i)$
- Assume that the training set contains 'p' positive examples and 'n' negative examples, then an estimate of the information contained in a correct answer is

$$I(p/(p+n), n/(p+n)) = - (p/(p+n))\log_2(p/(p+n)) - (n/(p+n))\log_2(n/(p+n))$$

Refer the given table of Attributes

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

This slide is adapted from the text book and from the set of slides available at *aima.eecs.berkeley.edu/slides-**ppt**/m18-**learning**.ppt*

Information content

- Since

  $I(p/(p+n), n/(p+n)) = -(p/(p+n))\log_2(p/(p+n)) - (n/(p+n))\log_2(n/(p+n))$

  – information $= -(6/12)\log_2(1/2) - (6/12)\log_2(1/2)$

  – $= -\log_2(1/2)$

  $= \log_2(1/2)^{-1}$

  $= \log_2(2)$

  $= 1$ bit

Generalize the splitting

- Let the attribute A divides the entire training set into sets E1, E2, … Ev. Where v is the total number of values A can be tested on.
- Assume that each set E_i contains p_i positive examples and n_i negative examples
- **Remainder (A)**

  $= \Sigma\ (p_i+n_i)/(p+n)\ I(p_i/(p_i+n_i),\ n_i/(p_i+n_i))$

  over i=1 to v

Gain(A)

- Gain(A)

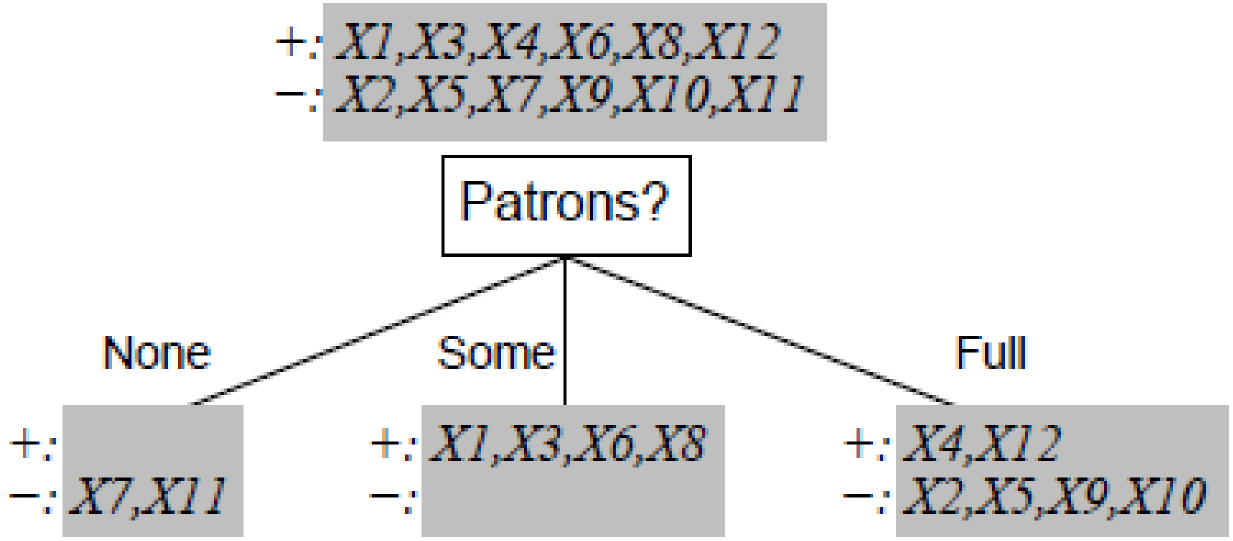
$$= I(p/p+n, n/p+n) - Remainder(A)$$

The heuristic to choose attribute A from a set of all attributes is the maximum gain

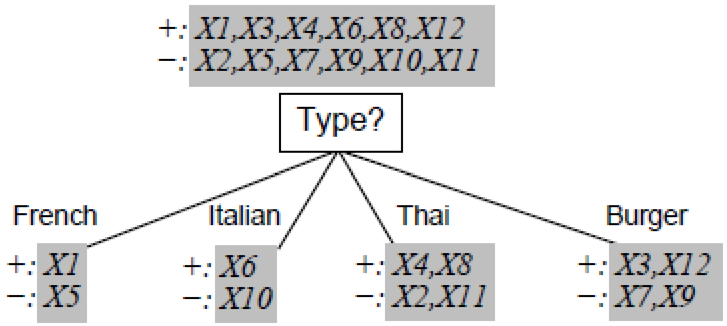
Compute

1. **Gain(Patrons)**
2. **Gain(type)**

Selecting patrons attribute



Selecting type as attribute



Gain(patron)

- 1 – ((2/12)I(o,1) + (4/12)I(1, 0) + (6/12) I(2/6, 4/6))
- Approximately equal to 0.541 bits

Refer the given table of Attributes and compute Gain

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	\$\$\$	F	T	French	0–10	T
X_2	T	F	F	T	Full	\$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	\$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	\$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	\$\$\$	F	T	French	>60	F
X_6	F	T	F	T	Some	\$\$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	\$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	\$\$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	\$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	\$\$\$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	\$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	\$	F	F	Burger	30–60	T

This slide is adapted from the text book and from the set of slides available at *aima.eecs.berkeley.edu/slides-**ppt**/m18-**learning**.ppt*

Decision Trees

- Learning is through a series of decisions taken with respect to the attribute at the non-leaf node.
- There can be **many trees** possible for the given training data.
- Finding the smallest DT is an NP-complete problem.
- Greedy selection of the attribute with largest gain to split the training data into two or more sub-classes may lead to approximately the smallest tree

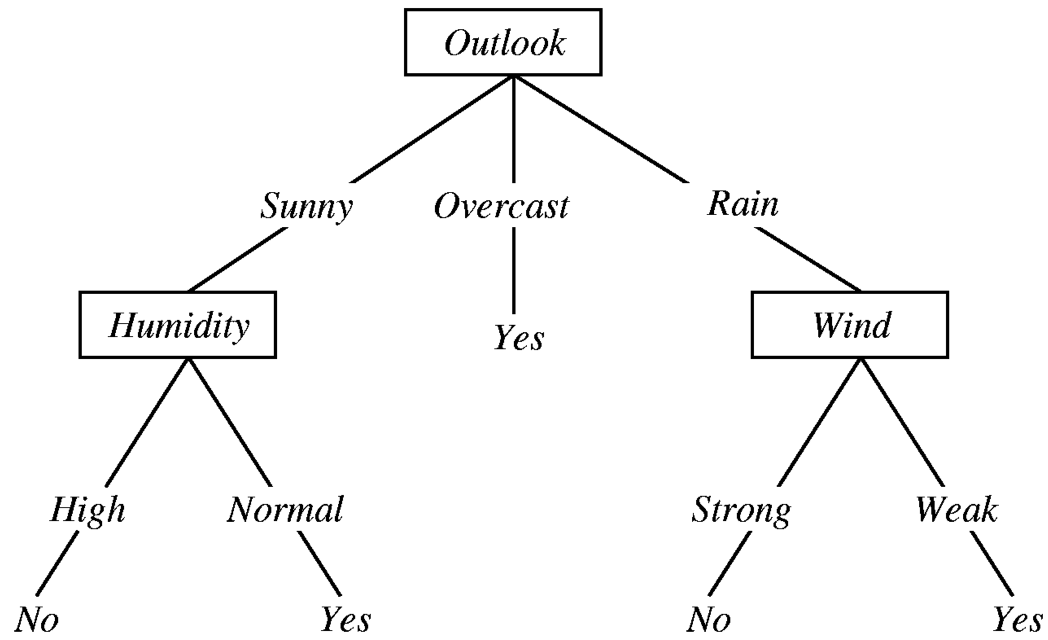
Decision Trees

- If the decisions are binary, then in the best case the decision eliminates almost half of the regions (leaves).
- If there are ‘b’ regions, then the correct region can be found in $\log_2(b)$ decisions in the best case.
- The height of the decision trees depends on the order of the attributes selected to split the training examples at each step.

Expressiveness of the DS

- A decision tree can represent a disjunction of conjunctions of constraints on the attribute values of instances.
  - Each path corresponds to a conjunction
  - The tree itself corresponds to a disjunction

Example



If (O=Sunny AND H=Normal) OR (O=Overcast) OR (O=Rain AND W=Weak)
then YES

- “A disjunction of conjunctions of constraints on attribute values”

Entropy

- It is the measure of the information content and is given by
  - $I = - \Sigma P(v_i)\log_2 P(v_i)$
  - Where v1,v2,..,vk are the values of the attribute on which the decisions bifurcate.

rec	Age	Income	Student	Credit_rating	Buys_computer
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31…40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31…40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<-=30	Medium	Yes	Excellent	Yes
r12	31…40	Medium	No	Excellent	Yes
r13	31…40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

Class Work

Remainder (A)
$$= \Sigma \, (p_i+n_i)/(p+n) \; I(p_i/(p_i+n_i), n_i/(p_i+n_i))$$
over i=1 to v

- Identify the examples belonging to the two sets constructed after the data is split on the basis of **attribute 'student'.**
- Compute the **total information content** of the training data.
- Compute the **information gain** if the training data is split on the basis of the attribute 'student'.
- Draw the **decision tree**, which may or may not be optimal.

Understand the examples

- Decisions are binary – yes / no
- Training data as <example, decision> pair
- <r1,no>, <r2,no>, <r3,yes>, <r4,yes> and so on
- Positive examples: r3, r4, r5, r7, r9, r10, r11, r12, r13
- Negative examples: r1,r2,r6, r8, r14
- Is the given training set sufficient to take any decision?
- Is the generalization capability of the given training set sufficient?

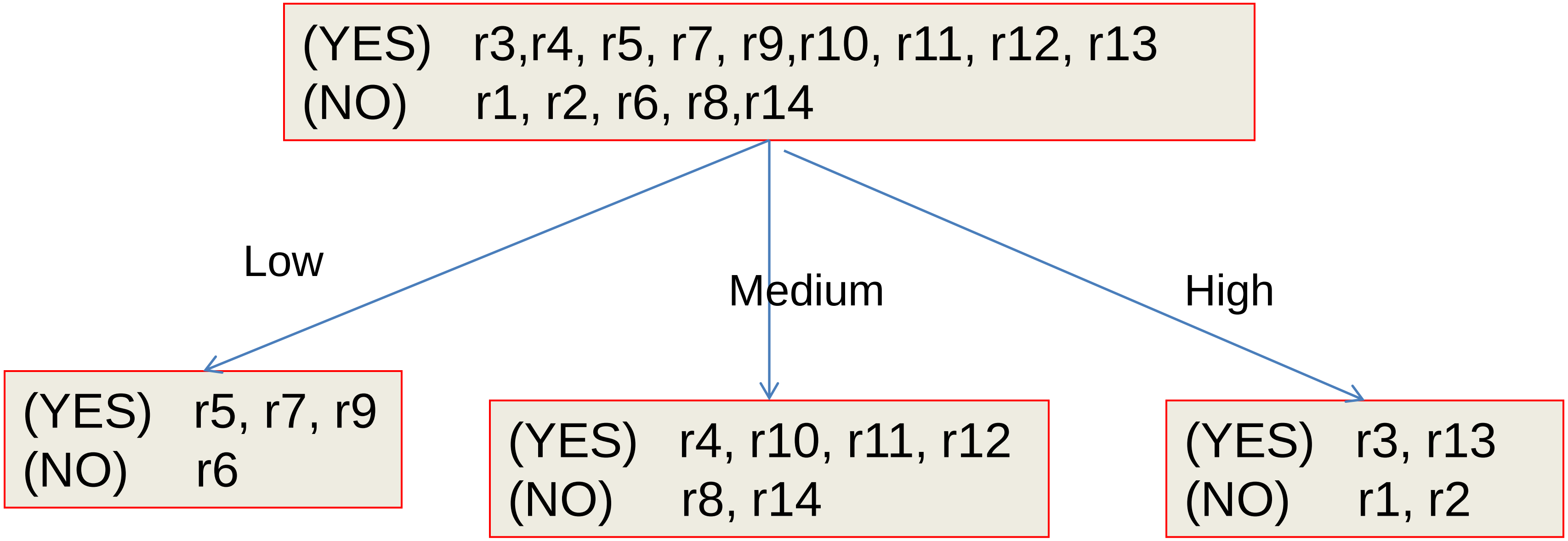
Information content of the given training data

- Here v1 = yes, v2 = no
- Positive examples: r3, r4, r5, r7, r9, r10, r11, r12, r13
- Negative examples: r1,r2,r6, r8, r14
- Total number of examamples = 14
- P(v1) = 9/14, P(v2)=5/14
- Information content is represented by the notion I(9/14, 5/14)
- Entropy = - $(P(v_1)\log_2(P(v_1)) + P(v_2)\log_2(P(v_2)))$

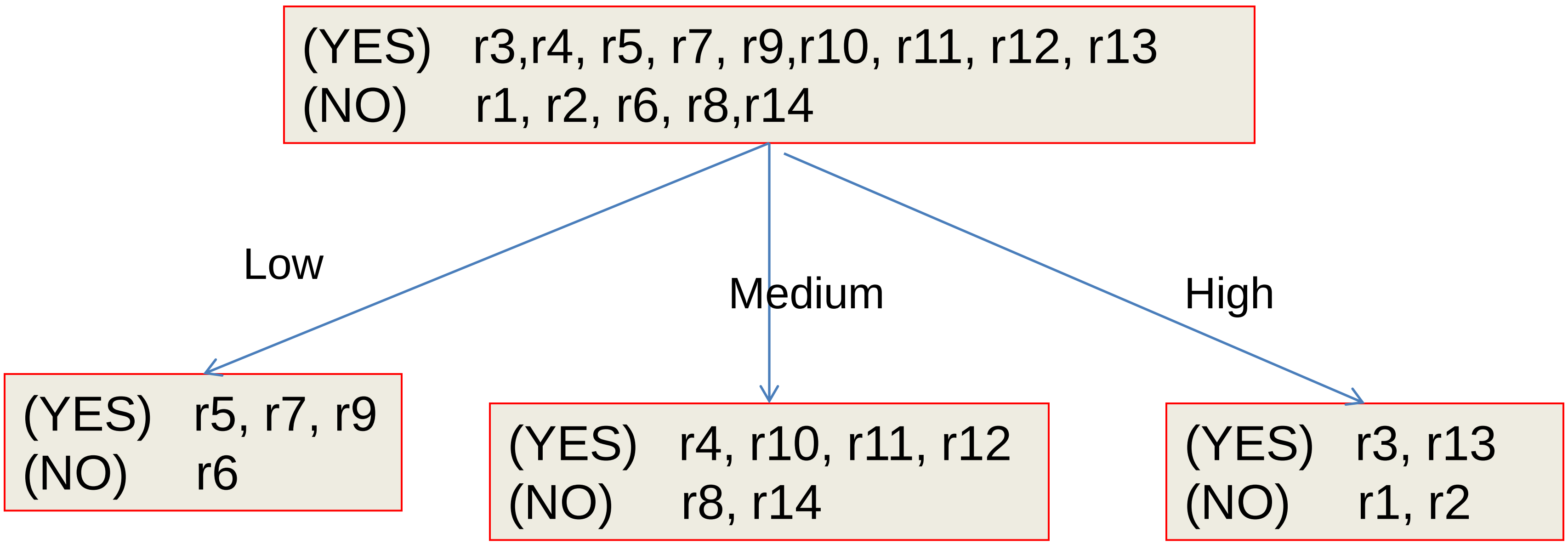
$$= -((9/14)* \log_2 (9/14) + (5/14)*\log_2 (5/14))$$

$$= 0.8108$$

Compute the significance of attribute ‘income’



Compute the significance of attribute 'income'



Observe that the split regions of examples possess mixed decisions, this shows the poor quality of the attribute 'income'

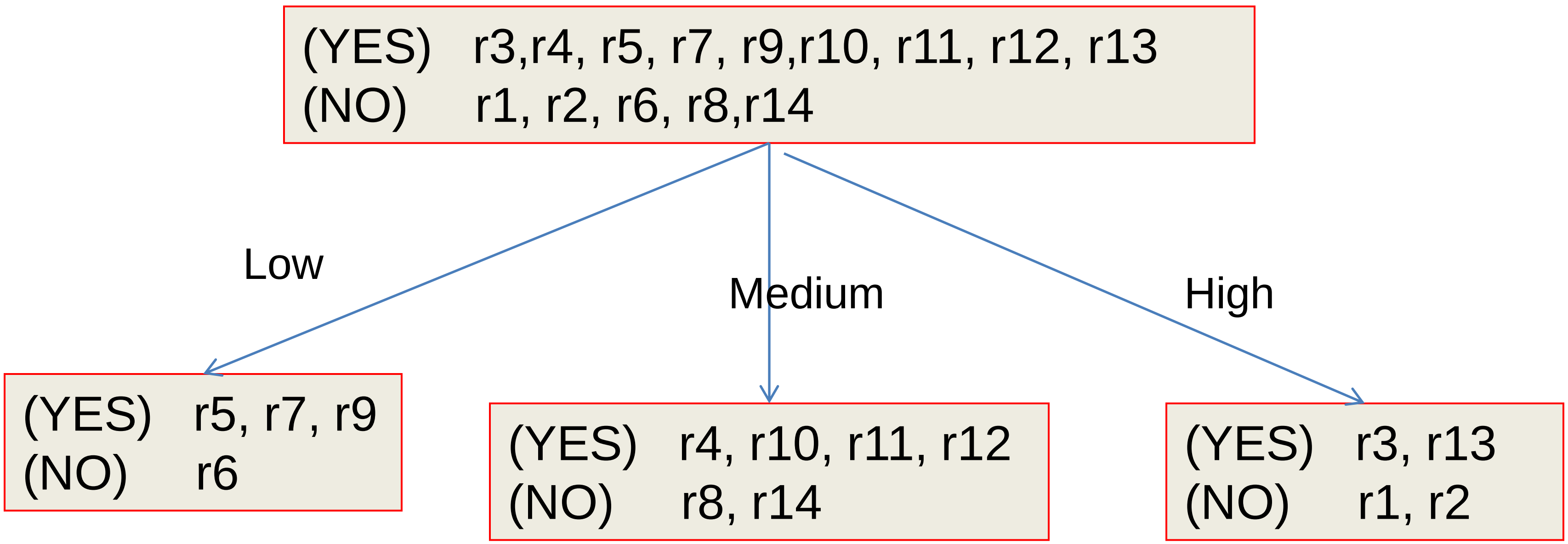
Recall Generalize the splitting

- Let the attribute A divides the entire training set into sets E1, E2, … Ev. Where v is the total number of values A can be tested on.
- Assume that each set E_i contains p_i positive examples and n_i negative examples
- **Remainder (A)**

$= \Sigma\ (p_i+n_i)/(p+n)\ I(p_i/(p_i+n_i),\ n_i/(p_i+n_i))$

over i=1 to v

Compute the significance of attribute ‘income’



$p_1 = 3$
$n_1 = 1$

$p_2 = 4$
$n_2 = 2$

$p_3 = 2$
$n_3 = 2$

Compute the Remainder information if attribute ‘income’ is used for splitting

- Remainder = $(4/14)*I(2/14,2/14)$

$$+ (6/14)*I(4/14,\ 2/14)$$

$$+(4/14)*I(2/14,2/14)$$

$$= 2*(4/14)\ (-(2/14)\ \log_2(2/14) - (2/14)\log_2 (2/14))$$

$$+(6/14)*(-(4/14)\log_2 (4/14)-(2/14)\log_2(2/14)$$

$$= 0.4583 + 0.0330 \quad = 0.4913$$

Review Session

- Mid Semester Syllabus
  - All topics and details discussed in Sessions 1-8 [Refer Slides and video contents]

- Not included
  - Decision Theory [Handout S. No. 2.2]
  - Expectation Maximization (EM) Algorithm [Handout S. No. 3.3]
  - Bias-variance decomposition [Handout S. No. 3.4]

What is learning?

- Learning (for humans) is experience from past.
- A machine can be programmed to gather experience in the form of facts, instances, rules etc.
- A machine with learning capability can predict about the new situation (seen or unseen) using its past experience.
- Examples:
  - As we humans can tell a person’s name seeing him/her second or fifth time, a machine can also do that.
  - As we humans can recognize a person’s voice even if not seeing person’s face, a machine can also be made to learn to do the same.

Class Experiment: Training

- Let
  - AA denote 5
  - BB denote 6
  - AAA denote 50
  - BBB denote 60
  - AAAA denote 500
  - BBBB denote 600
- Can you find out the equivalent numerical value of AAAAA? 5000: yes/no?
- Or of AABB? Not yet trained………

Artificial Intelligence: An intelligent car navigation system [An Example]

- A system to navigate a car to the airport works on its vision enabled using camera mounted at the front of the car.
- The system "sees" the lane limits, the vehicles on the way and controls the car from colliding. **[Vision]**
- It follows the road directions.
- It also follows the road rules.
- The system learns to handle unforeseen situations. For example if the traffic flow is restricted on a portion of the road temporarily, the system takes the alternative path.**[learning]**

More intelligence can be expected

- The system “listens” to the person sitting in the car to stop at a nearby hotel for a tea and “sees” around to find a hotel, keeps travelling till it finds one and stops the car. **[speech Recognition, Vision]**
- Understands the mood of the person and starts music to suit the mood of the person. **[Facial Expression]**
- Can answer the queries, such as “how far is Pilani?”, “What is the time”, “can I sleep for an hour?”, “Please wake me up when it is 11:00 in the morning?” **[Natural Language Processing]**

Other intelligent systems

- Smart home
  - Lights switch off if there is no one in the room
  - Curtain pull off at the sun rise
  - Dust bin is emptied before it is overflowing
  - Smart water taps, toilets etc.
- Smart office
  - Automatic meeting summary
  - Speaker recognition and summary generation
- Automatic answering machine

Other intelligent machines

- An airplane cockpit can have a intelligent system that takes automatic control when hijacked [context and speech understanding, NLP, vision]
- Medical diagnosis systems trained with expert guidance can diagnose the patients disease based on the xray, MRI images and other symptoms
- Automated theorem proving
- General problem solver

Intelligent Agent

- An intelligent agent is a system that perceives its environment and takes actions which maximize its chances of success.
- Artificial Intelligence aims to build intelligent agents or entities.

Machine Learning Applications

- Speech recognition
- Automatic news summary
- Spam email detection
- Credit card fraud detection
- Face recognition
- Function approximation
- Stock market prediction and analysis
- Etc.

Machine Learning

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. (Tom Mitchell)

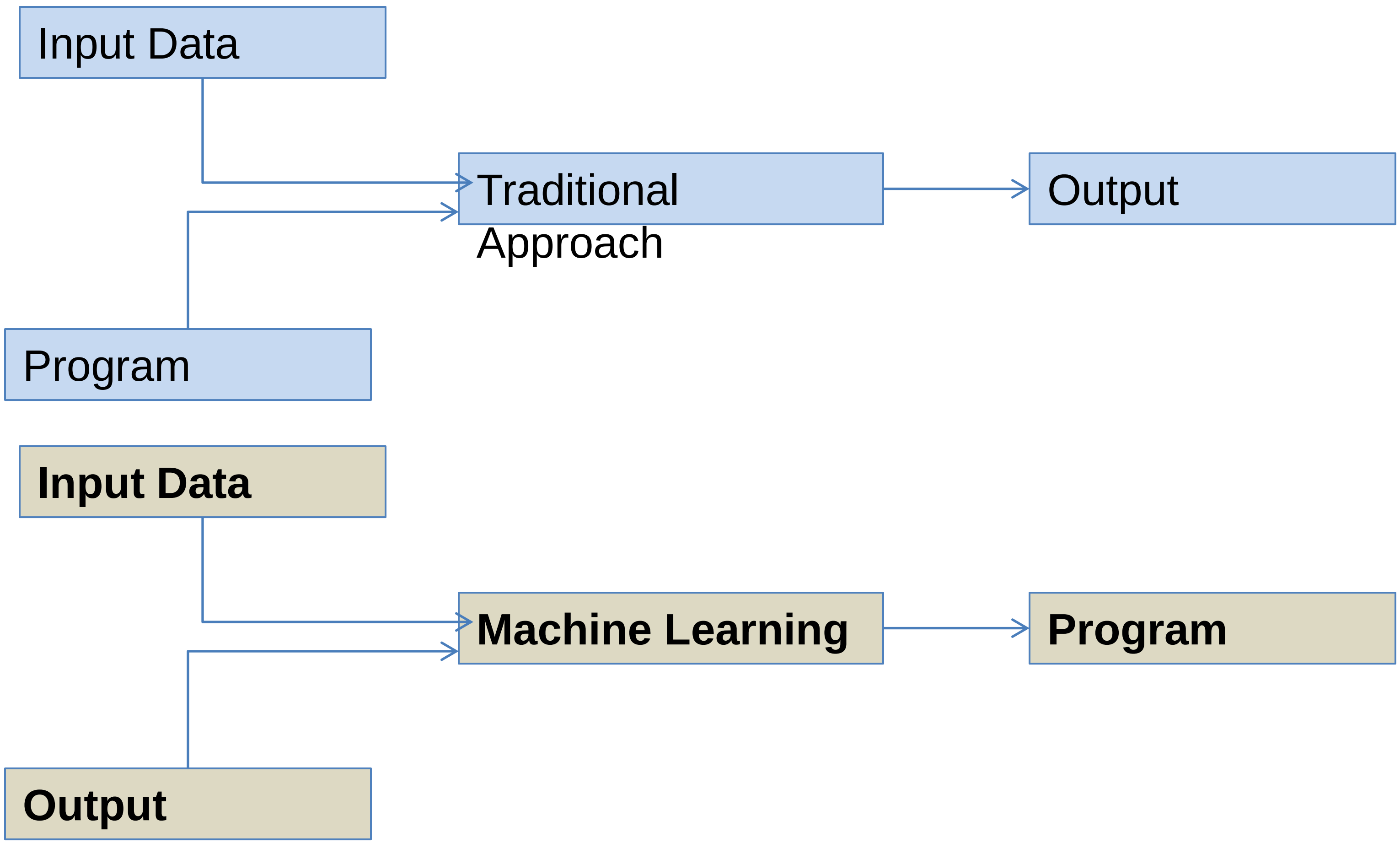
Learning From Observations

- Learning Element:
  - responsible for making improvements
- Performance Element:
  - responsible for selecting external actions
- The learning element uses feedback from the critic on how the agent is doing and determines how the performance element should be modified to do better in the future

Design of a learning Element

- Affected by three major issues:
  - Which components of the performance element are to be learned
  - What feedback is available to learn these components
  - What representation is used for the components.

Traditional Vs. Machine Learning



Training and testing: Prediction

X	Y
1	1
5	5
2	2
4	4
3	3

- Recall Learning: A machine with learning capability can predict about the new situation (seen or unseen) using its past experience.
- Prediction:

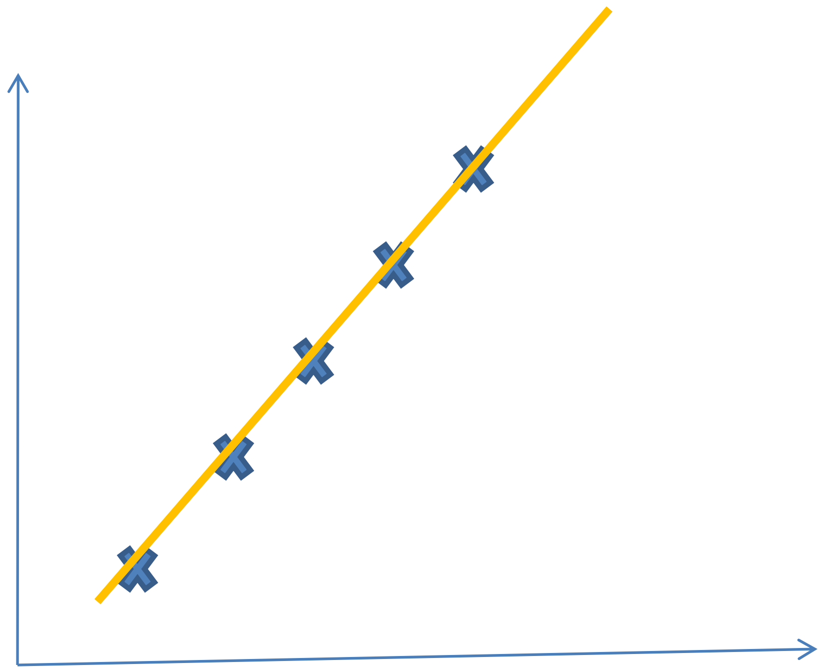
  Given values of x and y

  Predict value of y for x = 71
- Prediction is based on learning of the relationship between x and y
- Training data is the collection of (x,y) pairs
- Testing data is simply value of x for which value of y is required to be predicted.

Learning of a function from given sample data

Review Session

Straight Line



What did the system learn?

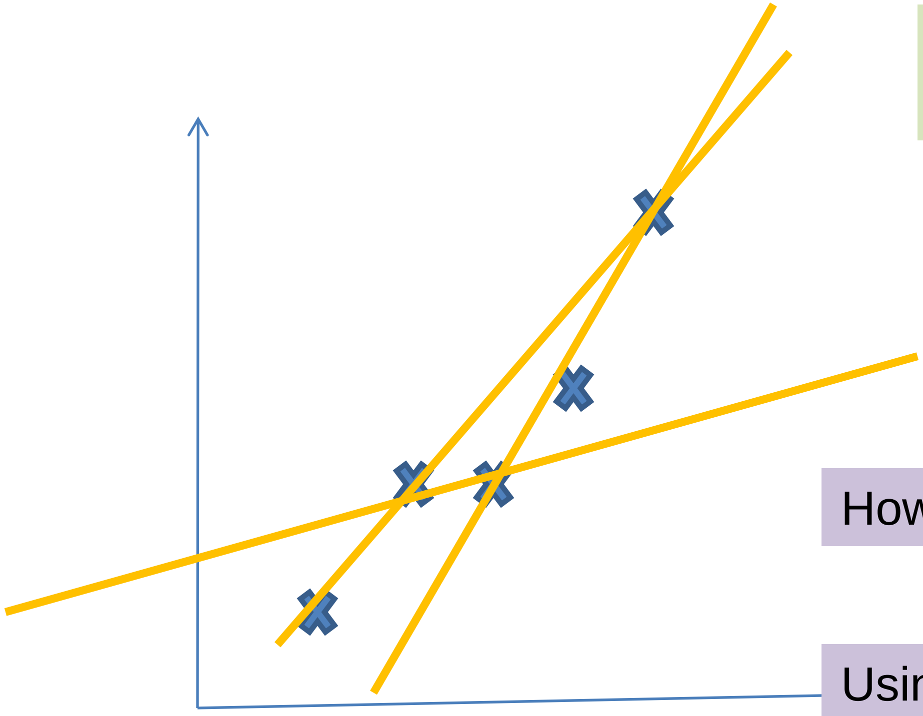
- Y = f(x)
- Y = x
- What is its generalization ability?
- Most accurate or we can say 100%
- What if the data to train the system changes slightly? The machine can be still made to learn.

X	Y
1	1
5	5
2	2
4	3
3	2

Learning of a function from given sample data-straight line learning

Review Session

Straight Line



Which line fits the best?

How?

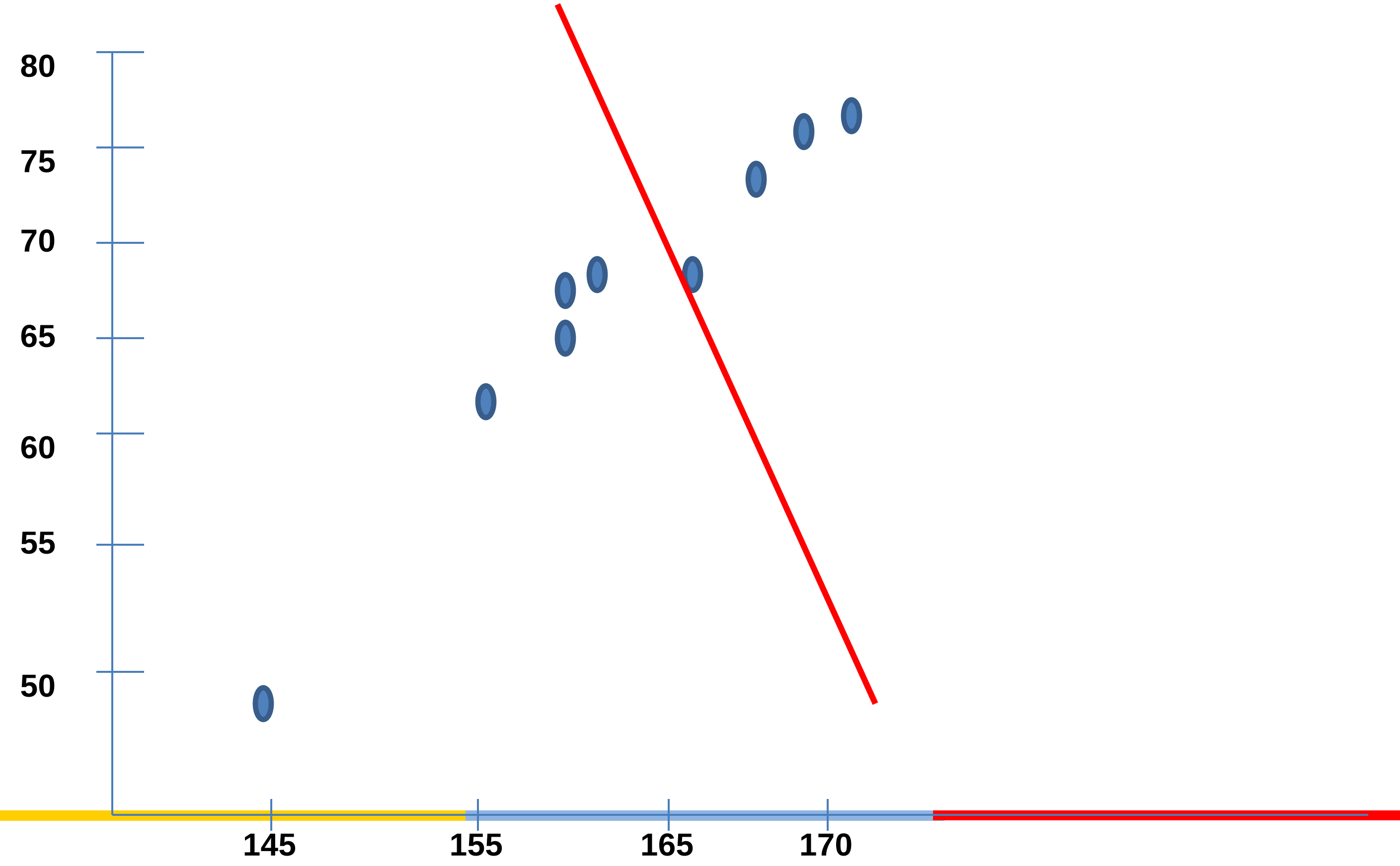
Using the data – known as training data i.e. (x,y) pair

Line is represented by parameters of slope and

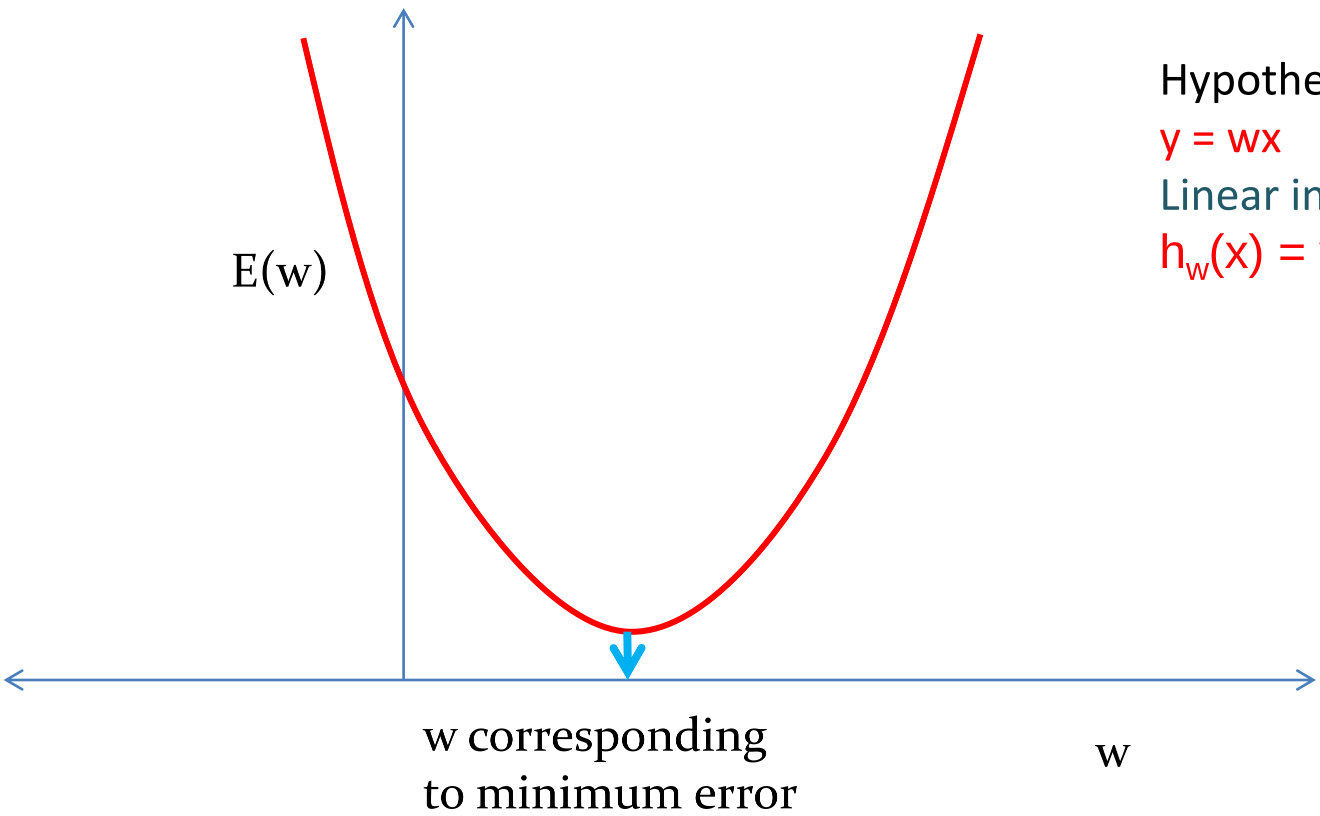
Machine must learn on its own- which is the best fit

Understanding ERROR

Which line(hypothesis) fits the given data best?



Plotting error when y=f(x)



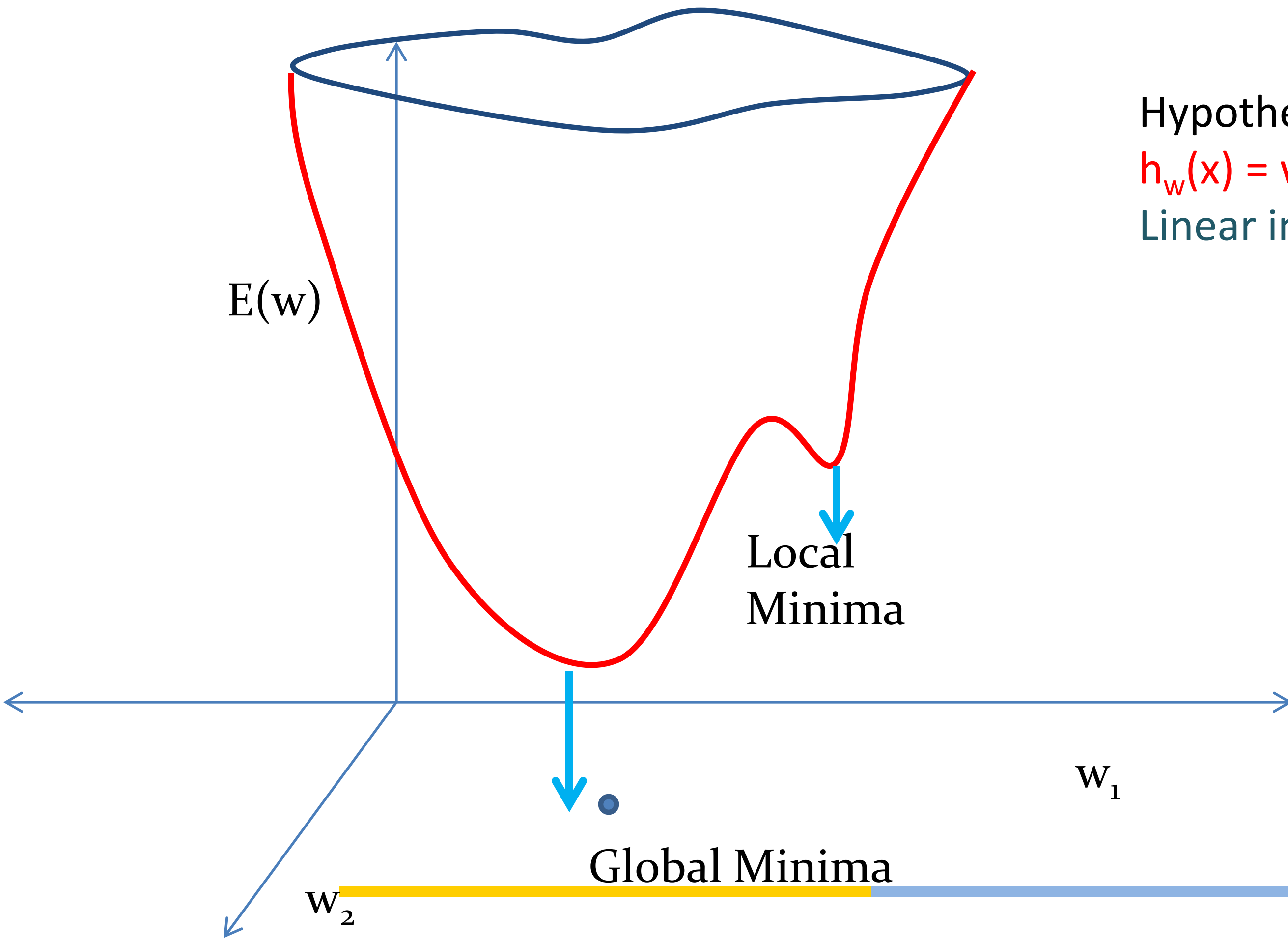
Hypothesis function

$y = wx$

Linear in one variable

$h_w(x) = wx$

Plotting error when $y=f(x_1,x_2)$



Hypothesis function

$h_w(x) = w_1x_1 + w_2x_2$ ....(1)

Linear in two variables

Uncertainty in real world

- Uncertainty in reaching New Delhi Airport in 5 hours from Pilani
  – Cab engine may or may not work at any moment
  – The route is diverted due to a procession on the way
  – The road condition is bad unexpectedly
  – The tire needs replacement

  Etc.
- A person having stomach ache can be told that he is suffering from ulcer, while in actual it may be gastritis or overeating

Recall

- Knowledge representation using Probability
- Random variables
- Atomic events
- Conditional probability
- Prior probability
- Marginalization
- Bayes' theorem and its application in problem solving
- Joint probability distribution (JPD) table and probabilistic inference

Bayes’ theorem

- Bayes’ theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probability of observing various data given the hypothesis, and the observed data itself.

Bayesian learning

Example 1: observation of sounds

Training with Observed data: $\{d_1, d_2, d_3\}$ =training data (say D)

d_1: cat sounds with 'ae'

d_2: pot sounds with 'aw'

d_3: mat sounds with 'ae'

Sounds 'ae' and 'aw' are the observed targets that we know.

Prior probabilities
P(sound = 'ae') = 0.5
P(sound = 'aw') = 0.5

features such as 'a' and 'o' are obtained through preprocessing of the given words – by parsing

Conditional probabilities are represented as
P('ae' |feature = 'a')=2/3
P('aw'| feature = 'o') = 1/3

OR
P('ae' |d1,d2,d3)=2/3
P('aw'|d1,d2,d3)=

OR
P('ae' |D)=2/3
P('aw'|D)= 1/3

Hypothesis

- In learning algorithms, the term hypothesis is used in contexts such as
  - Concept learning or classification: class label or category
  - Function approximation: a curve, a line or a polynomial
  - Decision making: a decision tree
- Plural of hypothesis: Hypotheses (multiple labels, multiple curves, multiple decision trees)
- Best Hypothesis (Always preferred) : Most appropriate class, best fit curve, smallest decision tree

Bayesian Learning

- Training : Through the computation of the probabilities as in previous two slides
- Testing : of unknown words
  - Example testing:

  Which sound does the word ‘cat’ make?

  Preprocess(cat) to get feature ‘a’ and compute P(h|a), where P(h|D) is known, where D is the set of 10 observations used to train the system and ‘h’ is the hypothesis.

  Compute probabilities P(ae | a), P(oo | a), P(a~|a) and P(aw |a) to obtain the likelihood of sound of cat.

Maximum a Posteriori (MAP) hypothesis

Review Session

- Consider a set of hypotheses H and the observed data used for training D
- Define

$$h_{MAP} = \underset{h \in H}{Arg\max}\, P(h \mid D)$$

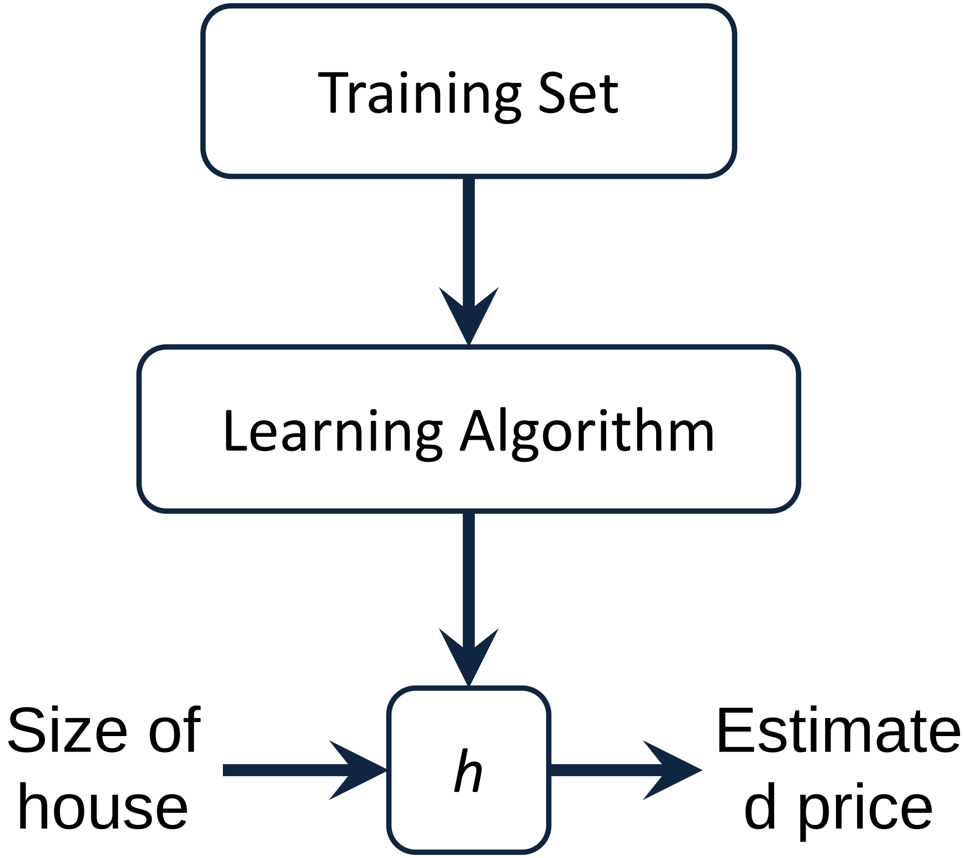
- The maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis.

Recall

- MAP algorithm
- Gibbs Algorithm
- Minimum Description Length Principle
- Information theory – entropy
- Bayes' Optimal Classifier
- Naïve Bayes' Classifier

What is Regression?

- The goal of **regression** is to predict the value of one or more continuous target variables ‘t’ given the value of a D-dimensional vector x of input variables.
- Polynomial curve fitting is an example of regression.



How do we represent *h* ?

Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

θ_i's: Parameters

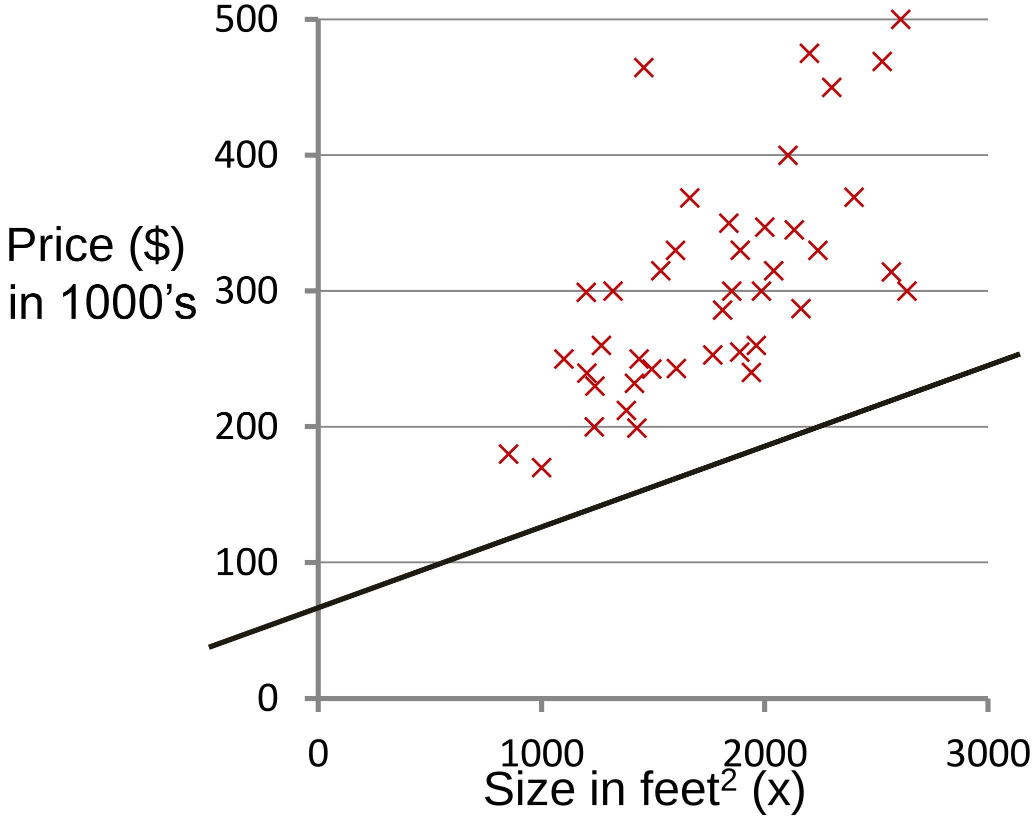
How to choose θ_i's ?

Linear regression with one variable.
Univariate linear regression.

Review Session

$h_\theta(x)$

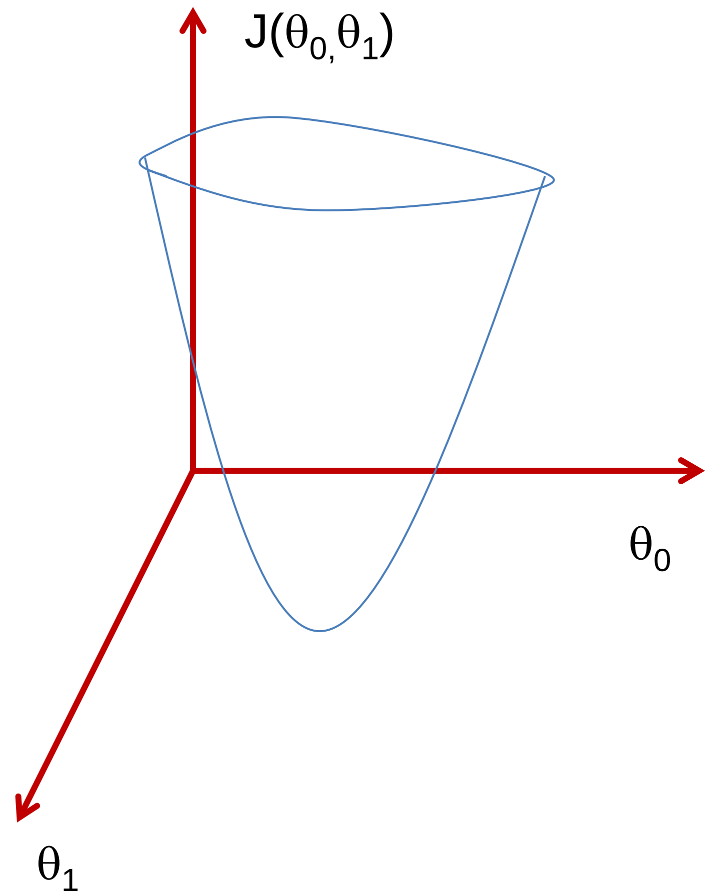
(for fixed θ_0, θ_1, this is a function of x)



$$h_\theta(x) = 50 + 0.06x$$

$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



Gradient descent algorithm

Learning Rate: α

$$\text{repeat until convergence } \{$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$

$$\}$$

Correct: Simultaneous update

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\theta_1 := \text{temp1}$$

Incorrect:

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_1 := \text{temp1}$$

Linear Regression

$$y(x,w) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_D x_D$$

Key Properties of Linear Regression

- y is a **linear** function of the parameters $w_0, w_1, w_2, \ldots w_D$
- y is a **linear** function of the input variables (features) $x_0, x_1, x_2, \ldots x_D$

Generalized Form of Linear Regression

- A notion of class of functions $\phi_i(x)$ is used to represent the regression function
- $y(x,w)= w_0 + w_1 x_1+ w_2 x_2+ w_3 x_3+ \ldots..+w_D x_D$

is represented as

$$y(x,w)= w_0 + w_1 \phi_1(x)+ w_2 \phi_2(x)+ w_3 \phi_3(x)+ \ldots..+w_D \phi_D(x)$$

Where $\phi_i(x)=x_i$

- $\phi_i(x)$ are called as basis functions for $i=1,2,3,..D$

Basis functions

- Linear basis functions $\phi_i(x)=x$ (Linear in x)

- Nonlinear basis functions

  $\phi_i(x)=x^2$ (Quadratic in x)

  $\phi_i(x)=x^3$ (Cubic in x)

What is linear in linear regression?

Review Session

- The following expression is linear in W

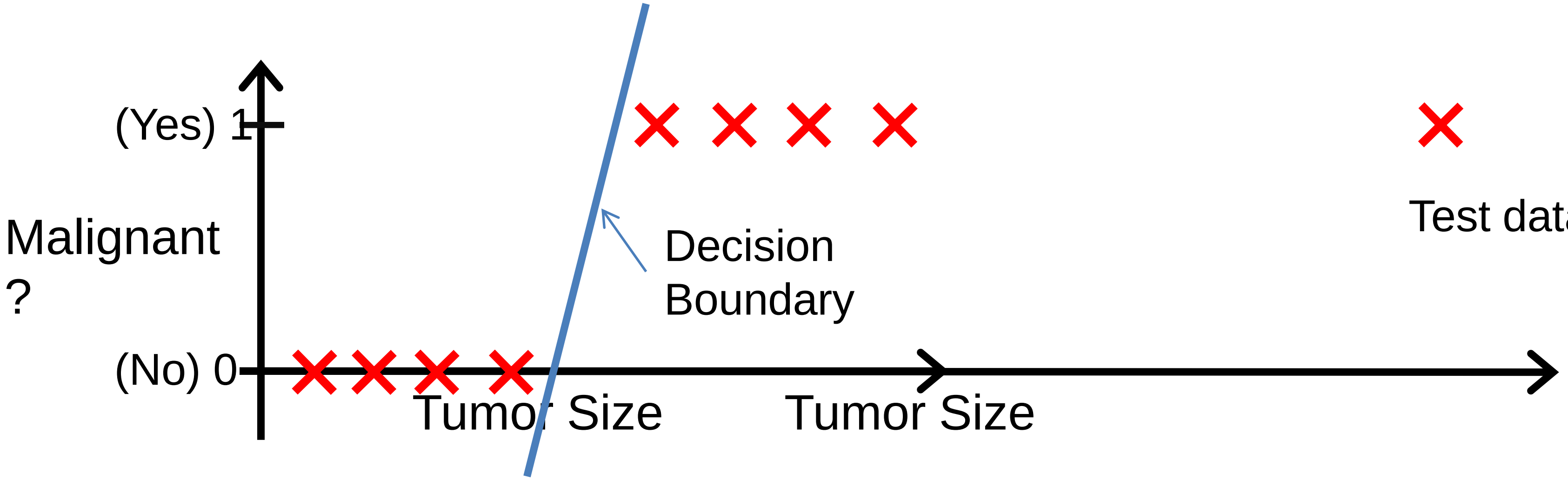
$y(x,w) = w_0 + w_1\,\phi_1(x) + w_2\,\phi_2(x) + w_3\,\phi_3(x) + \ldots\ldots + w_D\,\phi_D(x)$

- The basis functions may be linear or nonlinear in x

Classification

- The goal of classification is to take an input vector x and to assign it to one of K discrete classes C_k where k = 1, 2, 3, …, K
- Examples
  - Email: Spam / Not Spam?
  - Online Transactions: Fraudulent (Yes / No)?
  - Tumor: Malignant / Benign ?

Example of a Decision Boundary



Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "y = 1"

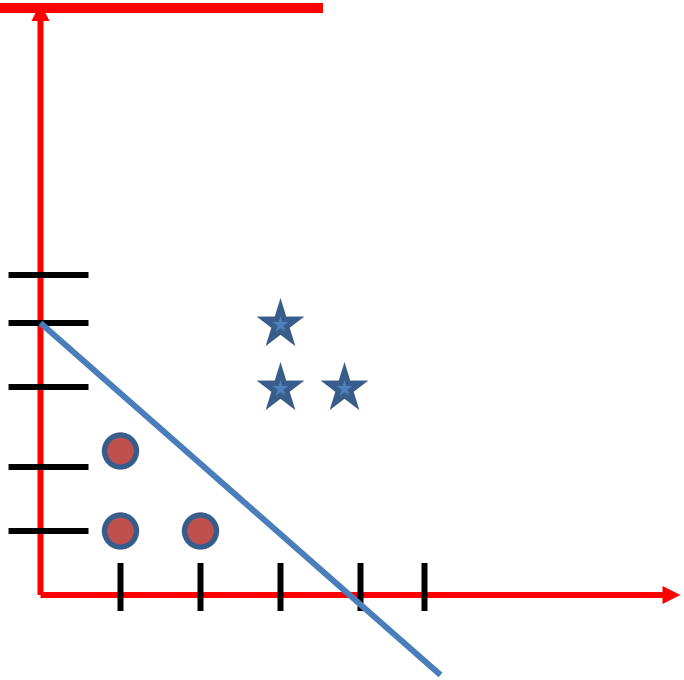
If $h_\theta(x) < 0.5$, predict "y = 0"

Solving Classification Problems

- Require the decision boundaries (or surfaces in hyper dimensional space) to be identified based on the training data.
- The decision boundary may be a line, a polynomial curve or a surface.
- The decision boundary can be represented as a hypothesis $h_\theta(x)$

Example

- Test vector <4,4>
- Compute $h(x) = x_1+x_2-1$ as $4+4-1= 7$
- Since h(x) > 4, then the test data belongs to class 2
- Test vector <2,1.5>
- h(x) = 2+1.5-1 = 2.5 <4
- Then it belongs to class 1



Recall

- Decision boundaries
- Binary and multi class classification
- Decision trees
- Gain and remainder
- Information content etc.

Bayes' Theorem based problem

- A box contains 10 red and 15 blue balls. Two balls are selected at random and are discarded without their colors being seen. If a third ball is drawn randomly and observed to be red, what is the probability that both of the discarded balls were blue?
- Atomic events for the selection of two balls

  RR: both balls were Red

  RB: One is red ball and the other is Blue

  BB: Both are blue balls

To find P(R | BB) : Probability that the third ball is red given that both the discarded balls were blue.

- Number of ways to select two balls = ${}^{25}C_2 = 300$
- Number of ways to select two red balls = ${}^{10}C_2 = 45$
- Number of ways to select two blue balls = ${}^{15}C_2 = 105$
- Number of ways to select one red and one blue ball = ${}^{10}C_1 * {}^{15}C_1 = 10*15 = 150$
- P(RR) = 45/300
- P(BR) = 150/300
- P(BB) = 105/300
- Therefore the probability that the third ball is red P(R) = P(RRR)+ P(RBB)+P(RBR)

Review Session

Probability of the third ball being red

P(R) = P(RRR)+ P(RBB)+P(RBR)

= P(R|RR)*P(RR) + P(R|BB)* P(BB)+ P(R|BR)*P(BR)

= (1/8)*(45/300) + (1/10)*(105/300)+ (1/9)*(150/300)

= 0.125*0.15 + 0.1*0.35 + 0.11*0.5

= 0.01875 + 0.035 + 0.0556 = 0.10935

Probability that the two discarded balls were blue given that the third ball is red

Review Session

- The expression is P(BB|R) and is given by

$$\frac{P(R|BB)P(BB)}{P(R|BB)P(BB)+P(R|BR)P(BR)+P(R|RR)P(RR)}$$

Using Bayes' Theorem

P(BB|R) = 0.035 / 0.10935 = 0.32 (Answer)