# Machine Learning

Syllabus

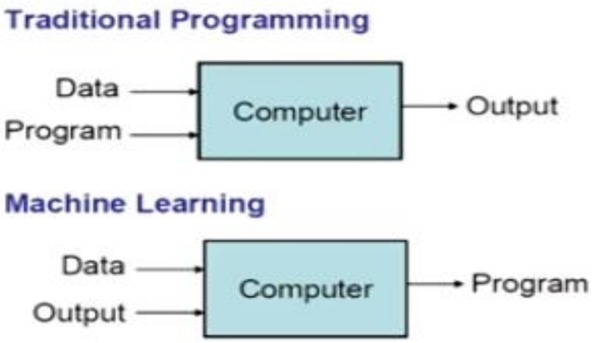
- Introduction
- Mathematics Preliminaries
- Bayesian Learning
- Linear Models for Classification
- Linear Models for regression
- Decision Trees
- Neural Networks
- Instance Based Learning
- Ensemble Learning
- SVM
- Unsupervised Learning

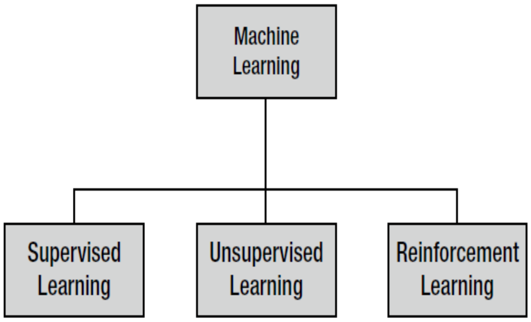
Introduction:

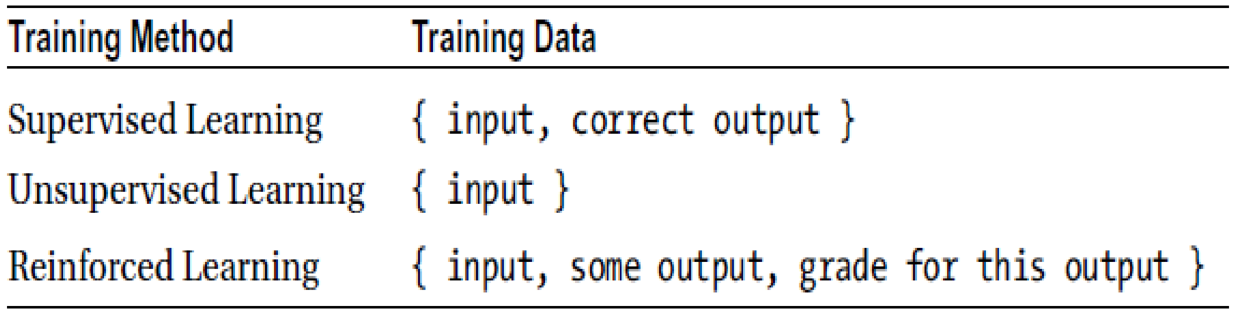
**Machine Learning and Examples of its Applications**

Machine Learning is a modeling technique that involves data. **Machine learning** is defined as an automated process that extracts patterns from data. To build the models used in predictive data analytics applications, we use **supervised machine learning**.

- Machine Learning is a modeling technique that involves data.
- 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, 1998)
-







| Training Method | Training Data |
|---|---|
| Supervised Learning | { input, correct output } |
| Unsupervised Learning | { input } |
| Reinforced Learning | { input, some output, grade for this output } |

Types of Data:

**Numeric:** True numeric values that allow arithmetic operations (e.g., price, age)

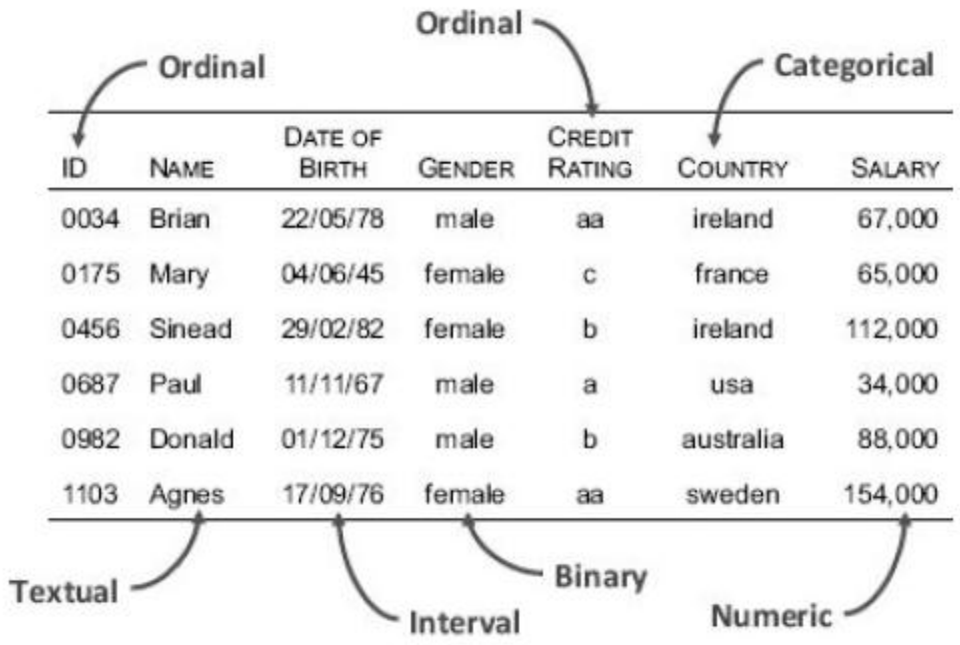
**Interval:** Values that allow ordering and subtraction, but do not allow other arithmetic operations (e.g., date, time)

**Ordinal:** Values that allow ordering but do not permit arithmetic (e.g., size measured as small, medium, or large)

**Categorical:** A finite set of values that cannot be ordered and allow no arithmetic (e.g., country, product type)

**Binary:** A set of just two values (e.g., gender)

**Textual:** Free-form, usually short, text data (e.g., name, address)



| ID | Name | Date of Birth | Gender | Credit Rating | Country | Salary |
|---|---|---|---|---|---|---|
| 0034 | Brian | 22/05/78 | male | aa | ireland | 67,000 |
| 0175 | Mary | 04/06/45 | female | c | france | 65,000 |
| 0456 | Sinead | 29/02/82 | female | b | ireland | 112,000 |
| 0687 | Paul | 11/11/67 | male | a | usa | 34,000 |
| 0982 | Donald | 01/12/75 | male | b | australia | 88,000 |
| 1103 | Agnes | 17/09/76 | female | aa | sweden | 154,000 |

Application areas:

Speech and Hand Writing Recognition)

•Robotics (Robot locomotion)

•Search Engines (Information Retrieval)

•Learning to Classify new astronomical structures
•Medical Diagnosis
•Learning to drive an autonomous vehicle
•Computational Biology/Bioinformatics
•Computer Vision (Object Detection algorithms)
•Detecting credit card fraud
•Stock Market analysis
•Game playing

Possible Issues in ML:

- What algorithms are available for learning a concept? How well do they perform?
- How much training data is sufficient to learn a concept with high confidence?
- When is it useful to use prior knowledge?
- Are some training examples more useful than others?
- What are the best tasks for a system to learn?
- What is the best way for a system to represent its knowledge?

## Learning

- Learning is defined as "any relatively permanent change in behaviour that occurs as a result of practice and experience". This definition has three important elements.
- Change in behaviour
- Change through practice or experience
- Change should last relatively permanent(longer)

Designing a Learning system

- Choosing the training experience
- Choosing the Target function
- Algorithm for learning from examples

Designing a Learning system- Training Experience

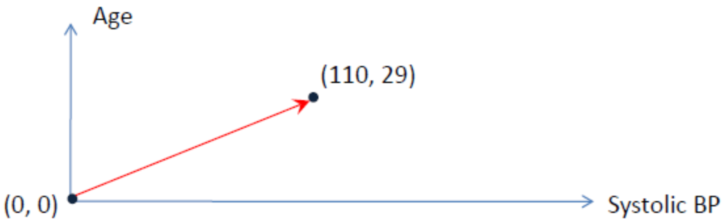
- One key attribute is whether the training experience provides ***direct*** or ***indirect*** feedback regarding the choices made by the performance system.
- A second important attribute of the training experience is the degree to which the learner controls the sequence of training examples[supervise/unsupervised].
- A third important attribute of the training experience is how well it represents the distribution of examples over which the final system performance ***P*** must be measured[train/test].
- Move to numerical domain (or) assign values, V: S-> R.
- More expressive the function, the closer it is to the truth but will need more training examples[regression equation].

## Mathematics Preliminaries

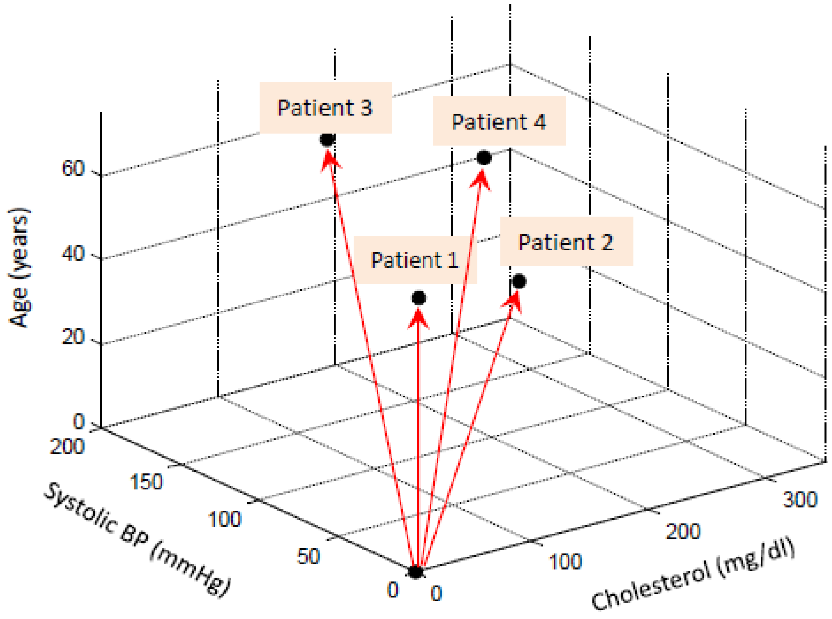
- **Linear Algebra.**

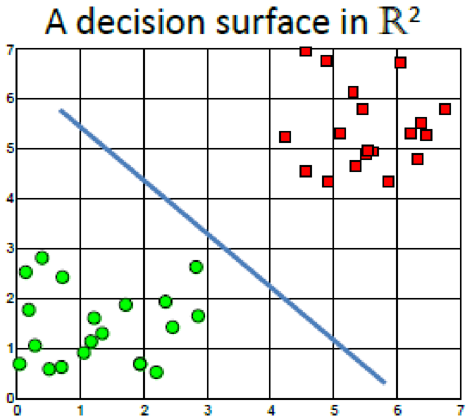
Vectors.

- Consider a patient described by 2 features:
- *Systolic BP*= 110 and *Age*= 29.
- This patient can be represented as a vector in R$^2$.

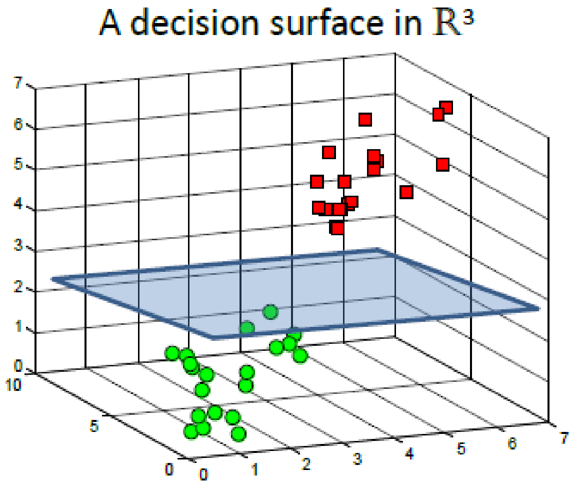


| Patient id | Cholesterol (mg/dl) | Systolic BP (mmHg) | Age (years) | Tail of the vector | Arrow-head of the vector |
|---|---|---|---|---|---|
| 1 | 150 | 110 | 35 | (0,0,0) | (150, 110, 35) |
| 2 | 250 | 120 | 30 | (0,0,0) | (250, 120, 30) |
| 3 | 140 | 160 | 65 | (0,0,0) | (140, 160, 65) |
| 4 | 300 | 180 | 45 | (0,0,0) | (300, 180, 45) |





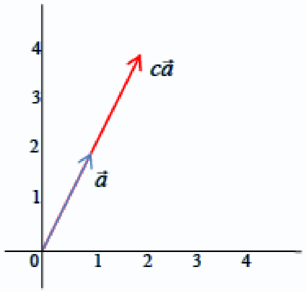
A decision surface in $\mathbb{R}^2$



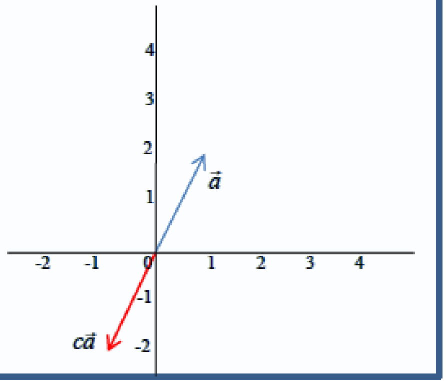
A decision surface in $\mathbb{R}^3$

- Vectors. Multiplying a vector by a scalar.

$$\vec{a} = (1,2)$$
$$c = 2$$
$$\vec{c}a = (2,4)$$



$$\vec{a} = (1,2)$$
$$c = -1$$
$$\vec{c}a = (-1,-2)$$

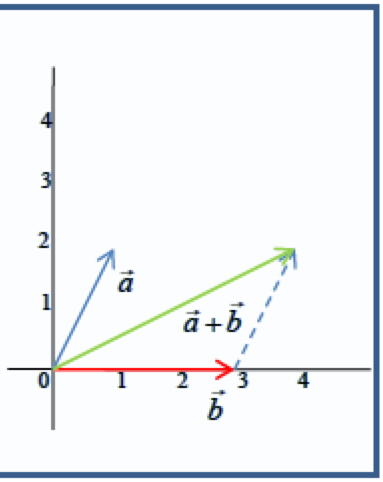


- Vectors. Vector length , addition, subtraction.

$$\vec{a} = (1,2)$$

$$\vec{b} = (3,0)$$

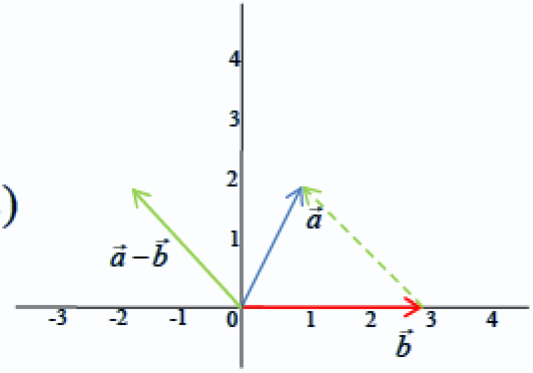
$$\vec{a} + \vec{b} = (4,2)$$



$$\vec{a} = (1,2)$$

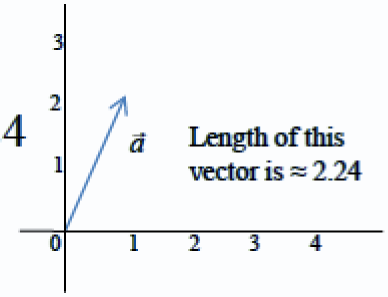
$$\vec{b} = (3,0)$$

$$\vec{a} - \vec{b} = (-2,2)$$

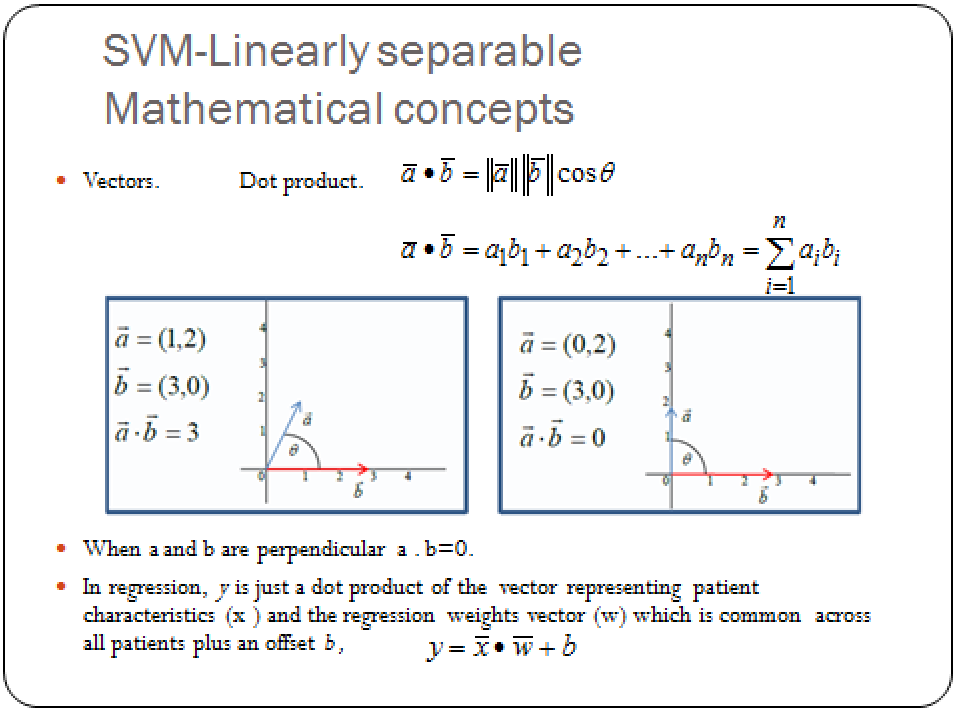


$$\vec{a} = (1,2)$$

$$\|\vec{a}\|_2 = \sqrt{5} \approx 2.24$$



Length of this vector is ≈ 2.24



## SVM-Linearly separable Mathematical concepts

- Vectors. Dot product. $\bar{a} \bullet \bar{b} = \|\bar{a}\| \|\bar{b}\| \cos\theta$

$$\bar{a} \bullet \bar{b} = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n = \sum_{i=1}^{n} a_i b_i$$

$\vec{a} = (1,2)$

$\vec{b} = (3,0)$

$\vec{a} \cdot \vec{b} = 3$

0
1
2
3
4
$\vec{a}$
$\theta$
$\vec{b}$

$\vec{a} = (0,2)$

$\vec{b} = (3,0)$

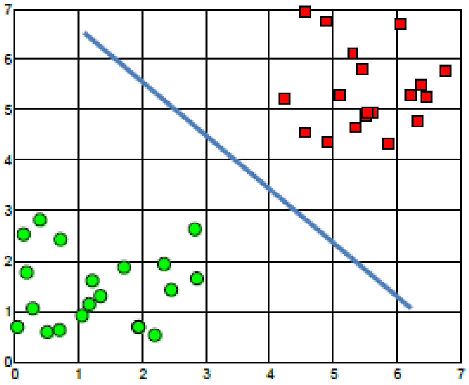
$\vec{a} \cdot \vec{b} = 0$

0
1
2
3
4
$\vec{a}$
$\theta$
$\vec{b}$

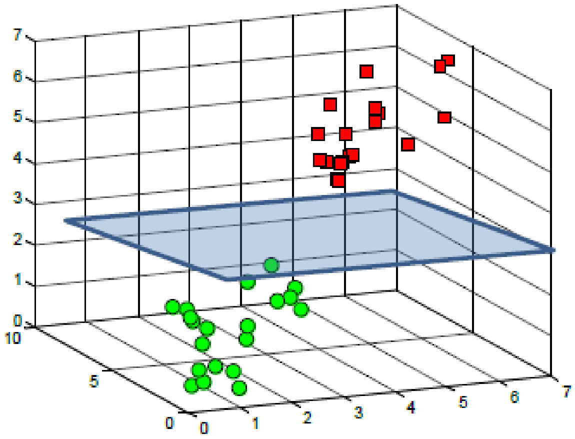
- When a and b are perpendicular a . b=0.
- In regression, $y$ is just a dot product of the vector representing patient characteristics (x ) and the regression weights vector (w) which is common across all patients plus an offset $b$, $y = \bar{x} \bullet \bar{w} + b$

- A hyperplane is a linear decision surface(binary classifier) that splits the space into two parts.

A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane



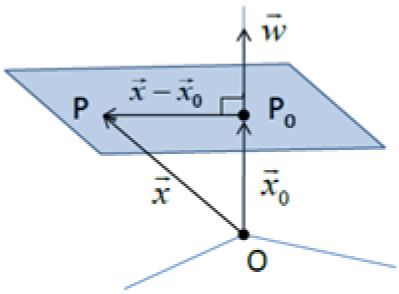
## SVM-Linearly separable Mathematical concepts-Hyperplanes

- An equation of a hyperplane is defined by a point ($P_0$) and a perpendicular vector to the plane (w ) at that point.
- Define vectors: $\overline{x}_0 = \overline{OP_0}$ *and* $\overline{x} = \overline{OP}$, where *P* is an arbitrary point on the hyperplane.
- A condition for *P* to be on the plane is that the vector $\overline{x} - \overline{x}_0$ is perpendicular to $\overline{w}$

$\vec{w}\cdot(\vec{x}-\vec{x}_0)=0$ or

$\vec{w}\cdot\vec{x}-\vec{w}\cdot\vec{x}_0=0$ define $b=-\vec{w}\cdot\vec{x}_0$

$$\boxed{\vec{w}\cdot\vec{x}+b=0}$$



## SVM-Linearly separable Mathematical concepts-Hyperplanes

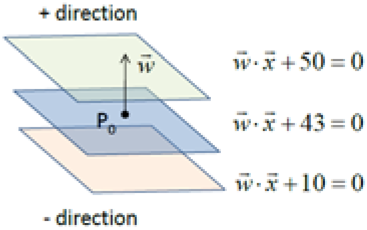
- What happens if the *b* coefficient changes?
- The hyperplane moves along the direction of $\overline{w}$ .
- We obtain "parallel hyperplanes".

$\vec{w}=(4,-1,6)$

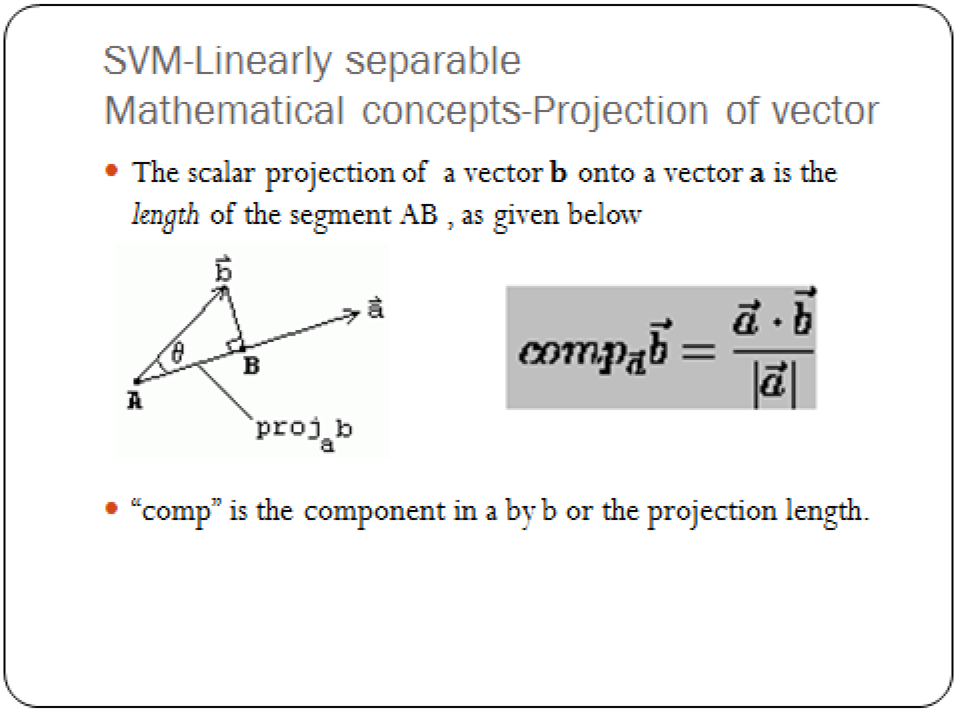
$P_0=(0,1,-7)$

$b=-\vec{w}\cdot\vec{x}_0=-(0-1-42)=43$

$$\boxed{\begin{aligned}&\Rightarrow \vec{w}\cdot\vec{x}+43=0\\&\Rightarrow (4,-1,6)\cdot\vec{x}+43=0\\&\Rightarrow (4,-1,6)\cdot(x_{(1)},x_{(2)},x_{(3)})+43=0\\&\Rightarrow 4x_{(1)}-x_{(2)}+6x_{(3)}+43=0\end{aligned}}$$


+ direction
$\vec{w}$
$P_0$
$\vec{w}\cdot\vec{x}+50=0$
$\vec{w}\cdot\vec{x}+43=0$
$\vec{w}\cdot\vec{x}+10=0$
- direction

- Distance between two parallel hyperplanes $\overline{x}\bullet\overline{w}+b_1=0$ *and* $\overline{x}\bullet\overline{w}+b_2=0$
- is equal to $D = |b_1 - b_2| \big/ \|\overline{w}\|$



## SVM-Linearly separable
## Mathematical concepts-Projection of vector

- The scalar projection of a vector **b** onto a vector **a** is the *length* of the segment AB , as given below

$$comp_{\vec{a}}\vec{b} = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|}$$

- “comp” is the component in a by b or the projection length.

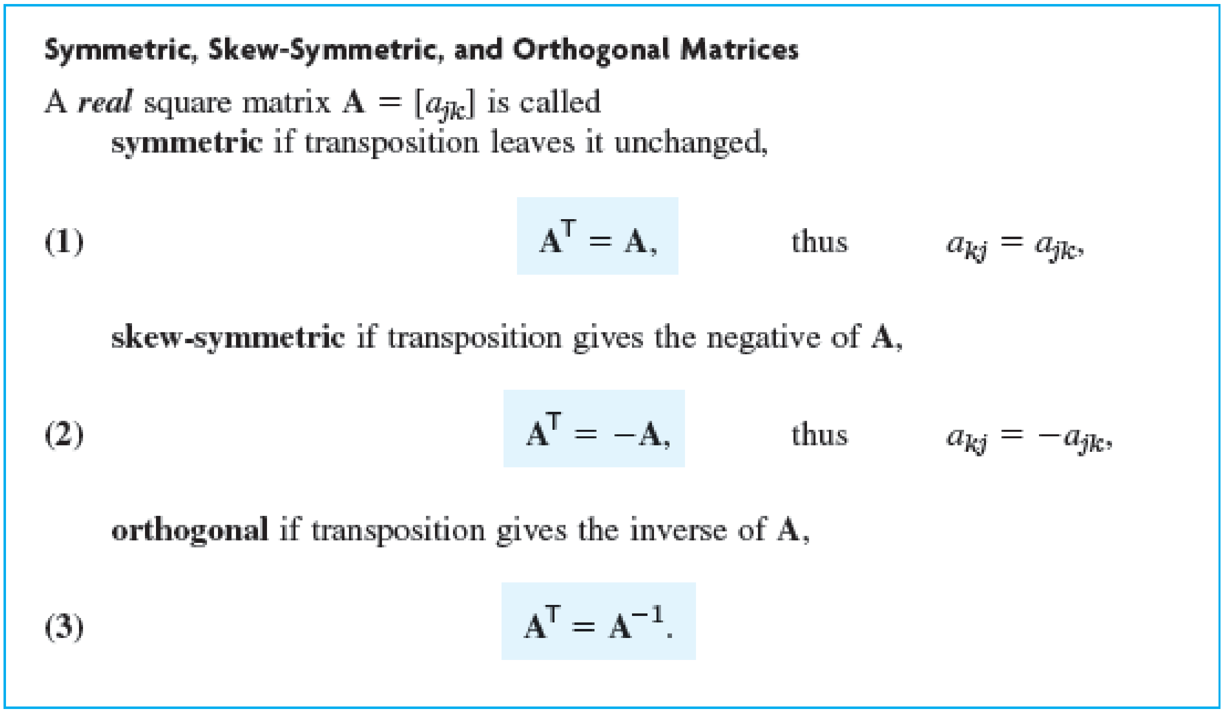
Types of matrices:

**Diagonal Matrix D. Scalar Matrix S. Unit Matrix I**

$$\mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} c & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & c \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

**Upper and Lower Triangular Matrices**

$$\underbrace{\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 4 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 6 \end{bmatrix},}_{\text{Upper triangular}} \quad \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 8 & -1 & 0 \\ 7 & 6 & 8 \end{bmatrix}, \begin{bmatrix} 3 & 0 & 0 & 0 \\ 9 & -3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 9 & 3 & 6 \end{bmatrix}.}_{\text{Lower triangular}}$$



### Symmetric, Skew-Symmetric, and Orthogonal Matrices

A ***real*** square matrix $\mathbf{A} = [a_{jk}]$ is called

**symmetric** if transposition leaves it unchanged,

(1) $$\mathbf{A}^\mathsf{T} = \mathbf{A}, \qquad \text{thus} \qquad a_{kj} = a_{jk},$$

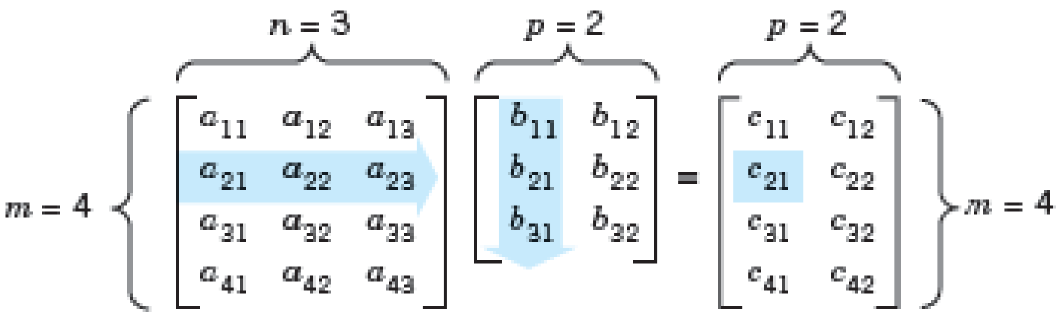
**skew-symmetric** if transposition gives the negative of $\mathbf{A}$,

(2) $$\mathbf{A}^\mathsf{T} = -\mathbf{A}, \qquad \text{thus} \qquad a_{kj} = -a_{jk},$$

**orthogonal** if transposition gives the inverse of $\mathbf{A}$,

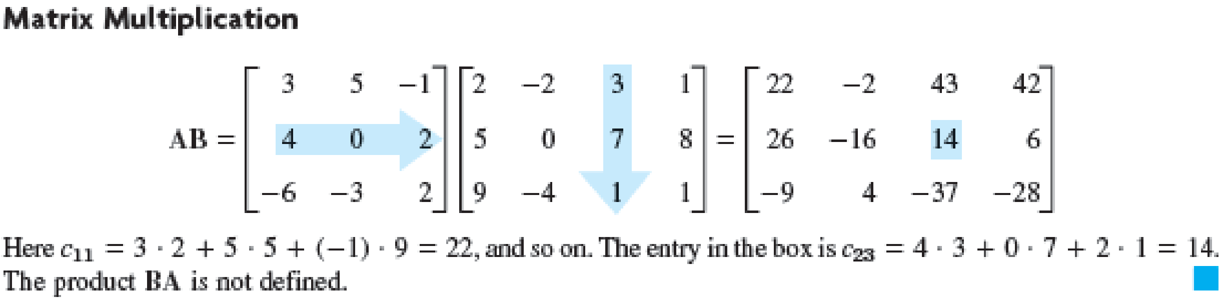
(3) $$\mathbf{A}^\mathsf{T} = \mathbf{A}^{-1}.$$

Operations on matrices:



$$m = 4\left\{\overbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix}}^{n=3} \overbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}}^{p=2} = \overbrace{\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \end{bmatrix}}^{p=2}\right\} m = 4$$

Notations in a product **AB** = **C**



### Matrix Multiplication

$$\mathbf{AB} = \begin{bmatrix} 3 & 5 & -1 \\ 4 & 0 & 2 \\ -6 & -3 & 2 \end{bmatrix} \begin{bmatrix} 2 & -2 & 3 & 1 \\ 5 & 0 & 7 & 8 \\ 9 & -4 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 22 & -2 & 43 & 42 \\ 26 & -16 & \boxed{14} & 6 \\ -9 & 4 & -37 & -28 \end{bmatrix}$$

Here $c_{11} = 3 \cdot 2 + 5 \cdot 5 + (-1) \cdot 9 = 22$, and so on. The entry in the box is $c_{23} = 4 \cdot 3 + 0 \cdot 7 + 2 \cdot 1 = 14$. The product **BA** is not defined. ■

Product is not commutative:

$$\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix} = \begin{bmatrix} 99 & 99 \\ -99 & -99 \end{bmatrix}.$$

(a) $(k\mathbf{A})\mathbf{B} = k(\mathbf{AB}) = \mathbf{A}(k\mathbf{B})$ *written* $k\mathbf{AB}$ *or* $\mathbf{A}k\mathbf{B}$

(b) $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$ *written* $\mathbf{ABC}$

(c) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

(d) $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$

Transposition:

$$\mathbf{A} = \begin{bmatrix} 5 & -8 & 1 \\ 4 & 0 & 0 \end{bmatrix}, \quad \text{then} \quad \mathbf{A}^{\mathsf{T}} = \begin{bmatrix} 5 & 4 \\ -8 & 0 \\ 1 & 0 \end{bmatrix}.$$

(a) $(\mathbf{A}^{\mathsf{T}})^{\mathsf{T}} = \mathbf{A}$

(b) $(\mathbf{A} + \mathbf{B})^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} + \mathbf{B}^{\mathsf{T}}$

(c) $(c\mathbf{A})^{\mathsf{T}} = c\mathbf{A}^{\mathsf{T}}$

(d) $(\mathbf{AB})^{\mathsf{T}} = \mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}.$

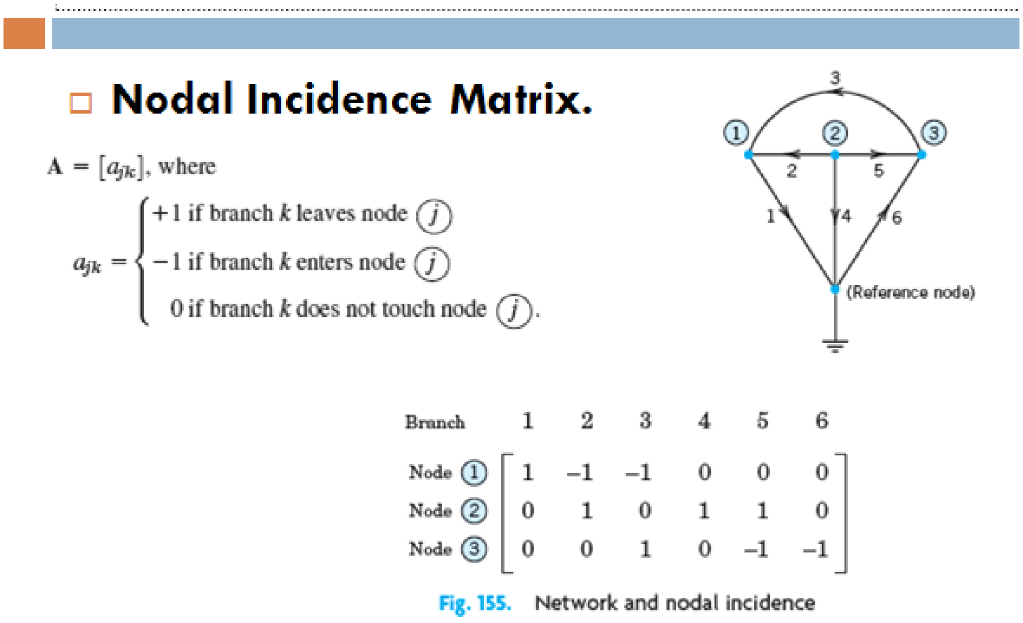
Ex:

- □ Sales figures for three products I, II, III in a store on Monday (Mon), Tuesday (Tues), may for each week…be arranged in a matrix

$$\begin{array}{cc} & \begin{array}{ccccccc}\text{Mon} & \text{Tues} & \text{Wed} & \text{Thur} & \text{Fri} & \text{Sat} & \text{Sun}\end{array} \\ \mathbf{A} = & \left[\begin{array}{ccccccc} 40 & 33 & 81 & 0 & 21 & 47 & 33 \\ 0 & 12 & 78 & 50 & 50 & 96 & 90 \\ 10 & 0 & 0 & 27 & 43 & 78 & 56 \end{array}\right] \begin{array}{c} \text{I} \\ \cdot\ \text{II} \\ \text{III} \end{array} \end{array}$$

Ex2:

- □ **Nodal Incidence Matrix. The network in picture** consists of six *branches (connections) and four nodes* (points where two or more branches come together). One node is the *reference node (grounded node, whose* voltage is zero). We number the other nodes and number and direct the branches. This we do arbitrarily. The network can now be described by a matrix. **A is called the *nodal incidence matrix of the network.*** Show that for the network in Fig. the matrix **A has** the given form.



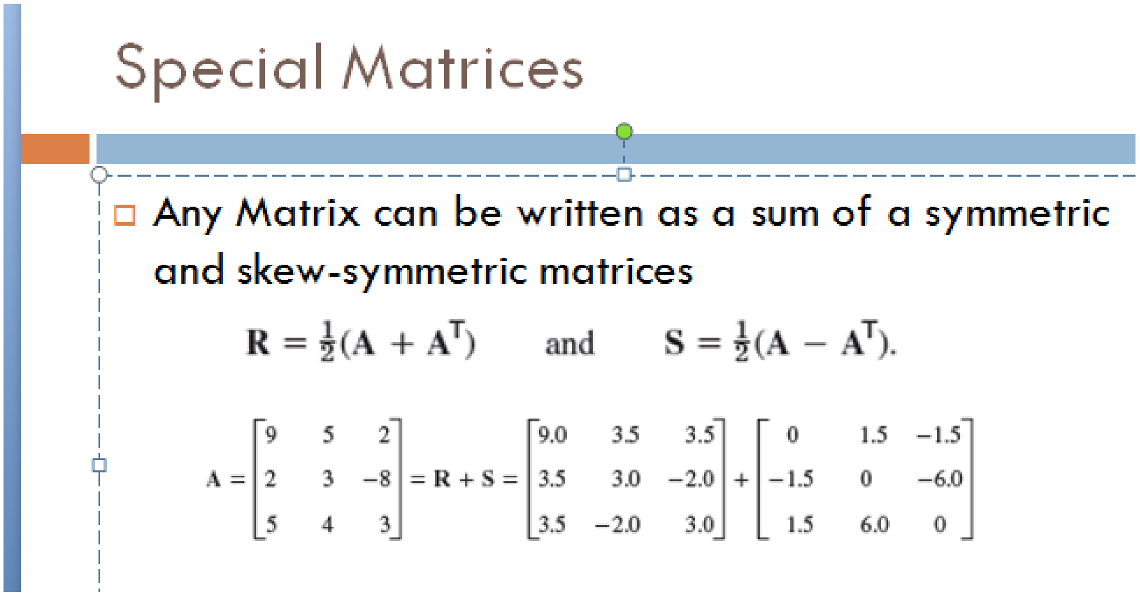
## □ Nodal Incidence Matrix.

$\mathbf{A} = [a_{jk}]$, where

$$a_{jk} = \begin{cases} +1 \text{ if branch } k \text{ leaves node } \textcircled{j} \\ -1 \text{ if branch } k \text{ enters node } \textcircled{j} \\ 0 \text{ if branch } k \text{ does not touch node } \textcircled{j}. \end{cases}$$

$$\begin{array}{cc} \text{Branch} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} \text{Node } \textcircled{1} \\ \text{Node } \textcircled{2} \\ \text{Node } \textcircled{3} \end{array} & \left[\begin{array}{cccccc} 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 \end{array}\right] \end{array}$$

Fig. 155. Network and nodal incidence



## Special Matrices

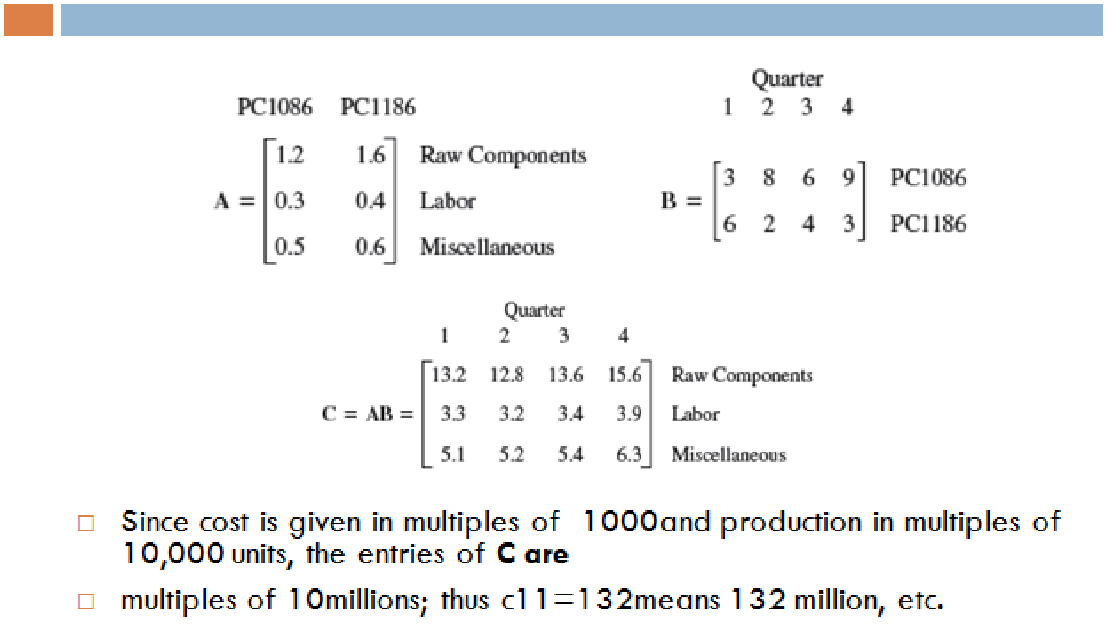
- Any Matrix can be written as a sum of a symmetric and skew-symmetric matrices

$$\mathbf{R} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathsf{T}}) \qquad \text{and} \qquad \mathbf{S} = \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^{\mathsf{T}}).$$

$$\mathbf{A} = \begin{bmatrix} 9 & 5 & 2 \\ 2 & 3 & -8 \\ 5 & 4 & 3 \end{bmatrix} = \mathbf{R} + \mathbf{S} = \begin{bmatrix} 9.0 & 3.5 & 3.5 \\ 3.5 & 3.0 & -2.0 \\ 3.5 & -2.0 & 3.0 \end{bmatrix} + \begin{bmatrix} 0 & 1.5 & -1.5 \\ -1.5 & 0 & -6.0 \\ 1.5 & 6.0 & 0 \end{bmatrix}$$

Matrix multiplication:

□ Supercomp Ltd produces two computer models PC1086 and PC1186. The matrix **A shows the cost per computer**(in thousands of dollars) and **B the production figures for the year 2010 (in multiples of 10,000 units.) Find a**matrix **C that shows the shareholders the cost per quarter (in millions of dollars) for raw material, labor, and**miscellaneous.

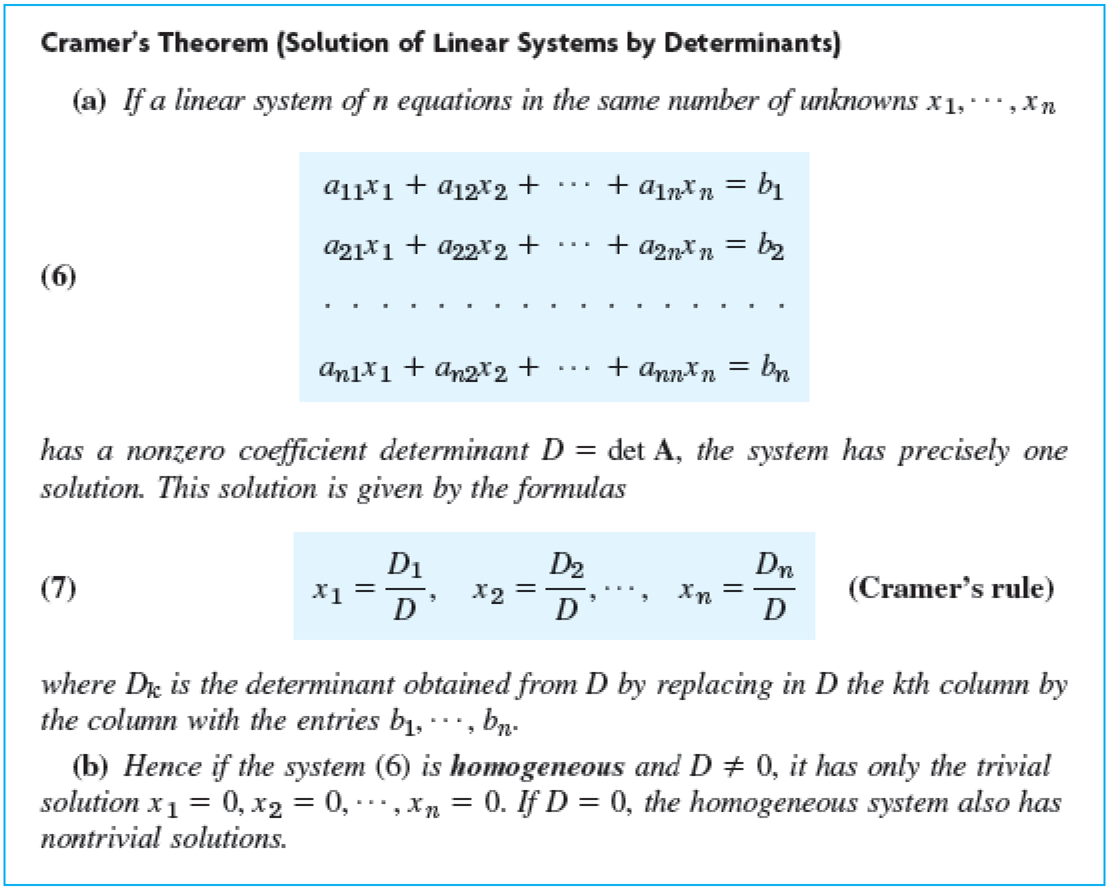


$$\mathbf{A} = \begin{array}{cc} \text{PC1086} & \text{PC1186} \\ \left[\begin{matrix} 1.2 & 1.6 \\ 0.3 & 0.4 \\ 0.5 & 0.6 \end{matrix}\right. & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right] \end{array} \begin{array}{l} \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array} \qquad \mathbf{B} = \begin{array}{c} \text{Quarter} \\ \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 3 & 8 & 6 & 9 \\ 6 & 2 & 4 & 3 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \text{PC1086} \\ \text{PC1186} \end{array}$$

$$\mathbf{C} = \mathbf{AB} = \begin{array}{c} \text{Quarter} \\ \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 13.2 & 12.8 & 13.6 & 15.6 \\ 3.3 & 3.2 & 3.4 & 3.9 \\ 5.1 & 5.2 & 5.4 & 6.3 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array}$$

- Since cost is given in multiples of  1000and production in multiples of 10,000 units, the entries of **C are**
- multiples of 10millions; thus c11=132means 132 million, etc.

Cramer’s rule:



**Cramer's Theorem (Solution of Linear Systems by Determinants)**

**(a)** *If a linear system of n equations in the same number of unknowns* $x_1, \cdots, x_n$

**(6)**
$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots&\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

*has a nonzero coefficient determinant* $D = \det \mathbf{A}$, *the system has precisely one solution. This solution is given by the formulas*

**(7)**
$$x_1 = \frac{D_1}{D}, \quad x_2 = \frac{D_2}{D}, \cdots, \quad x_n = \frac{D_n}{D} \qquad \textbf{(Cramer's rule)}$$

*where* $D_k$ *is the determinant obtained from* $D$ *by replacing in* $D$ *the kth column by the column with the entries* $b_1, \cdots, b_n$.

**(b)** *Hence if the system* (6) *is* ***homogeneous*** *and* $D \neq 0$, *it has only the trivial solution* $x_1 = 0, x_2 = 0, \cdots, x_n = 0$. *If* $D = 0$, *the homogeneous system also has nontrivial solutions.*

## Linear system of equations:

- □ Existence of solution: Coefficient matrix A and Augmented matrix have the same rank.
- □ Uniqueness: If the common rank is n, the order of the matrix.
- □ If common rank r<n, infinite solutions exist.
- □ Now we can use methods like Gauss-Elimination to obtain them.

# Calculus:

Limits:

consider the functions

$$f(x) = \frac{x^2 - 4}{x - 2} \quad \text{and} \quad g(x) = \frac{x^2 - 5}{x - 2}.$$

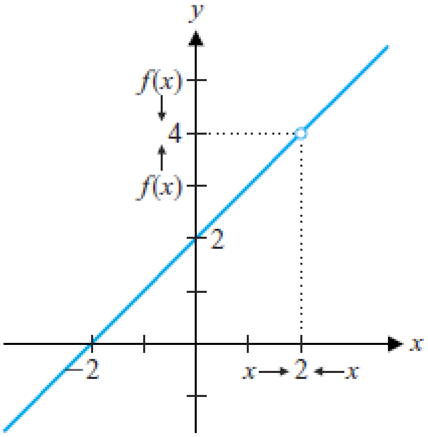


FIGURE 1.7a

$y = \dfrac{x^2 - 4}{x - 2}$

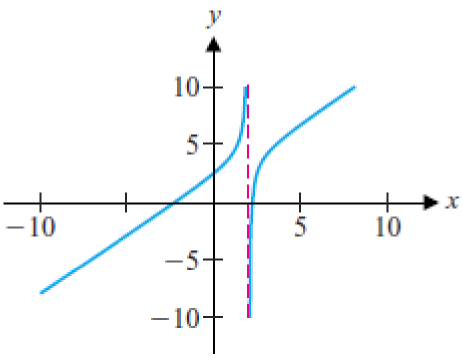


FIGURE 1.7b

$y = \dfrac{x^2 - 5}{x - 2}$

| $x$ | $f(x) = \dfrac{x^2 - 4}{x - 2}$ |
|---|---|
| 1.9 | 3.9 |
| 1.99 | 3.99 |
| 1.999 | 3.999 |
| 1.9999 | 3.9999 |

| $x$ | $f(x) = \dfrac{x^2 - 4}{x - 2}$ |
|---|---|
| 2.1 | 4.1 |
| 2.01 | 4.01 |
| 2.001 | 4.001 |
| 2.0001 | 4.0001 |

Notice that as you move down the first column of the table, the $x$-values get closer to 2, but are all less than 2. We use the notation $x \to 2^-$ to indicate that $x$ *approaches 2 from the left side*. Notice that the table and the graph both suggest that as $x$ gets closer and closer to 2 (with $x < 2$), $f(x)$ is getting closer and closer to 4. In view of this, we say that the **limit of $f(x)$ as $x$ approaches 2 from the left** is 4, written

$$\lim_{x \to 2^-} f(x) = 4.$$

Likewise, we need to consider what happens to the function values for $x$ close to 2 but larger than 2. Here, we use the notation $x \to 2^+$ to indicate that $x$ *approaches 2 from the right side*. We compute some of these values in the second table.

Again, the table and graph both suggest that as $x$ gets closer and closer to 2 (with $x > 2$), $f(x)$ is getting closer and closer to 4. In view of this, we say that the **limit of $f(x)$ as $x$ approaches 2 from the right** is 4, written

$$\lim_{x \to 2^+} f(x) = 4.$$

We call $\lim_{x \to 2^-} f(x)$ and $\lim_{x \to 2^+} f(x)$ **one-sided limits.** Since the two one-sided limits of $f(x)$ are the same, we summarize our results by saying that the **limit of $f(x)$ as $x$ approaches 2** is 4, written

$$\lim_{x \to 2} f(x) = 4.$$

The notion of limit as we have described it here is intended to communicate the behavior of a function *near* some point of interest, but not actually *at* that point. We finally observe that we can also determine this limit algebraically, as follows. Notice that since the expression in the numerator of $f(x) = \dfrac{x^2 - 4}{x - 2}$ factors, we can write

$$\begin{aligned} \lim_{x \to 2} f(x) &= \lim_{x \to 2} \frac{x^2 - 4}{x - 2} \\ &= \lim_{x \to 2} \frac{(x - 2)(x + 2)}{x - 2} && \text{Cancel the factors of } (x - 2). \\ &= \lim_{x \to 2} (x + 2) = 4, && \text{As } x \text{ approaches 2, } (x + 2) \text{ approaches 4.} \end{aligned}$$

where we can cancel the factors of $(x - 2)$ since in the limit as $x \to 2$, $x$ is *close* to 2, but $x \neq 2$, so that $x - 2 \neq 0$.

A limit exists if and only if both corresponding one-sided limits exist and are equal. That is,

$$\lim_{x \to a} f(x) = L, \text{ for some number } L, \text{ if and only if } \lim_{x \to a^-} f(x) = \lim_{x \to a^+} f(x) = L.$$

### THEOREM 3.1

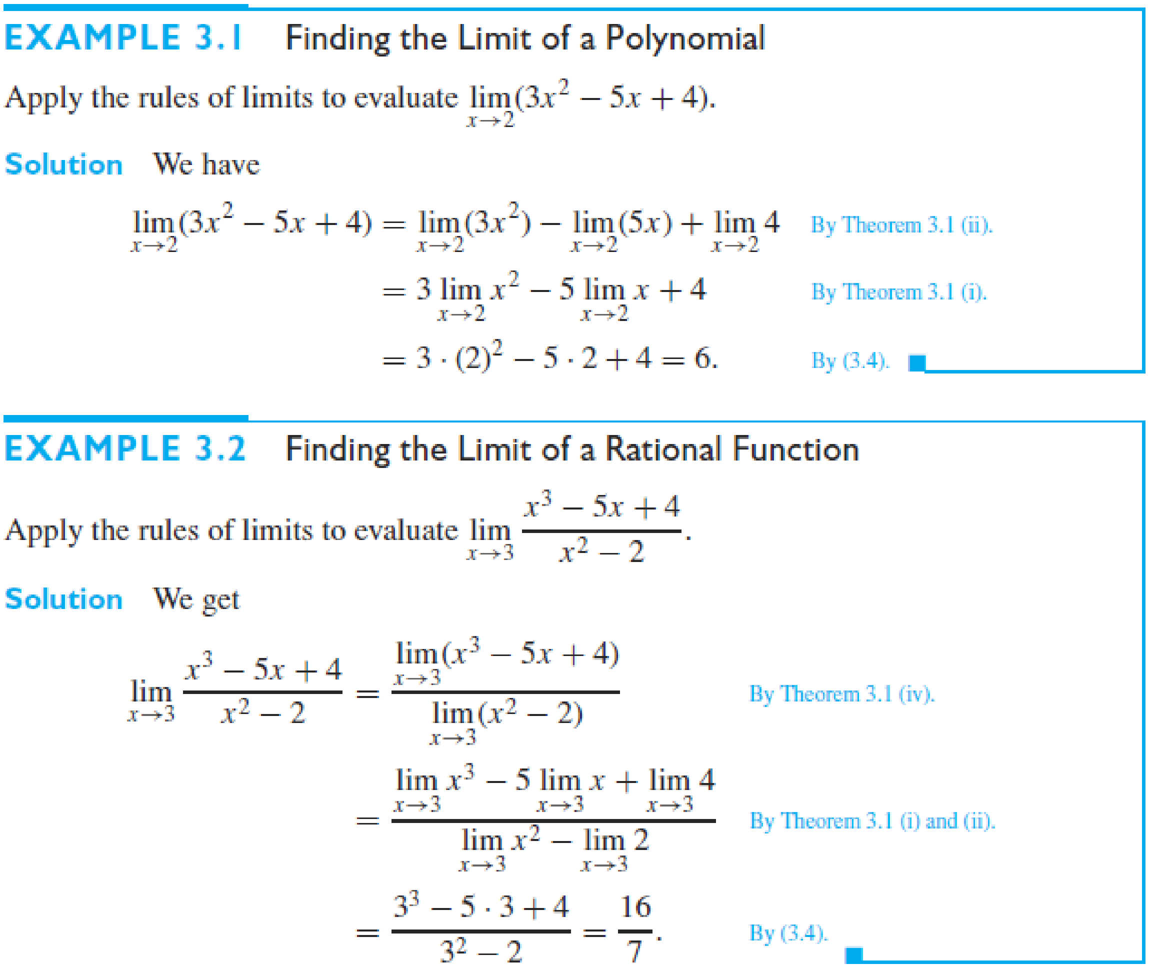
Suppose that $\lim_{x \to a} f(x)$ and $\lim_{x \to a} g(x)$ both exist and let $c$ be any constant. The following then apply:

(i) $\lim_{x \to a} [c \cdot f(x)] = c \cdot \lim_{x \to a} f(x)$,

(ii) $\lim_{x \to a} [f(x) \pm g(x)] = \lim_{x \to a} f(x) \pm \lim_{x \to a} g(x)$,

(iii) $\lim_{x \to a} [f(x) \cdot g(x)] = \left[\lim_{x \to a} f(x)\right]\left[\lim_{x \to a} g(x)\right]$ and

(iv) $\lim_{x \to a} \dfrac{f(x)}{g(x)} = \dfrac{\lim_{x \to a} f(x)}{\lim_{x \to a} g(x)}$ $\left(\text{if } \lim_{x \to a} g(x) \neq 0\right)$.



### EXAMPLE 3.1 Finding the Limit of a Polynomial

Apply the rules of limits to evaluate $\lim_{x\to 2}(3x^2 - 5x + 4)$.

**Solution** We have

$$\begin{aligned}\lim_{x\to 2}(3x^2 - 5x + 4) &= \lim_{x\to 2}(3x^2) - \lim_{x\to 2}(5x) + \lim_{x\to 2} 4 && \text{By Theorem 3.1 (ii).}\\ &= 3\lim_{x\to 2} x^2 - 5\lim_{x\to 2} x + 4 && \text{By Theorem 3.1 (i).}\\ &= 3\cdot(2)^2 - 5\cdot 2 + 4 = 6. && \text{By (3.4).}\end{aligned}$$

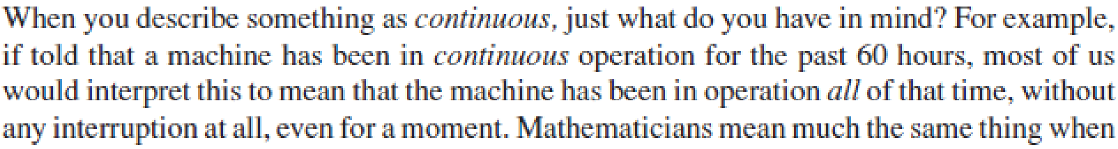
### EXAMPLE 3.2 Finding the Limit of a Rational Function

Apply the rules of limits to evaluate $\lim_{x\to 3}\dfrac{x^3 - 5x + 4}{x^2 - 2}$.

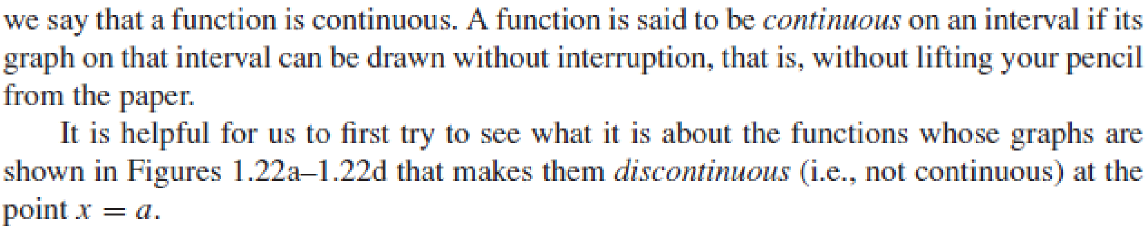
**Solution** We get

$$\begin{aligned}\lim_{x\to 3}\frac{x^3 - 5x + 4}{x^2 - 2} &= \frac{\lim_{x\to 3}(x^3 - 5x + 4)}{\lim_{x\to 3}(x^2 - 2)} && \text{By Theorem 3.1 (iv).}\\ &= \frac{\lim_{x\to 3} x^3 - 5\lim_{x\to 3} x + \lim_{x\to 3} 4}{\lim_{x\to 3} x^2 - \lim_{x\to 3} 2} && \text{By Theorem 3.1 (i) and (ii).}\\ &= \frac{3^3 - 5\cdot 3 + 4}{3^2 - 2} = \frac{16}{7}. && \text{By (3.4).}\end{aligned}$$

## Continuity:



When you describe something as *continuous,* just what do you have in mind? For example, if told that a machine has been in *continuous* operation for the past 60 hours, most of us would interpret this to mean that the machine has been in operation *all* of that time, without any interruption at all, even for a moment. Mathematicians mean much the same thing when



we say that a function is continuous. A function is said to be *continuous* on an interval if its graph on that interval can be drawn without interruption, that is, without lifting your pencil from the paper.

It is helpful for us to first try to see what it is about the functions whose graphs are shown in Figures 1.22a–1.22d that makes them *discontinuous* (i.e., not continuous) at the point $x = a$.

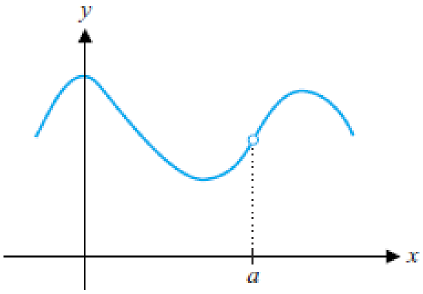


**FIGURE 1.22a**
$f(a)$ is not defined (the graph has a hole at $x = a$).

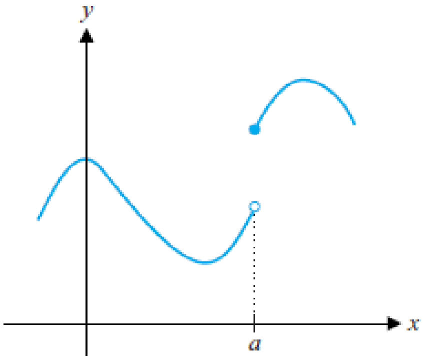


**FIGURE 1.22b**
$f(a)$ is defined, but $\lim_{x\to a} f(x)$ does not exist (the graph has a jump at $x = a$).

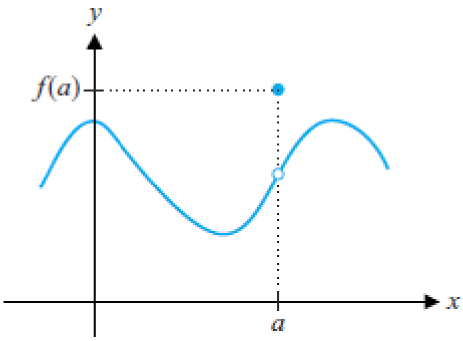


**FIGURE 1.22c**
$\lim_{x\to a} f(x)$ exists and $f(a)$ is defined, but $\lim_{x\to a} f(x) \neq f(a)$ (the graph has a hole at $x = a$).

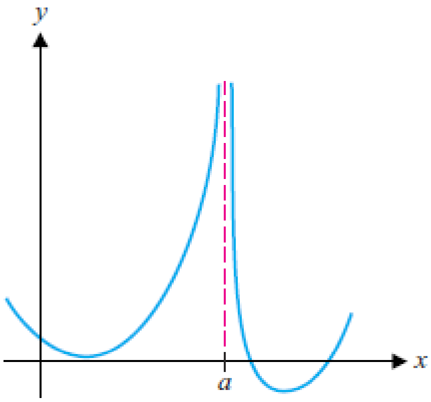


**FIGURE 1.22d**
$\lim_{x\to a} f(x)$ does not exist (the function "blows up" at $x = a$).

### DEFINITION 4.1

A function $f$ is **continuous** at $x = a$ when
(i) $f(a)$ is defined, (ii) $\lim_{x\to a} f(x)$ exists and (iii) $\lim_{x\to a} f(x) = f(a)$.
Otherwise, $f$ is said to be **discontinuous** at $x = a$.

Determine where $f(x) = \dfrac{x^2 + 2x - 3}{x - 1}$ is continuous.

**Solution** Note that

$$\begin{aligned} f(x) &= \frac{x^2 + 2x - 3}{x - 1} = \frac{(x-1)(x+3)}{x-1} && \text{Factoring the numerator.} \\ &= x + 3, \text{ for } x \neq 1. && \text{Canceling common factors.} \end{aligned}$$

This says that the graph of $f$ is a straight line, but with a hole in it at $x = 1$, as indicated in Figure 1.23. So, $f$ is discontinuous at $x = 1$, but continuous elsewhere. ■

### THEOREM 4.2

Suppose that $f$ and $g$ are continuous at $x = a$. Then all of the following are true:

(i) $(f \pm g)$ is continuous at $x = a$,
(ii) $(f \cdot g)$ is continuous at $x = a$ and
(iii) $(f/g)$ is continuous at $x = a$ if $g(a) \neq 0$.

### DEFINITION 4.2

If $f$ is continuous at every point on an open interval $(a, b)$, we say that $f$ is **continuous on** $(a, b)$**.** Following Figure 1.27, we say that $f$ is **continuous on the closed interval** $[a, b]$, if $f$ is continuous on the open interval $(a, b)$ and

$$\lim_{x \to a^+} f(x) = f(a) \quad \text{and} \quad \lim_{x \to b^-} f(x) = f(b).$$

Finally, if $f$ is continuous on all of $(-\infty, \infty)$, we simply say that $f$ is **continuous.** (That is, when we don't specify an interval, we mean continuous everywhere.)

# 1. Partial Derivatives

The derivative of a function of one variable, such as $y(x)$, tells us the gradient of the function: how $y$ changes when $x$ increases. If we have a function of more than one variable, such as:

$$z(x,y) = x^3 + 4xy + 5y^2$$

we can ask, for example, how $z$ changes when $x$ increases but $y$ doesn't change. The answer to this question is found by thinking of $z$ as a function of $x$, and differentiating, treating $y$ as if it were a constant parameter:

$$\frac{\partial z}{\partial x} = 3x^2 + 4y$$

This process is called partial differentiation. We write $\frac{\partial z}{\partial x}$ rather than $\frac{dz}{dx}$, to emphasize that $z$ is a function of another variable as well as $x$, which is being held constant.

$$\boxed{\frac{\partial z}{\partial x} \text{ is called the partial derivative of } z \text{ with respect to } x}$$

Exercises 7.1: Find the partial derivatives with respect to $x$ and $y$ of the functions:

(1) $f(x,y) = 3x^2 - xy^4$

(3) $g(x,y) = \dfrac{\ln x}{y}$

(2) $h(x,y) = (x+1)^2(y+2)$

## 1.1. Second-order Partial Derivatives

For the function in the previous section:

$$z(x,y) = x^3 + 4xy + 5y^2$$

we found:

$$\frac{\partial z}{\partial x} = 3x^2 + 4y$$
$$\frac{\partial z}{\partial y} = 4x + 10y$$

These are the *first-order* partial derivatives. But we can differentiate again to find *second-order* partial derivatives. The second derivative with respect to $x$ tells us how $\frac{\partial z}{\partial x}$ changes as $x$ increases, still keeping $y$ constant.

$$\frac{\partial^2 z}{\partial x^2} = 6x$$

Similarly:

$$\frac{\partial^2 z}{\partial y^2} = 10$$

# 2. Economic Applications of Partial Derivatives, and Euler's Theorem

## 2.1. The Marginal Products of Labour and Capital

Suppose that the output produced by a firm depends on the amounts of labour and capital used. If the production function is

$$Y(K, L)$$

the partial derivative of $Y$ with respect to $L$ tells us the the marginal product of labour:

$$MPL = \frac{\partial Y}{\partial L}$$

The marginal product of labour is the amount of extra output the firm could produce if it used one extra unit of labour, but kept capital the same as before.
Similarly the marginal product of capital is:

$$MPK = \frac{\partial Y}{\partial K}$$

Examples 2.1: For a firm with production function $Y(K, L) = 5K^{\frac{1}{3}}L^{\frac{2}{3}}$:

(i) Find the marginal product of labour.

$$MPL = \frac{\partial Y}{\partial L} = \frac{10}{3}K^{\frac{1}{3}}L^{-\frac{1}{3}}$$

(ii) What is the MPL when $K = 64$ and $L = 125$?

$$MPL = \frac{10}{3}(64)^{\frac{1}{3}}(125)^{-\frac{1}{3}} = \frac{10}{3} \times 4 \times \frac{1}{5} = \frac{8}{3}$$

(iii) What happens to the marginal product of labour as the number of workers increases?

Differentiate MPL with respect to $L$: $\quad \dfrac{\partial^2 Y}{\partial L^2} = -\dfrac{10}{9}K^{\frac{1}{3}}L^{-\frac{4}{3}} < 0$

So the MPL decreases as the labour input increases – the firm has diminishing returns to labour, if capital is held constant. This is true for *all* values of $K$ and $L$.

# 5. The Chain Rule and Implicit Differentiation

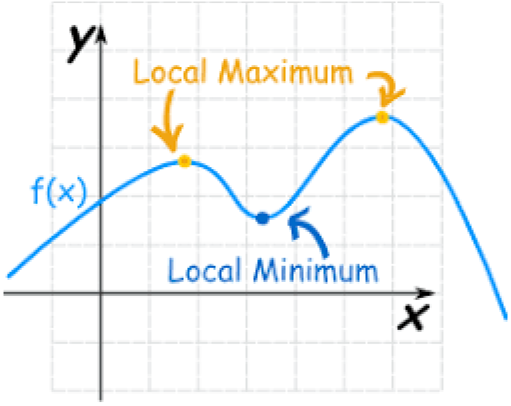
## 5.1. The Chain Rule for Functions of Several Variables

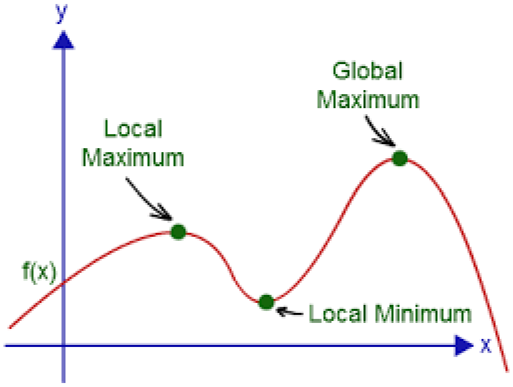
If $z$ is a function of two variables, $x$ and $y$, and both $x$ and $y$ depend on another variable, $t$ (time, for example), then $z$ also depends on $t$. We have:

If $z = z(x, y)$, and $x$ and $y$ are functions of $t$, then:

$$\frac{dz}{dt} = \frac{\partial z}{\partial x}\frac{dx}{dt} + \frac{\partial z}{\partial y}\frac{dy}{dt}$$

**Maxima-Minima:**



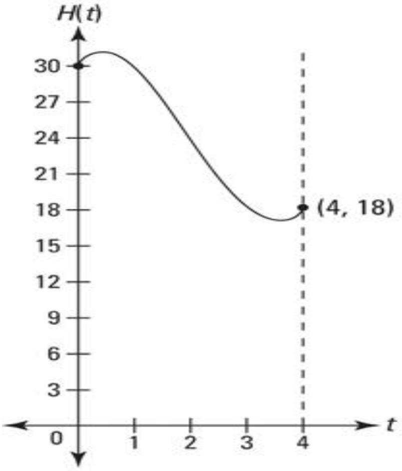


Procedure:

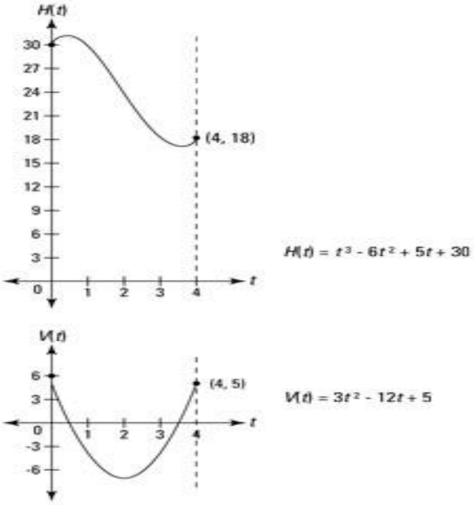
Step1: Find the first derivative of f(x), equate it to zero and solve.

Step2: Substitute the points obtained in step 1 on the second derivative. If value is –ve, then the point is max and if +ve, then the point is minimum.

Example: A yo-yo moves straight up and down. Its height above the ground, as a function of time, is given by the function below, where *t* is in seconds and *H*(*t*) is in inches. At *t* = 0, it's 30 inches above the ground, and after 4 seconds, it's at height of 18 inches.



$H(t) = t^3 - 6t^2 + 5t + 30$



- **Velocity,** *V*(*t*) is the derivative of position (height, in this problem) and acceleration, *A*(*t*), is the derivative of velocity. Thus V(t)=3t$^2$-12t+5
- 
- **Maximum and minimum height** of *H*(*t*) occur at the local extrema you see in the above figure. To locate them, set the derivative of *H*(*t*) — that's *V*(*t*) — equal to zero and solve.
- These are the times when the yo-yo reaches its maximum and minimum heights. Plug these numbers into *H*(*t*) to obtain the heights:
- *H* (0.47) ≈ 31.1
- *H* (3.53) ≈ 16.9
- So, the yo-yo gets as high as about 31.1 inches above the ground at t ≈ 0.47 seconds and as low as about 16.9 inches at *t* ≈ 3.53 seconds.
- **Maximum and minimum velocity** of the yo-yo during the interval from 0 to 4 seconds are determined with the derivative of *V*(*t*): Set the derivative of *V*(*t*) — that's *A*(*t*) — equal to zero and solve:
- V'(t)=A(t). Hence A(t)=6t-12, which when equated to zero, we obtain t=2.
- Now, evaluate *V*(*t*) at the critical number, 2, and at the interval's endpoints, 0 and 4:

## Decision theory

- Broad types are
- Normative and Descriptive:
- A normative decision theory is a theory about how decisions should be made, and a descriptive theory is a theory about how decisions are actually made.
- Decision process:
- 1.Identification of the problem
- 2. Obtaining necessary information
- 3. Production of possible solutions
- 4. Evaluation of such solutions
- 5. Selection of a strategy for performance
- Suppose we have an input vector **x** together with a corresponding vector **t** of target variables, and our goal is to predict **t** given a new value for **x**.
- From probability perspective, we are talking about the joint distribution p(x,t).
- Determination of *p*(**x***,* **t**) from a set of training data is *inference.*
- Taking a specific action based on the predicted/expected values of t form *Decision theory*.

## Information theory

- The theory studying information gathered from known values of random variables is information theory.
- If X is a random variable with pmf p(x), then information is the quantity h(x) which is defined by Shannon as
- h(x) =-log2p(x).
- Low probability events *x* correspond to high information content.

- suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit in the process is obtained by taking the expectation of with respect to the distribution *p*(*x*) and is given by $H(x) = -\sum_{x} p(x) * \log_2 p(x)$
- called the Entropy of the random variable X.
- Consider a random variable *x* having 8 possible states, each of which is equally likely. In order to communicate the value of *x* to a receiver, we would need to transmit a message of length 3 bits. Notice that the entropy of this variable is given by
- H(x)=-8*[1/8]*log2[1/8]= 3 bits.

## Probability

## Basic Concepts

**Random Experiment**: An experiment is said to be a random experiment, if it's out-come can't be predicted with certainty.

Example; If a coin is tossed, we can't say, whether head or tail will appear. So it is a random experiment.

**Sample Space**: The set of all possible out-comes of an experiment is called the sample space. It is denoted by 'S' and its number of elements are n(s).

Example; In throwing a dice, the number that appears at top is any one of 1,2,3,4,5,6. So here:S ={1,2,3,4,5,6} and n(s) = 6

Similarly in the case of a coin, S={Head,Tail} or {H,T} and n(s)=2.

The elements of the sample space are called sample points or event points.

**Event**: Every subset of a sample space is an event. It is denoted by 'E'.

Example: In throwing a dice S={1,2,3,4,5,6}, the appearance of an event number will be the event E={2,4,6}.

Clearly E is a sub set of S.

**Simple event**; An event, consisting of a single sample point is called a simple event.

Example: In throwing a dice, S={1,2,3,4,5,6}, so each of {1},{2},{3},{4},{5} and {6} are simple events.

**Compound event**: A subset of the sample space, which has more than on element is called a mixed event.

Example: In throwing a dice, the event of appearing of odd numbers is a compound event, because E={1,3,5} which has '3' elements.

**Equally likely events**: Events are said to be equally likely, if we have no reason to believe that one is more likely to occur than the other.

Example: When a dice is thrown, all the six faces {1,2,3,4,5,6} are equally likely to come up.

**Exhaustive events**: When every possible out come of an experiment is considered.

Approaches of Probability

- □ Classical approach
- □ Frequency approach
- □ Subjective approach
- □ Axiomatic approach
    - ▣ P(A)>=0;
    - ▣ P(S)=1;
    - ▣ P(A+B)<=P(A)+P(B)

**Classical definition of probability***:*

If 'S' be the sample space, then the probability of occurrence of an event 'E' is defined as:

$$P(E) = n(E)/N(S) = \frac{\text{number of elements in 'E'}}{\text{number of elements in sample space 'S'}}$$

Example: Find the probability of getting a tail in tossing of a coin.

Solution:
Sample space S = {H,T} and n(s) = 2
Event 'E' = {T} and n(E) = 1
therefore P(E) = n(E)/n(S) = ½
Note: This definition is not true, if
(a) The events are not equally likely.
(b) The possible outcomes are infinite.

**Sure event**: Let 'S' be a sample space. If E is a subset of or equal to S then E is called a sure event.
Example: In a throw of a dice, S={1,2,3,4,5,6}
Let $E_1$=Event of getting a number less than '7'.
So '$E_1$' is a sure event.So, we can say, in a sure event n(E) = n(S)

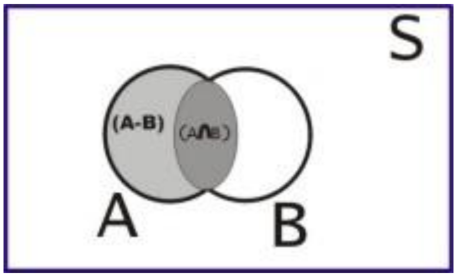
**Mutually exclusive or disjoint event**: If two or more events can't occur simultaneously, that is no two of them can occur together.

**Addition Theorem of Probability :**

If ‘A’ and ‘B’ by any two events, then the probability of occurrence of at least one of the events ‘A’ and ‘B’ is given by:

P(A or B) = P(A) + P(B) – P (A and B)

P(A∪B) = P(A) + P(B) – P (A∩B)



**Problems based on addition theorem of probability:**
Working rule :

(i) A ∪B denotes the event of occurrence of at least one of the event ‘A’ or ‘B’
(ii) A ∩B denotes the event of occurrence of both the events ‘A’ and ‘B’.
(iii) P(A∪B) or P(A+B) denotes the probability of occurrence of at least one of the event ‘A’ or ‘B’.
(iv) P(∩B) or P(AB) denotes the probability of occurrence of both the event ‘A’ and ‘B’.

--------------------x-------------------x---------------------x--------------------x---------------x-----

Ex.: The probability that a contractor will get a contract is '2/3' and the probability that he will get on other contract is 5/9 . If the probability of getting at least one contract is 4/5, what is the probability that he will get both the contracts ?

Sol.: Here P(A) = 2/3, P(B) = 5/9

P(A∪b) = 4/5, (P(A∩B) = ?

By addition theorem of Probability:

P(A∪B) = P(A) + P(B) - P(A∩B)= 4/5 = 2/3 + 5/9 - P(A∩B)

or 4/5 = 11/9 – P(A∩B)

or P(A∩B) = 11/9 – 4/5 = (55-36) / 45

P(A∩B) = 19/45

**Multiplication theorem:**

Let A and B be two independent events. Then multiplication theorem states that,

P[AB]= P[A]. P[B].

Note: P[AB] can also be represented by P[A and B] or P[A∩B].

Example:

Let a problem in statistics be given to two students whose probability of solving it are 1/5 and 5/7. What is the probability that both solve the problem.

Solution:

Let A= event that the first person solves the problem.

B= event that the second person solves the problem.

It is given that P[A]=1/ 5; P[B]=5/7.

Since A and B are independent, using multiplication theorem

P[AB]= P[A]. P[B]. = 1/5*5/7= 1/7.

**Conditional probability:**

Probability of dependent events is termed conditional probability. Let A and B be 2 events, A depending on B. Then,

$$P[A/B] = \frac{\mathrm{P[A \cap B]}}{P[B]}$$

Example:

Let a file contain 10 papers numbered 1 to 10. A paper is selected at random. What is the probability that it is 10 given that it is at least 5.

Solution:

From the problem we can see that,

Sample space ={1,2,3,4,5,6,7,8,9,10}

A- Event that number is 10 ={10}.

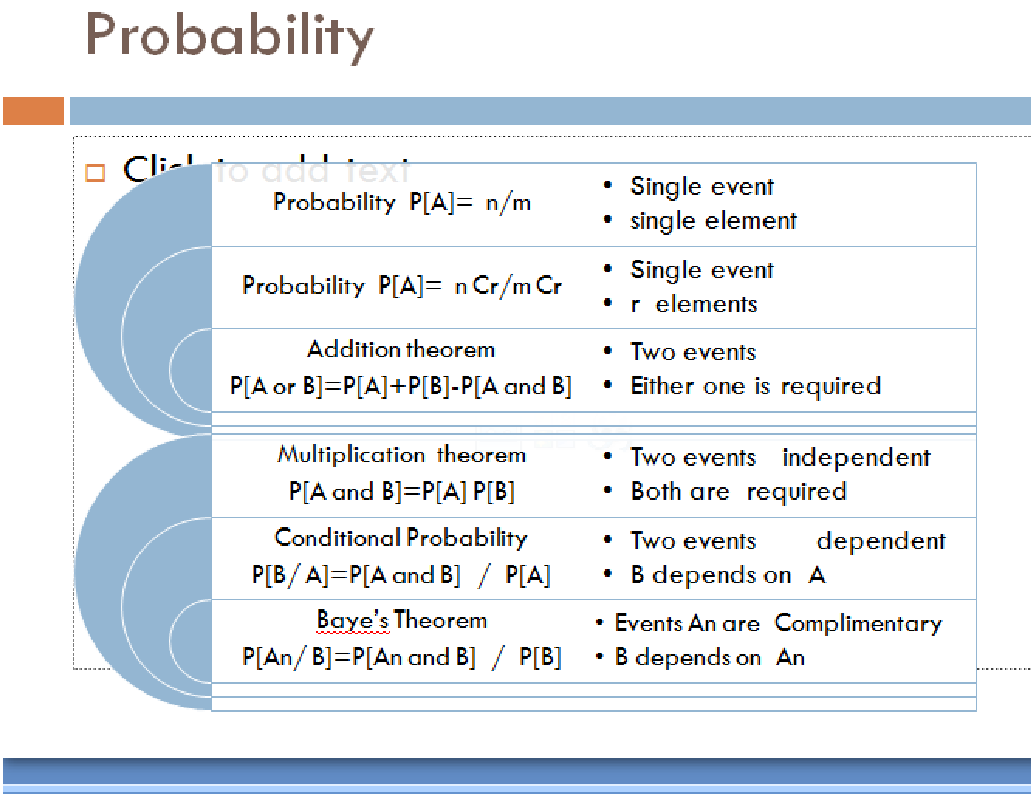
B- Event that number is at least 5 ={5,6,7,8,9,10}.

A∩B={10}.

P[A]= 1/10;   P[B] =6/10;  P[ A∩B] =1/10.

Therefore,

$$P[A/B] = \frac{P[A\cap B]}{P[B]} = \frac{1/10}{6/10} = \frac{1}{6}$$



Probability

□ Click to add Text

| | |
|---|---|
| Probability  P[A]= n/m | • Single event<br>• single element |
| Probability  P[A]= n Cr/m Cr | • Single event<br>• r elements |
| Addition theorem<br>P[A or B]=P[A]+P[B]-P[A and B] | • Two events<br>• Either one is required |
| Multiplication theorem<br>P[A and B]=P[A] P[B] | • Two events  independent<br>• Both are required |
| Conditional Probability<br>P[B/A]=P[A and B] / P[A] | • Two events  dependent<br>• B depends on A |
| Baye's Theorem<br>P[An/B]=P[An and B] / P[B] | • Events An are Complimentary<br>• B depends on An |

### 4. Random Variable

A random variable is a function that maps the set of events to $R^n$. By convention random variables are written as upper case Roman letters from the end of the alphabet like X.

For example, define the random variable X to be the sum of the two dice. For every element in the sample space, we can specify the value of X.

S={

(1; 1) = 2 (1; 2) = 3 (1; 3) = 4 (1; 4) = 5 (1; 5) = 6 (1; 6) = 7
(2; 1) = 3 (2; 2) = 4 (2; 3) = 5 (2; 4) = 6 (2; 5) = 7 (2; 6) = 8
(3; 1) = 4 (3; 2) = 5 (3; 3) = 6 (3; 4) = 7 (3; 5) = 8 (3; 6) = 9
(4; 1) = 5 (4; 2) = 6 (4; 3) = 7 (4; 4) = 8 (4; 5) = 9 (4; 6) = 10
(5; 1) = 6 (5; 2) = 7 (5; 3) = 8 (5; 4) = 9 (5; 5) = 10 (5; 6) = 11
(6; 1) = 7 (6; 2) = 8 (6; 3) = 9 (6; 4) = 10 (6; 5) = 11 (6; 6) = 12

}

If we know the probabilities of a set of events, we can calculate the probabilities that a random variable defined on those set of events takes on certain values. For example

$P(X = 2) = P((1; 1)) = 1/36$

$P(X = 5) = P((1; 4); (2; 3); (3; 2); (4; 1g) = 1/9.$

$P(X = 7) = P( (1; 6); (2; 5); (3; 4); (4; 3); (5; 2); (6; 1) ) = 1/6$

$P(X = 12) = P((6; 6)) = 1/36.$

The expression for P(X = 5) should be familiar, since we calculated it above as the probability of the event that the two dice sum to five. Much of the theory of probability is concerned with defining functions of random variables
and calculating the likelihood with which they take on their values.

So now we know something about what a random variable is. Now we see it a bit more closely. Random variables can be broadly classified into two types,

.Discrete r.v ---- these take only integer values

.continuous r.v ---- these can take any value

**Expectation Value**

Once we know the probability distribution of a random variable we can use it to predict the average outcome of functions of that variable. This is done using expectation values. The expectation value of a random variable X is defined to be

$E[(X)] =\sum_x xp(x)$ if X is discrete

$=\int xf(x)dx$ if X is continuous

The X defined in the previous section has the following mean value

$E[X] = 2P(X = 2) + 3P(X = 3) + 4P(X = 4)$ _ _ _ $+ 12P(X = 12)$

$= 7$

You can think of expectation values as taking a weighted average of the values of X
where more likely values get a higher weight than less likely values.

Note: If X is continuous we do the same process where we replace ∑ by ∫ .

**7. Variance**

Once we know the probability distribution of a random variable we can use it to predict the variance of that variable. This is done using expectation values, as

$V(X) = E[X^2]-\{E[X]\}^2$

where

E[X] =$\sum_x xp(x)$ if X is discrete

=$\int xf(x)dx$ if X is continuous

E[X$^2$] =$\sum_x x^2p(x)$ if X is discrete

=$\int x^2f(x)dx$ if X is continuous

1**. BERNOULLI:**

Bernoulli trials are trials with 2 outcomes, success and failure, with

- □ A coin is tossed
- □ A die is tossed
- □ We write an examination

Probabilities p and q=1-p respectively. Its probability mass function is given by,

$$p(x) = \begin{cases} p & x=1 \\ q & x=0 \end{cases}$$

E(X) = p

VAR(X) = p(1-p)=pq .

2. **BINOMIAL:**

The random variable X denoting the number of successes in a fixed Number of independent Bernoulli trials is called a binomial random variable and its distribution is Binomial distribution as defined below

$$p(x) = nCr\ p^r q^{n-r}$$

E(X)=np

VAR(X)=np(1-p)=npq.

**Example**: A bag contains 50 balls of which 35 are of red colour and15 are black. 5times a ball is randomly selected , colour is noted and replaced. Find the probability that 2 times black balls are selected.

Solution:

Here every time we draw a ball it is a Bernoulli trial, as we have only 2 possibilities.

So, n=5; p=15/50; q=1-p=35/50; x=2.

P(X=2)={nCxp$^x$q$^{n-x}$

= 5C2 15/50$^2$ 35/50$^{5-2}$

**GEOMETRIC:**

The random variable X denoting the number of Bernoulli trials required to achieve the first success is called a geometric random variable and its distribution is geometric distribution.

P(X=x) = $\begin{cases} pq^{x-1} & x = 1, 2, 3 \dots\dots. \\ 0 & \end{cases}$

**Example**: A bag contains 50 balls of which 35 are of red colour and15 are black. A ball is randomly selected, if it is red it is replaced and again we select and continue till we get a black for the first time. Find the probability that we need to select 7 times before black balls is obtained.

Solution:

Here every time we draw a ball it is a Bernoulli trial, as we have only 2 possibilities.

So, x=7; p=15/50; q=1-p=35/50;

P[X=7]=$\frac{15}{50}\,\frac{35}{50}^{7-1}$

4**. POISSON:**

The random variable X whose pmf is,

P(X=x) = $\begin{cases} (e^{-\lambda}\,\lambda^{x}) \,/\, x! & x = 1, 2, 3 \dots\dots. \\ 0 & \end{cases}$

E(X) = VAR(X) = λ.

**Example**: A bag contains 50 balls of which 35 are of red colour and15 are black. 20 times a ball is randomly selected , colour is noted and replaced. Find the probability that 2 times black balls are selected.

Solution:

Here every time we draw a ball it is a Bernoulli trial, as we have only 2 possibilities.

So, n=20; p=15/50; λ=np=6; x=2.

P(X=x) = $\begin{cases} \frac{(e^{-\lambda}\,\lambda^{x})}{x!} = \frac{(e^{-6}\,6^{2})}{2!} \end{cases}$

**DISCUSS ABOUT CONTINUOUS DISTRIBUTIONS:**

(a) **UNIFORM:**

A random variable X is uniformly distributed on the interval ( a, b) if its pdf is given by,

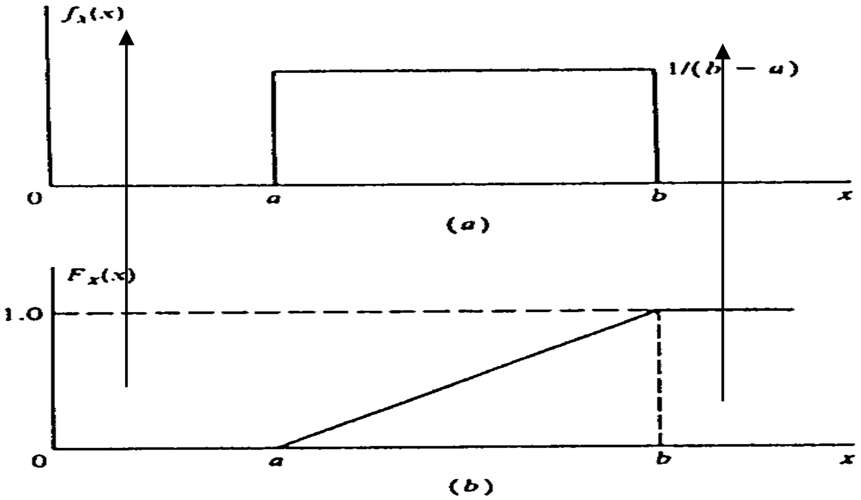
f(x) = $\begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & else \end{cases}$

Its cdf is,

F(x) = $\begin{cases} 0 & x < a \\ \frac{(x-a)}{(b-a)} & a \leq x \leq b \\ 1 & x > b \end{cases}$ {0

E(X) = (a+b)/2

V(X)=(b-a)$^{2}$/12



**Example:**

If a wheel is spun and then allowed to come to rest, the point on the circumference of the wheel that is located opposite a certain fixed marker could be considered the value of a random variable X that is uniformly distributed over the circumference of the wheel. One could then compute the probability that X will fall in any given arc.
If we assume that it is uniform in the interval[3,6], we can obtain,

Average point of outcome, E[X]= [a+b]/12 = [3+6]/12=9/12=3/4.
Variance var[X]= $[b-a]^2/12= [6-3]^2/12=6/12=1/2$.

2. **EXPONENTIAL:**

A Random variable X is said to be exponentially distributed if its pdf is given by,

f(x) = { $\lambda e^{-\lambda x}$ $x \geq 0$
{ 0 otherwise

Where λ – parameter.

f(x) = { 0 $x < 0$
{ $1- \lambda e^{-\lambda x}$ $x \geq 0$

E(X) = 1/λ.
V(X) = $1/\lambda^2$.

Exponential distribution is useful in representing lifetime of items, model interarrival times when arrivals are completely random and service times which are highly variable.

Exponential distribution has a property called memory less property given by,

P(X > s + t / X > s) = p(X > t)

This is why we are able to use exponential to model lifetimes.

**Example:**

Let us assume that a company is manufacturing burettes whose lifetime is assumed to be exponential with average life, 950 days. What is the probability that it is in working condition for up to 1000 days.

Solution:

It is given that , X= Lifetime of the burette , is exponential with

average life 950 days i.e $\lambda=950$.

P[life time is up to 1000 days] = P[0<X<1000] = $\int_0^{1000} \lambda\, e^{-\lambda x}$

$= \int_0^{1000} 950\, e^{-950x}$

$= 950\, [e^{-950x}/-950]_0^{1000}$ .

**3. NORMAL:**

A normal variable X with mean $\mu$( $-\infty < \mu < \infty$ ) and variance $\sigma^2 > 0$ has a normal distribution if its pdf is,

$$f(x) = ( 1/\sqrt{2\pi} ) \exp [ -1/2 ( x-\mu/\sigma)^2 ] \qquad -\infty < x < \infty$$

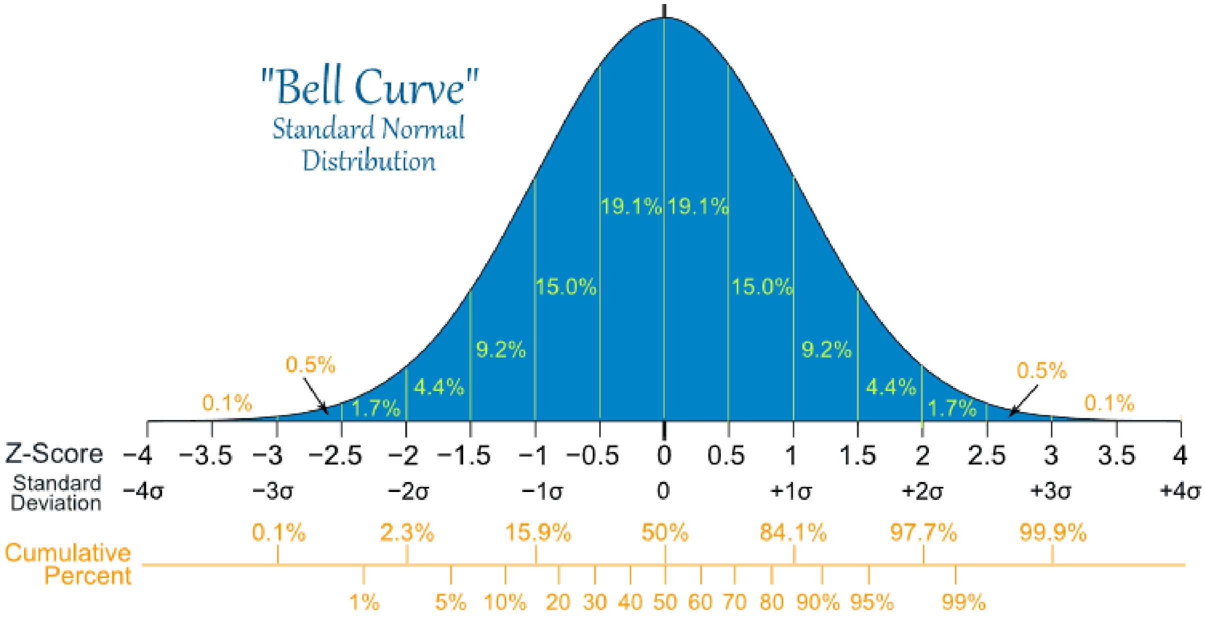
A normal distribution is used when we are having a sum of many random variables. A normal random variable with $\mu = 0$ and $\sigma = 1$ is called a standard normal r.v. Its curve is symmetrically distributed about the average $\mu = 0$.

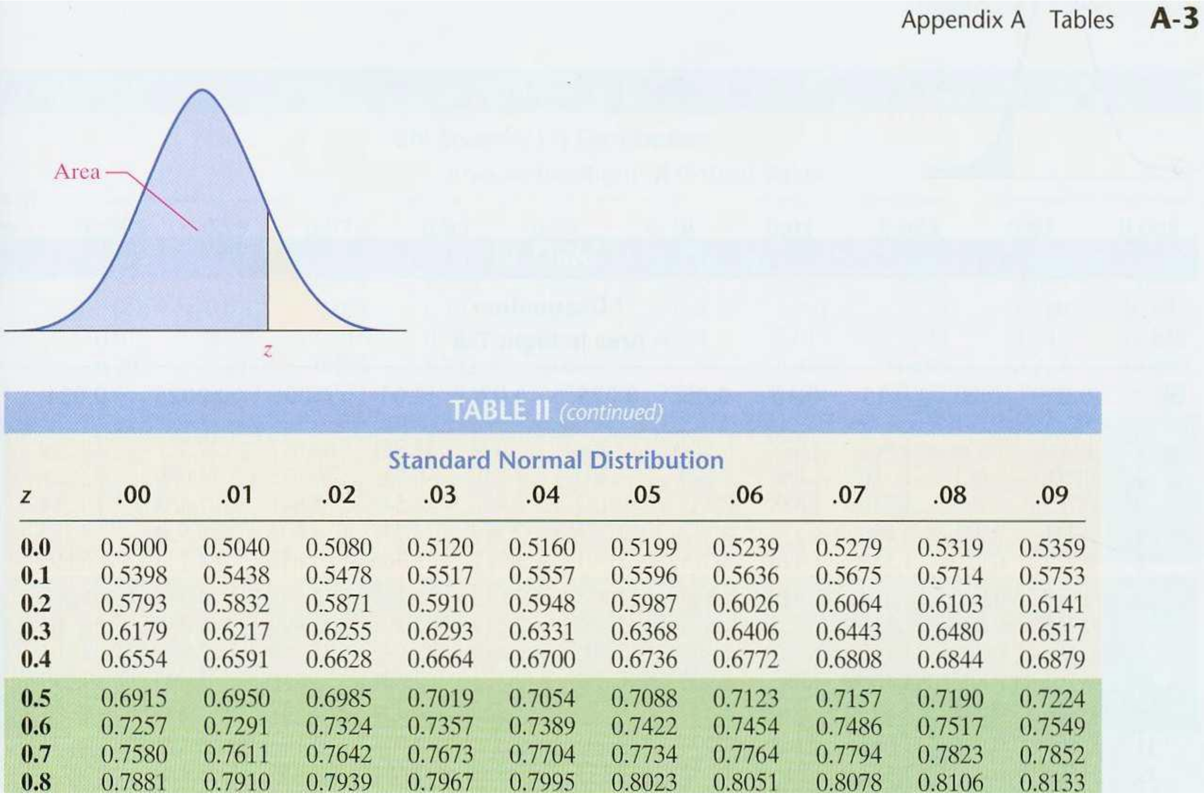
We Standardize a normal distribution by

- □ Z=[X-μ]/sigma

$$p_Z(z) = (2\pi)^{-1/2} e^{-z^2/2} \qquad -\infty < z < \infty$$

- □ Which will give us the pdf





Appendix A Tables A-3

**TABLE II** *(continued)*

Standard Normal Distribution

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0.0** | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| **0.1** | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| **0.2** | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| **0.3** | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| **0.4** | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| **0.5** | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| **0.6** | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| **0.7** | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| **0.8** | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |

**Example:**

Let us assume that heights of students in II M.Pharm is normally distributed with an average of 165 cm and a standard deviation of 10 cms. What is the probability that a student's height is less than 175 cms.

Solution:

Let, X= Height of students in II M.Pharm.

It is normal with, mean μ= 165; standard deviation σ =10.

P[ a student's height is less than 175 cms]=P[-∞<X<175]

First, we should convert X into Z by

$$Z= x- \mu/ \sigma.$$

We have x=175, μ= 165; σ =10.

$$Z= 175- 165/ 10 =1.$$

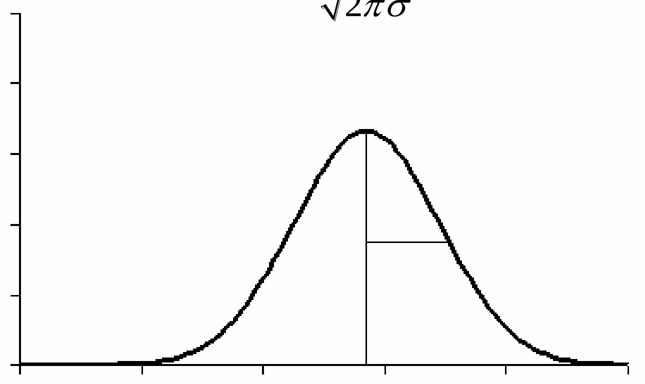
So when X=175; Z=1 and so

$$P[-\infty<X<175] = P[-\infty<Z<1]= P[-\infty<Z<0]+ P[0<Z<1].$$

The Normal distribution

(mean $\mu$, standard deviation $\sigma$)

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



=0.5+0.34 = 0.84.

**Note:**

1. The same question may have the following variations:

   P[ a student's height is more than 175 cms]=P[175<X<-∞]
   = P[0<X<-∞]- P[0<X<175] =0.5- table value

   P[ a student's height is between 165 and 175 cms]=P[165 < X <175]
   =P[0 < X <175]- P[ 0< X <165]=table value for 175 – table value for 165

## Bayesian Decision Theory:

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification
Decision making when all the probabilistic information is known
For given probabilities the decision is optimal
When new information is added, it is assimilated in optimal fashion for improvement of decisions
Example:
Suppose we have a conveyor belt carrying fish of two types Sea bass and salmon and we need the machine to identify them and pack.
Soln:
Let w=State of nature, so that w1= sea bass and w2=salmon.
a Priori probability P(w1): probability that next fish in line is sea bass;
P(w2): probability that next fish in line is salmon;

We need features for classification like Length, Lightness, width, Number and shape of fins, Position of the mouth. If we have character x, then P(wi/x) is the conditional probability after measuring the feature, also called ‘a Posteriori’. There is one more conditional probability here, which is P(x/wi) which is called the ‘Likelihood’ and the probability(x) , called the ‘Evidence’, makes up therequired set for the Bayesian inference

$$\boxed{P(\omega_j|x) = \frac{p(x\,|\,\omega_j)P(\omega_j)}{p(x)}}$$

$$\boxed{\text{Posterior} = \frac{\text{Likelihood} \ * \ \text{Prior}}{\text{Evidence}}}$$

The procedure is very simple. Calculate the ‘a Posteriori’ probabilities and classify the element into the class with the high ‘a Posteriori’ w1 if P(w1/x)>=P(w2/x), else w2.

Naïve Baye’s:

- Prior, conditional and joint probability for random variables
  - Prior probability: P(x)
  - Conditional probability: $P(x_1\,|\,x_2),\, P(x_2\,|\,x_1)$
  - Joint probability: $\mathbf{x} = (x_1, x_2),\, P(\mathbf{x}) = P(x_1\,, x_2)$
  - Relationship: $P(x_1\,, x_2) = P(x_2\,|\,x_1)P(x_1) = P(x_1\,|\,x_2)P(x_2)$
  - Independence: $P(x_2\,|\,x_1) = P(x_2),\, P(x_1\,|\,x_2) = P(x_1),\, P(x_1\,, x_2) = P(x_1)P(x_2)$
- Bayesian Rule $P(c\,|\,\mathbf{x}) = \dfrac{P(\mathbf{x}\,|\,c)P(c)}{P(\mathbf{x})} \qquad Posterior = \dfrac{Likelihood \times Prior}{Evidence}$

**Naive Bayes Examples**

We are going to use the Iris dataset we used in the previous blogs to illustrate how Naive Bayes works. Let’s suppose we measure an Iris Setosa and find the following measurements:

- Sepal length = 7 cm
- Sepal width = 3 cm
- Petal length = 5 cm
- Petal width = 2 cm

From our data we know that each class—Iris versicolor, Iris virginica, and Iris setosa—represents one-third of the data. Following Naive Bayes, we need to calculate the following conditional probabilities and categorize our measured flower with the class that has the highest probability. To do so, we are going to use the Gaussian measure of likelihood:

- P( Setosa | 7,3,5,2)

- P( Versicolor | 7,3,5,2)
- P( Virginica | 7,3,5,2)

Let's do one calculation with Versicolor:

- P( Versicolor | 7,3,5,2) = (P( 7,3,5,2 | Versicolor ) * P( Versicolor ) )/ P( 7,3,5,2)

Assuming independence and using the Gaussian distribution of conditional class probabilities, we can calculate the following:

- P( 7,3,5,2 | Versicolor ) = P(7 | Versicolor ) * P(3 | Versicolor ) * P(5 | Versicolor ) * P(2 | Versicolor )

To calculate each of the conditional class probabilities, we have to find the average and standard deviation of each of the features operating under the assumption that they are Versicolor. The average and standard deviation are calculated as follows.

| | Sepal Length | Sepal Width | Petal Length |
|---|---|---|---|
| Average | 5.936 | 2.77 | 4.26 |
| Standard Deviation | 0.51 | 0.31 | 0.46 |

Plugging those values into a Gaussian distribution, we can calculate: ( N stands for the normal distribution)

P( 7,3,5,2 | Versicolor ) = P(7 | Versicolor ) * P(3 | Versicolor ) * P(5 | Versicolor ) * P(2 | Versicolor )

P( 7,3,5,2 | Versicolor ) = N(7 | 5.936, 0.51) * N(3 | 2.77, 0.31) * N(5 | 4.26, 0.46) * N(2 | 1.32, 0.19)

P( 7,3,5,2 | Versicolor ) = 0.089 * 0.97 * 0.24 * 0.05

P( 7,3,5,2 | Versicolor ) = 0.001

P( Versicolor ) = 50/150

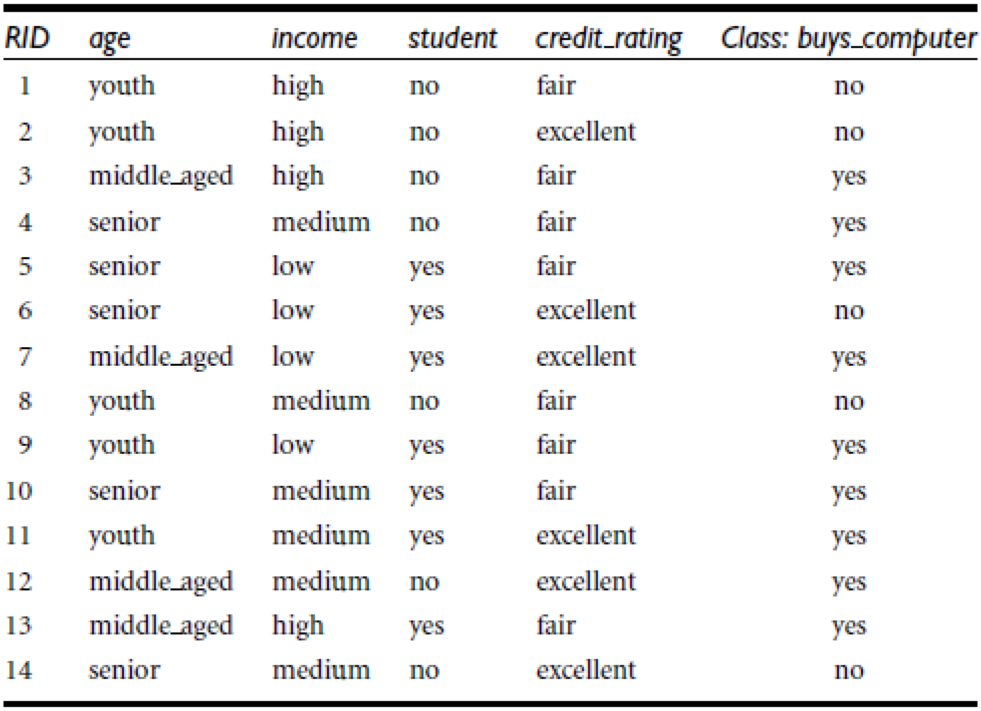
At last we can calculate: P( 7,3,5,2 | Versicolor ) 0.001* 0.33 = .0003

From here, we would need to calculate the same process for Setosa and Virginica in order to determine in which class the original flower is most likely to belong.

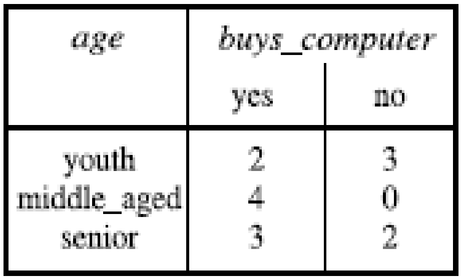
## Naïve Bayes classifier

- Classification-Assigning an input to one of the classes which has the maximum Bayes' posterior probability.
- The class *Ci* for which *P*(*Ci*/***X***) is maximized is called the *maximum posteriori hypothesis*.
- *P*(*Ci*/***X***) =*P*(***X***/*Ci*)*P*(*Ci*)/*P*(***X***)

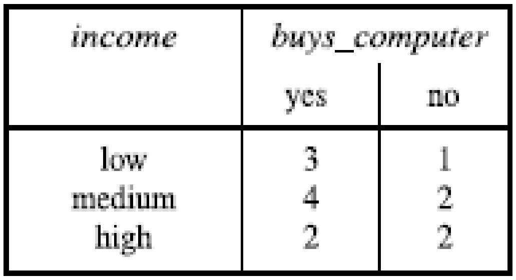
Example:



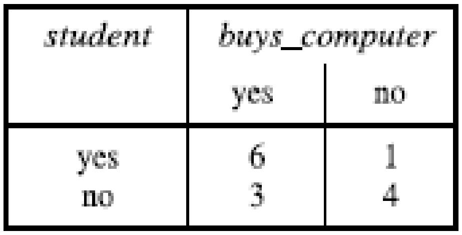
| RID | age | income | student | credit_rating | Class: buys_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle_aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle_aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle_aged | medium | no | excellent | yes |
| 13 | middle_aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |



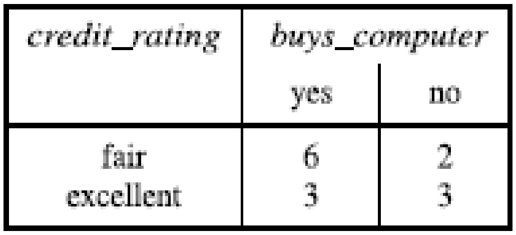
| age | buys_computer | |
|---|---|---|
| | yes | no |
| youth | 2 | 3 |
| middle_aged | 4 | 0 |
| senior | 3 | 2 |



| income | buys_computer | |
|---|---|---|
| | yes | no |
| low | 3 | 1 |
| medium | 4 | 2 |
| high | 2 | 2 |



| student | buys_computer | |
|---|---|---|
| | yes | no |
| yes | 6 | 1 |
| no | 3 | 4 |

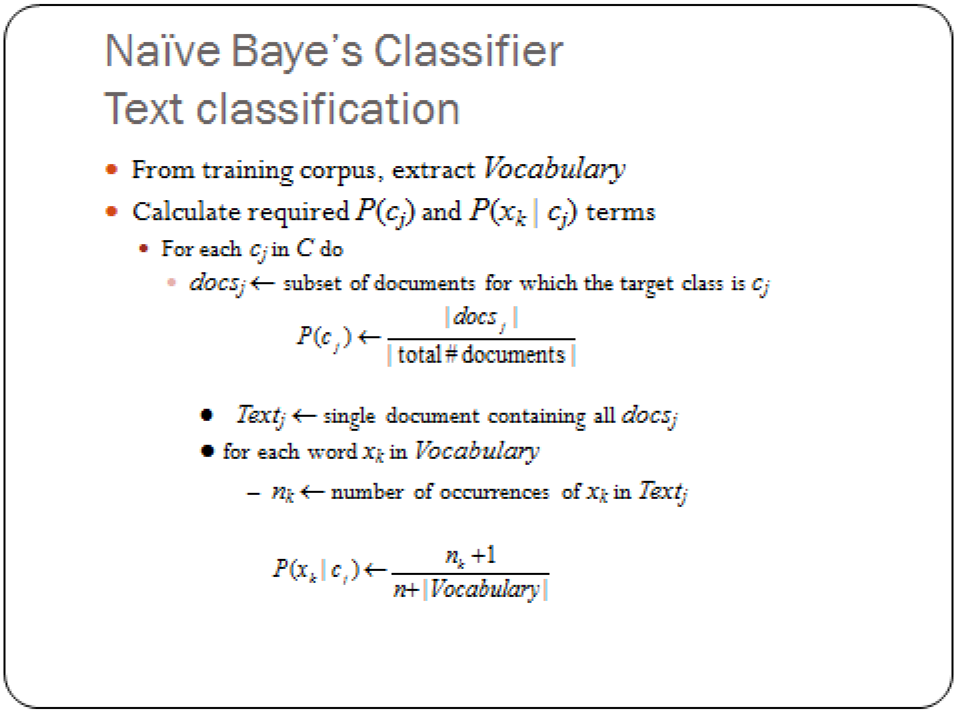


| credit_rating | buys_computer | |
|---|---|---|
| | yes | no |
| fair | 6 | 2 |
| excellent | 3 | 3 |

- ***X*** = (*age = youth, income = medium, student = yes, credit rating = fair*)
- We need to maximize *P*(***X***/*Ci*)*P*(*Ci*), for *i* = 1, 2. *P*(*Ci*), the prior probability of each
- class, can be computed based on the training tuples:
- *P*(*buys computer = yes*) = 9/14 = 0.643
- *P*(*buys computer = no*) = 5/14 = 0.357
- ***X*** = (*age = youth, income = medium, student = yes, credit rating = fair*)
- To compute *P(**X**/Ci*), for *i* = 1, 2, we compute the following conditional probabilities:
- *P*(*age = youth / buys computer = yes*) = 2/9 = 0.222
- *P*(*age = youth / buys computer = no*) = 3/5 = 0.600
- *P*(*income = medium / buys computer = yes*) = 4/9 = 0.444

- *P*(*income = medium /buys computer = no*) = 2/5 = 0.400
- *P*(*student = yes / buys computer = yes*) = 6/9 = 0.667
- *P*(*student = yes / buys computer = no*) = 1/5 = 0.200
- *P*(*credit rating = fair / buys computer = yes*) = 6/9 = 0.667
- *P*(*credit rating = fair / buys computer = no*) = 2/5 = 0.400
- Using the above probabilities, we obtain
- *P*(***X***/*buys computer = yes*) =
- *P*(*age = youth / buys computer = yes*) *
- *P*(*income = medium / buys computer = yes*)*
- *P*(*student = yes / buys computer = yes*) *
- *P*(*credit rating = fair / buys computer = yes*)
- = 0.222*0.444*0.667*0.667 = 0.044.
- Similarly,
- *P*(***X***/*buys computer = no*) = 0.600*0.400*0.200*0.400 = 0.019.
- To find the class, *Ci*, that maximizes *P*(***X***/*Ci*)*P*(*Ci*), we compute
- *P*(***X***/*buys computer = yes*)*P*(*buys computer = yes*) = 0.044*0.643 = 0.028
- *P*(***X***/*buys computer = no*)*P*(*buys computer = no*)
- =0.019*0.357 = 0.007
- Therefore, the naïve Bayesian classifier predicts *buys computer = yes* for tuple ***X***.

Naïve Baye's Classifier
Text classification



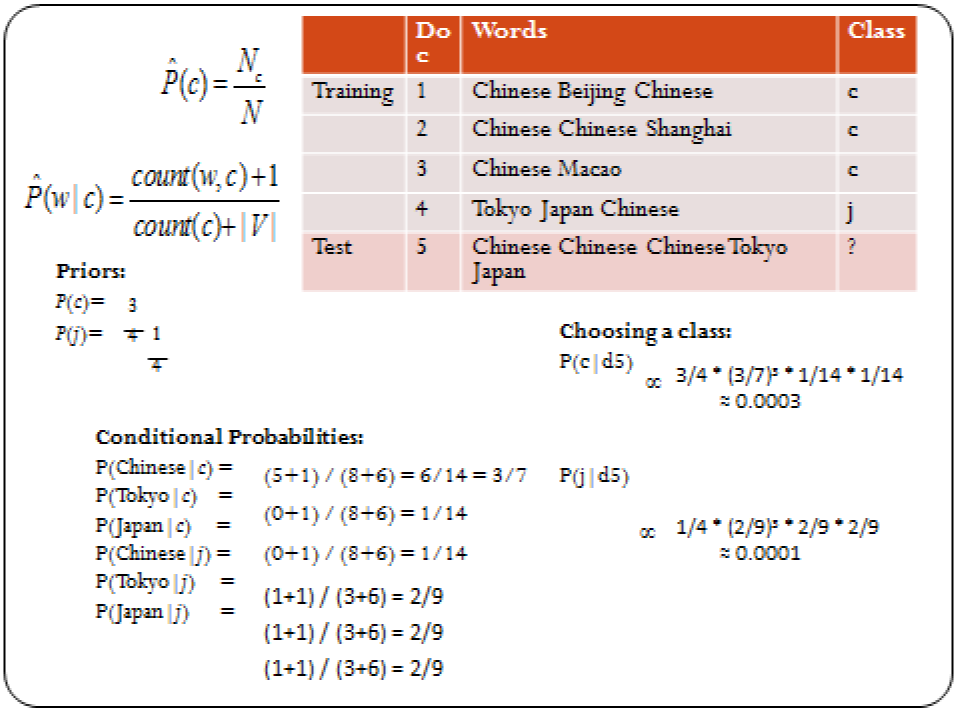
## Naïve Baye's Classifier Text classification

- From training corpus, extract *Vocabulary*
- Calculate required $P(c_j)$ and $P(x_k \mid c_j)$ terms
  - For each $c_j$ in $C$ do
    - $docs_j \leftarrow$ subset of documents for which the target class is $c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- $Text_j \leftarrow$ single document containing all $docs_j$
- for each word $x_k$ in *Vocabulary*
  - $n_k \leftarrow$ number of occurrences of $x_k$ in $Text_j$

$$P(x_k \mid c_j) \leftarrow \frac{n_k + 1}{n + |Vocabulary|}$$



$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

|  | Doc | Words | Class |
|---|---|---|---|
| Training | 1 | Chinese Beijing Chinese | c |
|  | 2 | Chinese Chinese Shanghai | c |
|  | 3 | Chinese Macao | c |
|  | 4 | Tokyo Japan Chinese | j |
| Test | 5 | Chinese Chinese Chinese Tokyo Japan | ? |

**Priors:**

$P(c)=$ 3

$P(j)=$ 4 1

4

**Choosing a class:**

P(c | d5) $\propto$ 3/4 * $(3/7)^3$ * 1/14 * 1/14 ≈ 0.0003

**Conditional Probabilities:**

P(Chinese | c) = (5+1) / (8+6) = 6 / 14 = 3 / 7

P(Tokyo | c) = (0+1) / (8+6) = 1 / 14

P(Japan | c) = (0+1) / (8+6) = 1 / 14

P(Chinese | j) = (1+1) / (3+6) = 2/9

P(Tokyo | j) = (1+1) / (3+6) = 2/9

P(Japan | j) = (1+1) / (3+6) = 2/9

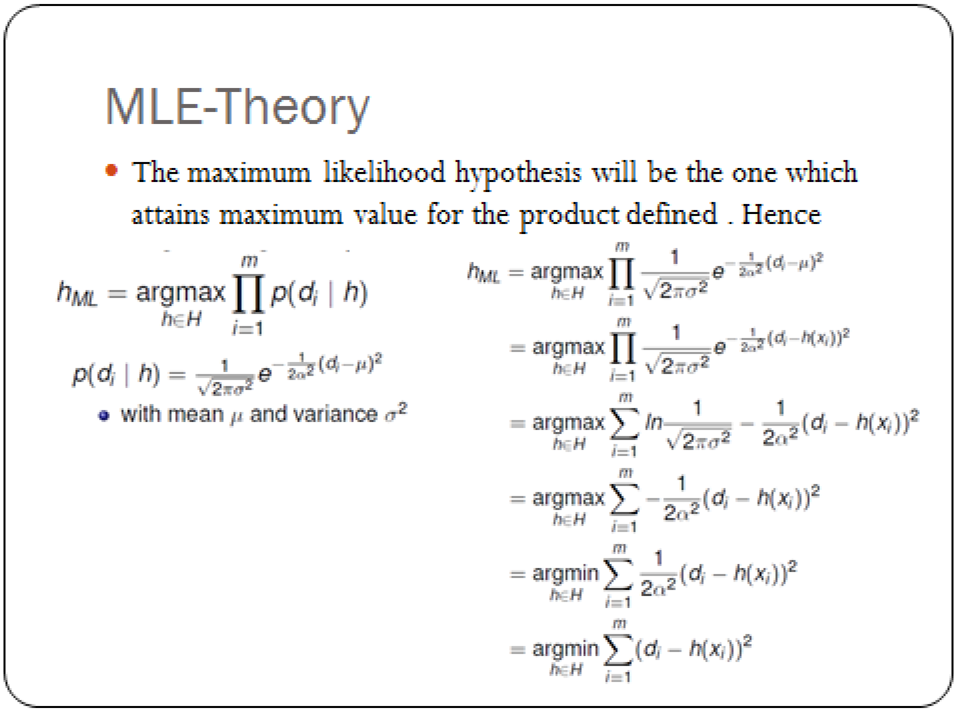
P(j | d5) $\propto$ 1/4 * $(2/9)^3$ * 2/9 * 2/9 ≈ 0.0001

## MLE-MAP:

- The learner considers some set of candidate hypotheses H and it is interested in finding the ***most probable hypothesis*** h **in** H given the observed data D.
- Any such maximally probable hypothesis is called a ***maximum a posteriori (MAP) hypothesis h<sub></sub>*** ***$h_{MAP}$***.
- Using Bayes theorem

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{\arg\max}\, P(h \mid D) \\ &= \underset{h \in H}{\arg\max}\, \frac{P(D \mid h)P(h)}{P(D)} \\ &= \underset{h \in H}{\arg\max}\, P(D \mid h)P(h) \end{aligned}$$

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



## MLE-Theory

- The maximum likelihood hypothesis will be the one which attains maximum value for the product defined . Hence

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} p(d_i \mid h)$$

$$p(d_i \mid h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\alpha^2}(d_i - \mu)^2}$$

- with mean $\mu$ and variance $\sigma^2$

$$\begin{aligned} h_{ML} &= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\alpha^2}(d_i - \mu)^2} \\ &= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\alpha^2}(d_i - h(x_i))^2} \\ &= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\alpha^2}(d_i - h(x_i))^2 \\ &= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2\alpha^2}(d_i - h(x_i))^2 \\ &= \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} \frac{1}{2\alpha^2}(d_i - h(x_i))^2 \\ &= \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} (d_i - h(x_i))^2 \end{aligned}$$

Example:

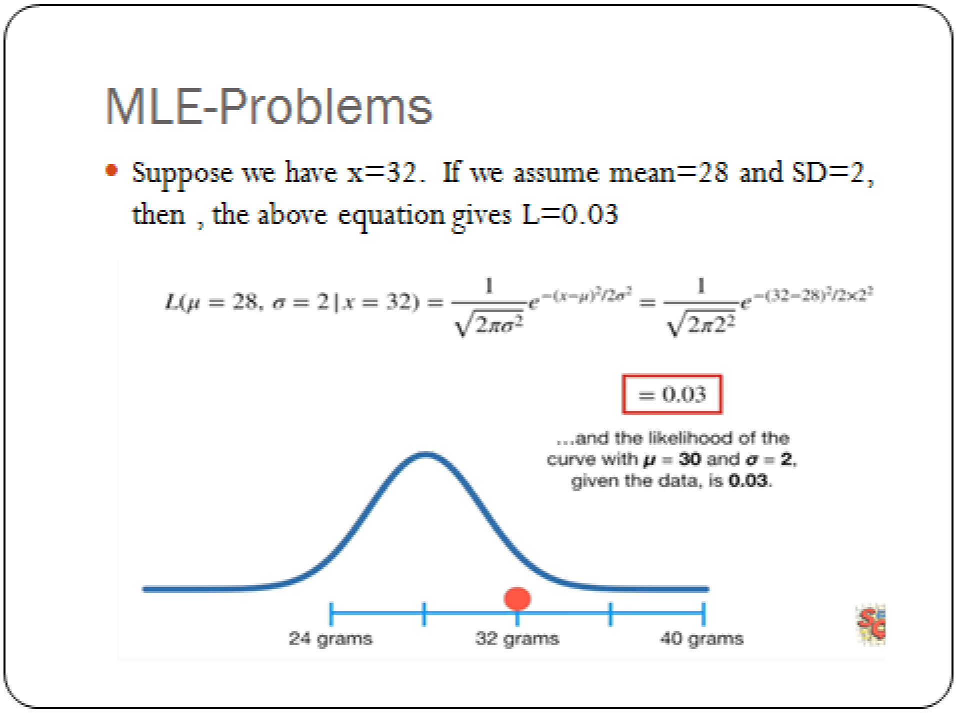
- Suppose a teacher has two hypotheses
- h1-> Student is attentive in class
- h2-> Student is not attentive in class
- Let p(attn)=0.67 and p(not attn)=0.33
- p(pass|attn)=0.90; p(fail|attn)=0.10
- p(pass|not attn)=0.15; p(fail|not attn)=0.85.
- If a student has passed the exam, what can we say anything about his attentiveness

| **Event** | **Prior** | **Posterior** |
|---|---|---|
| h1 | 0.67 | 0.92 |
| h2 | 0.33 | 0.075 |

MLE:

- Difference between probability and Likelihood.
- Assume we are in Gaussian domain. That is , say, heights of students in this class is normally distributed with mean 160 and SD=25.
- Then we say probability of a randomly selected student’s height to be 165 is (from area ideas)=0.20.
- On the other hand , if we do not know the exact parameters of the distribution and we have a student’s height is known as 165. Then we’ll ask, what is the Likelihood that it is N(160,25)?

- We can have many normal distributions around the sample and we calculate the likelihood of all. Finally we select the one with the maximum likelihood as the best approximate.
- Notationally $P(x \mid \mu, \sigma) =$
  $L(\mu, \sigma \mid x) =$
- Suppose we have x=32. If we assume mean=28 and SD=2, then , the above equation gives L=0.03



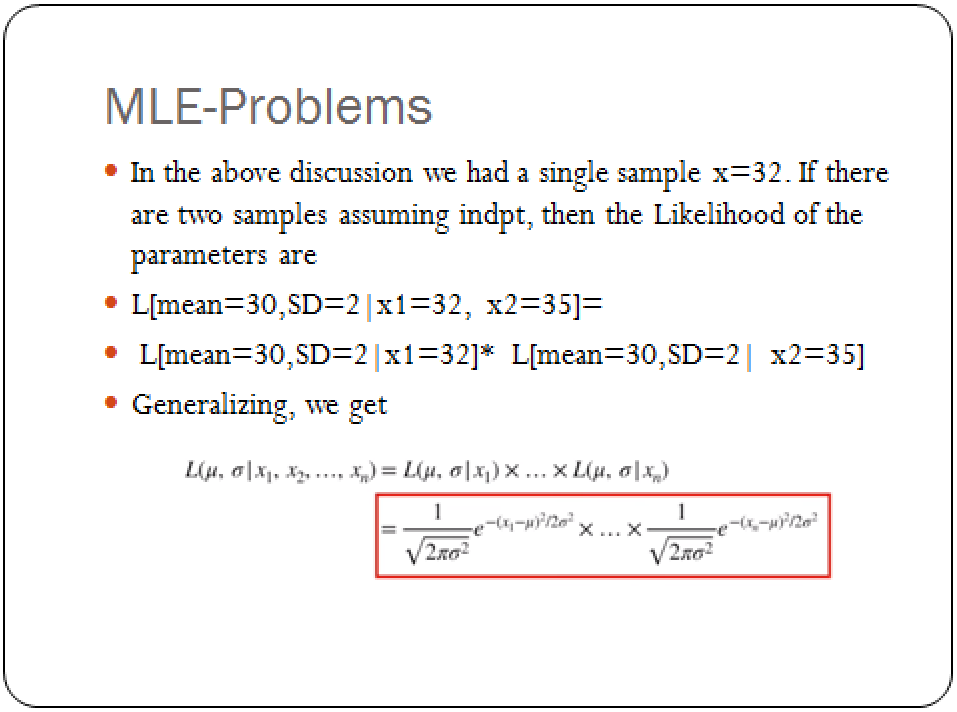
## MLE-Problems

- Suppose we have x=32. If we assume mean=28 and SD=2, then , the above equation gives L=0.03

$$L(\mu = 28, \sigma = 2 \mid x = 32) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi 2^2}} e^{-(32-28)^2/2\times 2^2}$$

$$\boxed{= 0.03}$$

- Suppose we move it to mean=30 and SD=2, then , the above equation gives L=0.12 which is much better than the earlier one.
- We can calculate L for various values of mean , plot it to find the normal like curve , giving the maximum L value and the corresponding parameters are selected as the best.
- In the above we varied mean, keeping SD fixed to find ML for mean. Similarly we can do again to find ML of SD by varying it, keeping the mean as fixed.



## MLE-Problems

- In the above discussion we had a single sample x=32. If there are two samples assuming indpt, then the Likelihood of the parameters are
- L[mean=30,SD=2|x1=32, x2=35]=
- L[mean=30,SD=2|x1=32]* L[mean=30,SD=2| x2=35]
- Generalizing, we get

$$L(\mu,\sigma|x_1,x_2,\ldots,x_n)=L(\mu,\sigma|x_1)\times\ldots\times L(\mu,\sigma|x_n)$$

$$\boxed{=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x_1-\mu)^2/2\sigma^2}\times\ldots\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x_n-\mu)^2/2\sigma^2}}$$

MLE-Example:

- Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form P[X=0]=(1-p), P[X=1]=p. Find the maximum likelihood estimate of p.
- Soln :
- L(p)=P[X=0] P[X=1] P[X=0] P[X=0] P[X=1] P[X=0]
- =(1-p) p (1-p) (1-p) p (1-p)
- =$(1-p)^3p^2$.
- Log L(p)=log[$(1-p)^3p^2$.]=3log(1-p)+2logp
- Log L(p)=log[$(1-p)^3p^2$.]=3log(1-p)+2logp.
- To find minimum, find derivative w.r.t p, equate it to zero to get p=2/5.
- Is this maximum or minimum?
- To find it, get the second derivative , substitute to find out.

MLE Ex2:

- Find the maximum likelihood of the Poisson distributed sample 5,9,3,12,14.

Poisson is $P[X=k]=\dfrac{e^{-\lambda}\lambda^x}{x!}$

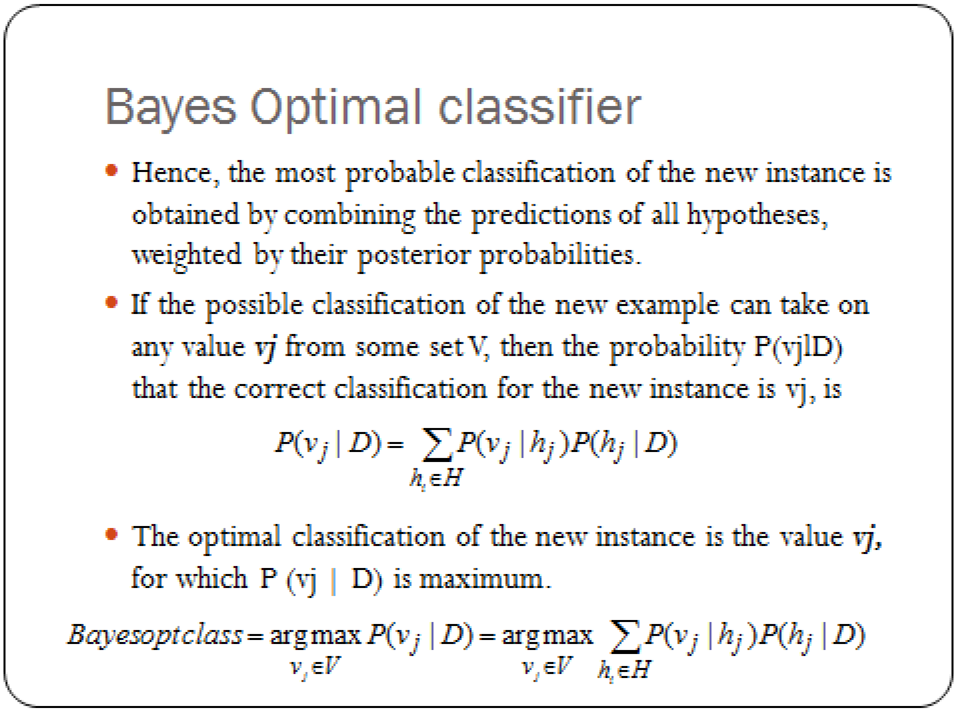
$$L(\lambda)=\frac{e^{-\lambda}\lambda^5}{5!}\,\frac{e^{-\lambda}\lambda^9}{9!}\,\frac{e^{-\lambda}\lambda^3}{2!}\,\frac{e^{-\lambda}\lambda^{12}}{12!}\,\frac{e^{-\lambda}\lambda^{14}}{14!}$$

$$=\frac{e^{-5\lambda}\lambda^{43}}{5!9!2!12!14!}$$

Taking log, finding first derivative, we get lambda=43/5.

## Bayes Optimal classifier:

- MAP ,ML, Bayes.. discuss the most probable or representative for the data. But that is to get the model ready. Its purpose , like classification, needs the models efficiency on any new instance and the most probable classification need not be the most likely one or MAP.
- consider a hypothesis space containing three hypotheses, hl, h2, and h3 with posteriors 0.4,0.3 and 0.3 resply. The MAP is h1. Suppose a new instance x1 is classified positive by h1 and negative by h2 and h3, then the most likely classification is negative (h2+h3)



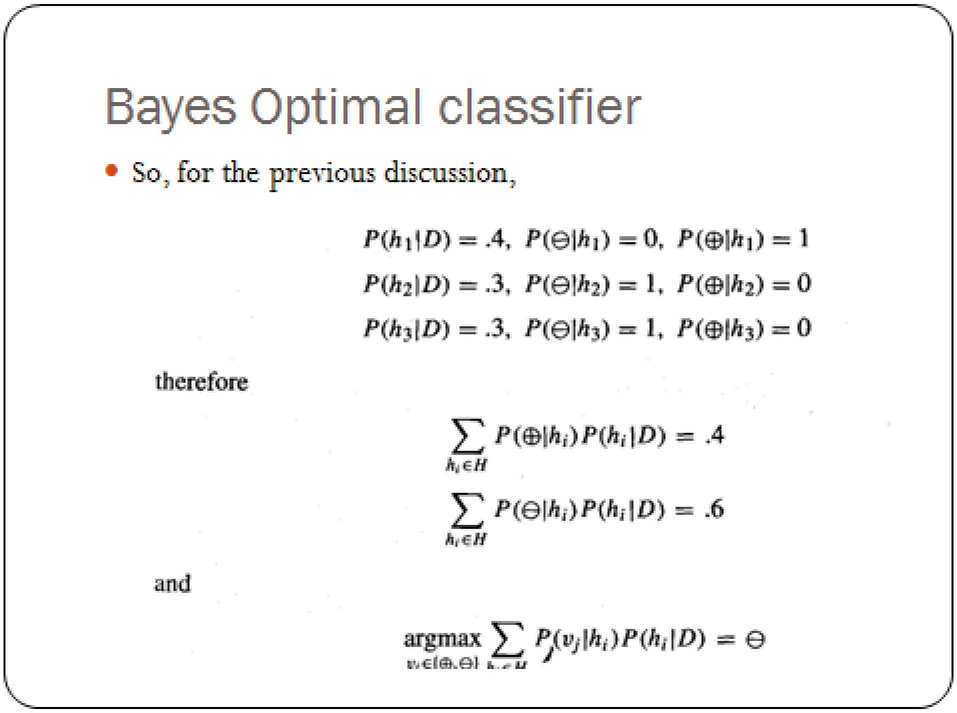
## Bayes Optimal classifier

- Hence, the most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities.
- If the possible classification of the new example can take on any value **vj** from some set V, then the probability P(vjlD) that the correct classification for the new instance is vj, is

$$P(v_j \mid D) = \sum_{h_i \in H} P(v_j \mid h_j) P(h_j \mid D)$$

- The optimal classification of the new instance is the value **vj,** for which P (vj | D) is maximum.

$$Bayesoptclass = \underset{v_j \in V}{\arg\max}\, P(v_j \mid D) = \underset{v_j \in V}{\arg\max} \sum_{h_i \in H} P(v_j \mid h_j) P(h_j \mid D)$$



## Bayes Optimal classifier

- So, for the previous discussion,

$$P(h_1|D) = .4,\ P(\ominus|h_1) = 0,\ P(\oplus|h_1) = 1$$
$$P(h_2|D) = .3,\ P(\ominus|h_2) = 1,\ P(\oplus|h_2) = 0$$
$$P(h_3|D) = .3,\ P(\ominus|h_3) = 1,\ P(\oplus|h_3) = 0$$

therefore

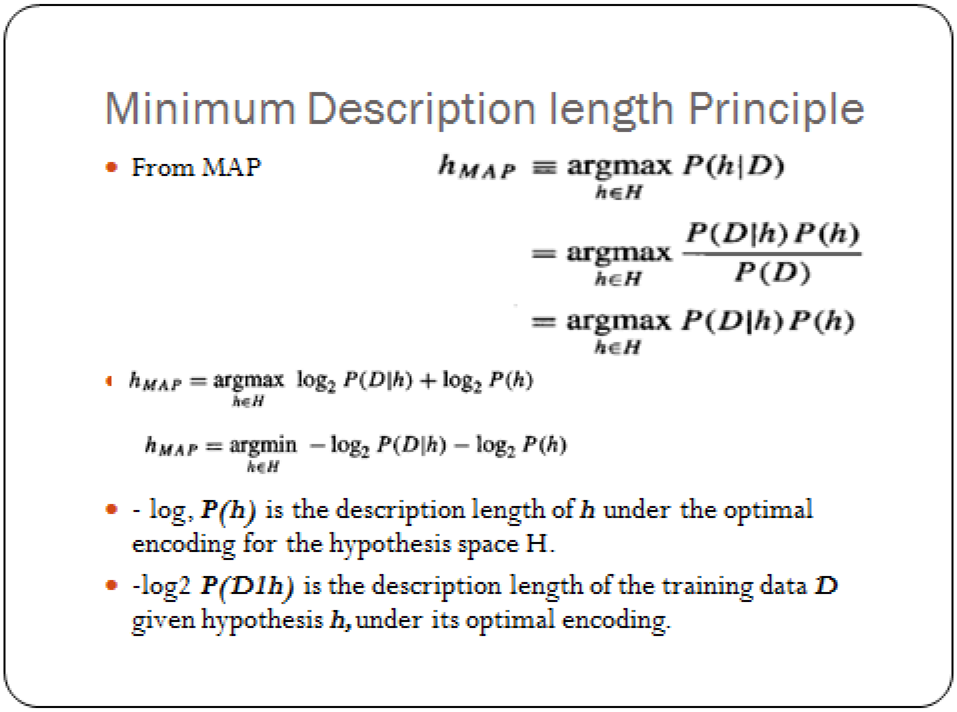
$$\sum_{h_i \in H} P(\oplus|h_i) P(h_i|D) = .4$$
$$\sum_{h_i \in H} P(\ominus|h_i) P(h_i|D) = .6$$

and

$$\underset{v_j \in \{\oplus,\ominus\}}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D) = \ominus$$

**Gibbs Algorithm:**

- 1. Choose a hypothesis ***h*** from H at random, according to the posterior probability distribution over H.
- 2. Use ***h*** to predict the classification of the next instance ***x.***



## Minimum Description length Principle

- From MAP $$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D) = \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)} = \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)$$
- $h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \ \log_2 P(D|h) + \log_2 P(h)$

  $h_{MAP} = \underset{h \in H}{\operatorname{argmin}} \ -\log_2 P(D|h) - \log_2 P(h)$
- - log, ***P(h)*** is the description length of ***h*** under the optimal encoding for the hypothesis space H.
- -log2 ***P(D1h)*** is the description length of the training data ***D*** given hypothesis ***h,*** under its optimal encoding.

- From MAP
- Shannon –
- the optimal code (i.e., the code that minimizes the expected message length) assigns - log, ***pi*** bitst to encode message ***i .***
- minimize the expected code length we should assign shorter codes to messages that are more probable.
- number of bits required to encode message ***i*** using code ***C*** as the ***description length of message i with respect to C,*** which we denote by ***Lc(i)***

## Regression:

Regression is the procedure to obtain the type of relation existing between the variables under discussion.

The term **linear model** is used in different ways according to the context. The most common occurrence is in connection with regression models and the term is often taken as synonymous with Linear regression model. The designation "linear" is used to identify a subclass of models for which substantial reduction in the complexity of the related Statistical theory is possible.

Let us consider two variables, X and Y. Since we are theoretically considering their relation, keeping each as an independent variable we 'll derive an equation.

**Regression line of X on Y[X depending on Y]**

$$X-\bar{X} = b_{xy} [Y-\bar{Y}]$$

Where,

$\bar{X}$ - mean of X $\bar{Y}$ - mean of Y

$b_{xy}$ – regression coefficient of X on Y $=\frac{\sum xy}{\sum y^2}$

$x$= X-$\bar{X}$ $y$ =Y-$\bar{Y}$

**Regression line of Y on X[Y depending on X]**

$$Y-\bar{Y}=b_{yx} [ X-\bar{X}]$$

Where,

$\bar{X}$ - mean of X $\bar{Y}$ - mean of Y

$b_{yx}$ – regression coefficient of X on Y $=\frac{\sum xy}{\sum x^2}$

$x$= X-$\bar{X}$ $y$ =Y-$\bar{Y}$

Note:

1. The regression coefficients $b_{xy}$ and $b_{yx}$ are of the same sign.
2. The correlation coefficient and the regression coefficients are connected by .r= $\sqrt{[b_{xy}\ b_{yx}]}$

Example: Calculate the regression lines for the following data.

| | | | | |
|---|---|---|---|---|
| X:6 | 2 | 10 | 4 | 8 |
| Y:9 | 11 | 5 | 8 | 7 |

Solution:

| X | Y | $x$= X-$\bar{X}$ | $y$ =Y-$\bar{Y}$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|---|---|
| 6 | 9 | 0 | 1 | 0 | 1 | 0 |
| 2 | 11 | -4 | 3 | 16 | 9 | -12 |
| 10 | 5 | 4 | -3 | 16 | 9 | -12 |
| 4 | 8 | -2 | 0 | 4 | 0 | 0 |
| 8 | 7 | 2 | -1 | 4 | 1 | -2 |
| ∑=30 | ∑=40 | ∑=0 | ∑=0 | ∑=40 | ∑=20 | ∑=-26 |

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6 \;\; ; \bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$$

Regression coefficients

$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{-26}{20} = -1.3$

$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{-26}{40} = -0.65$

**Regression line of X on Y[X depending on Y]**

X-6 =-1.3 [Y-8]

X =-1.3Y+1.64

**Regression line of Y on X[Y depending on X]**

Y-8=-0.65 [ X-6]

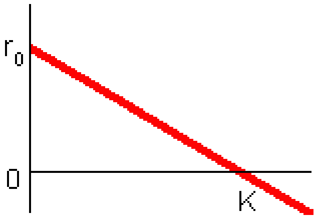
Y= -0.65X+11.9

# Logistic Regression:

**Logistic Model**

Logistic model was developed by Belgian mathematician Pierre Verhulst (1838) who suggested that the rate of population increase may be limited, i.e., it may depend on population density:

$$r = r_0(1 - \frac{N}{K})$$



At low densities (N < < 0), the population growth rate is maximal and equals to ro. Parameter ro can be interpreted as population growth rate in the absence of intra-specific competition.

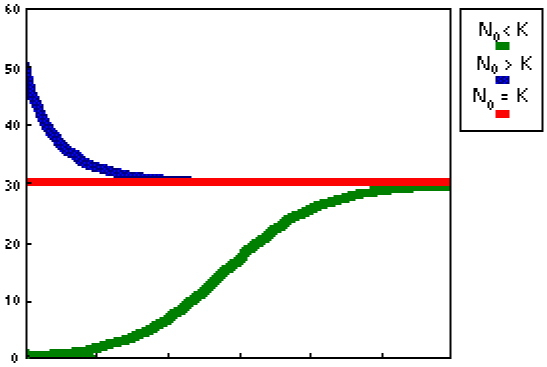
Population growth rate declines with population numbers, N, and reaches 0 when N = K. Parameter K is the upper limit of population growth and it is called carrying capacity. It is usually interpreted as the amount of resources expressed in the number of organisms that can be supported by these resources. If population numbers exceed K, then population growth rate becomes negative and population numbers decline. The dynamics of the population is described by the differential equation:

$$\frac{dN}{dt} = rN = r_0N(1 - \frac{N}{K})$$

which has the following solution:

$$N_t = \frac{N_0 \cdot K}{N_0 + (K - N_0) \cdot \exp(- r_0 \cdot t)}$$

**Three possible model outcomes**



1. Population increases and reaches a plateau (No < K). This is the logistic curve.
2. Population decreases and reaches a plateau (No > K)
3. Population does not change (No = K or No = 0)

Logistic model has two equilibria: N = 0 and N = K. The first equilibrium is unstable because any small deviation from this equilibrium will lead to population growth. The second equilibrium is stable because after small disturbance the population returns to this equilibrium state.

Logistic model combines two ecological processes: reproduction and competition. Both processes depend on population numbers (or density). The rate of both processes corresponds to the mass-action law with coefficients: ro for reproduction and ro/K for competition.

**Interpretation of parameters of the logistic model**

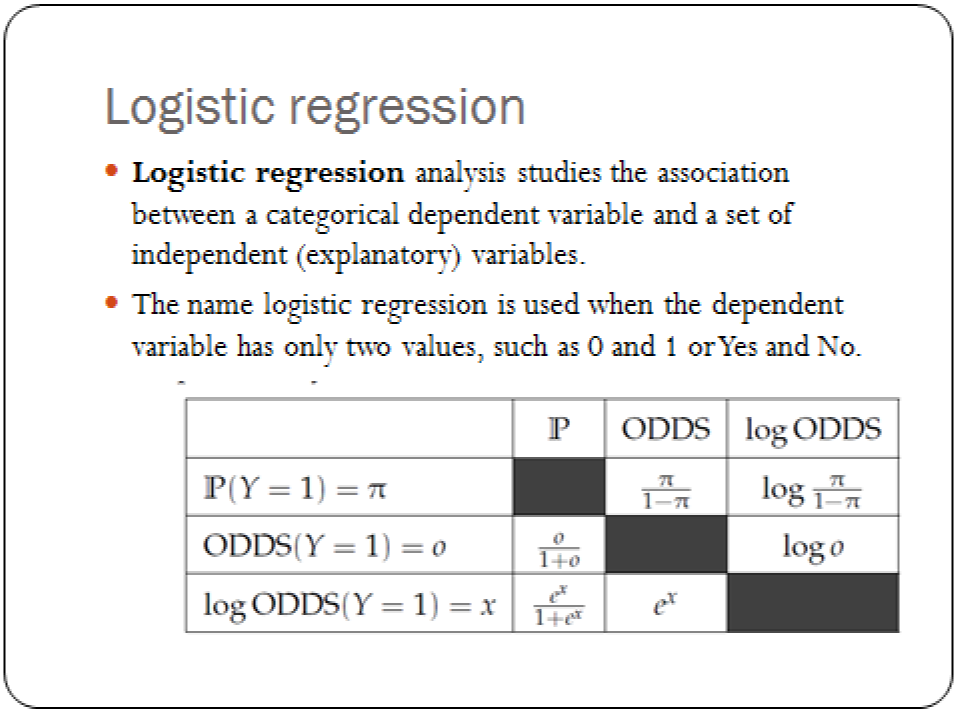
<u>Parameter ro</u> is relatively easy to interpret: this is the maximum possible rate of population growth which is the net effect of reproduction and mortality (excluding density-dependent mortality). Slowly reproducing organisms (elephants) have low ro and rapidly reproducing organisms (majority of pest insects) have high ro. The problem with the logistic model is that parameter ro controls not only population growth rate, but population decline rate (at N > K) as well. Here biological sense becomes not clear. It is not obvious that organisms with a low reproduction rate should die at the same slow rate. If reproduction is slow and mortality is fast, then the logistic model will not work.

<u>Parameter K</u> has biological meaning for populations with a strong interaction among individuals that controls their reproduction. For example, rodents have social structure that controls reproduction, birds have territoriality, plants compete for space and light. However, parameter K has no clear meaning for organisms whose population dynamics is determined by the balance of reproduction and

mortality processes (e.g., most insect populations). In this case the equilibrium population density does not necessary correspond to the amount of resources; thus, the term "carrying capacity" becomes confusing. For example, equilibrium density may depend on mortality caused by natural enemies.

Logistic regression:

- Odds: The odds in favor of an event is the ratio of the number of ways the outcome **can** occur to the number of ways the outcome **cannot** occur.
- Difference between fraction and ration.
- Probability(event)= No.of favorable outcomes/Total no.of outcomes
- Odds(event)= No.of Favorable outcomes/No.of Unfavorable outcomes
- So , odds=p/(1-p).
- Logit function is logit(p)=log[p/1-p]
- Logistic function is logistic (x)=exp(x)/1+exp(x).



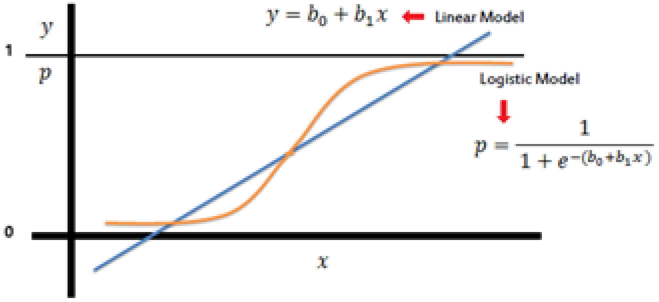
## Logistic regression

- **Logistic regression** analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables.
- The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 orYes and No.

| | $\mathbb{P}$ | ODDS | log ODDS |
|---|---|---|---|
| $\mathbb{P}(Y=1)=\pi$ | ■ | $\frac{\pi}{1-\pi}$ | $\log\frac{\pi}{1-\pi}$ |
| $\text{ODDS}(Y=1)=o$ | $\frac{o}{1+o}$ | ■ | $\log o$ |
| $\log\text{ODDS}(Y=1)=x$ | $\frac{e^x}{1+e^x}$ | $e^x$ | ■ |

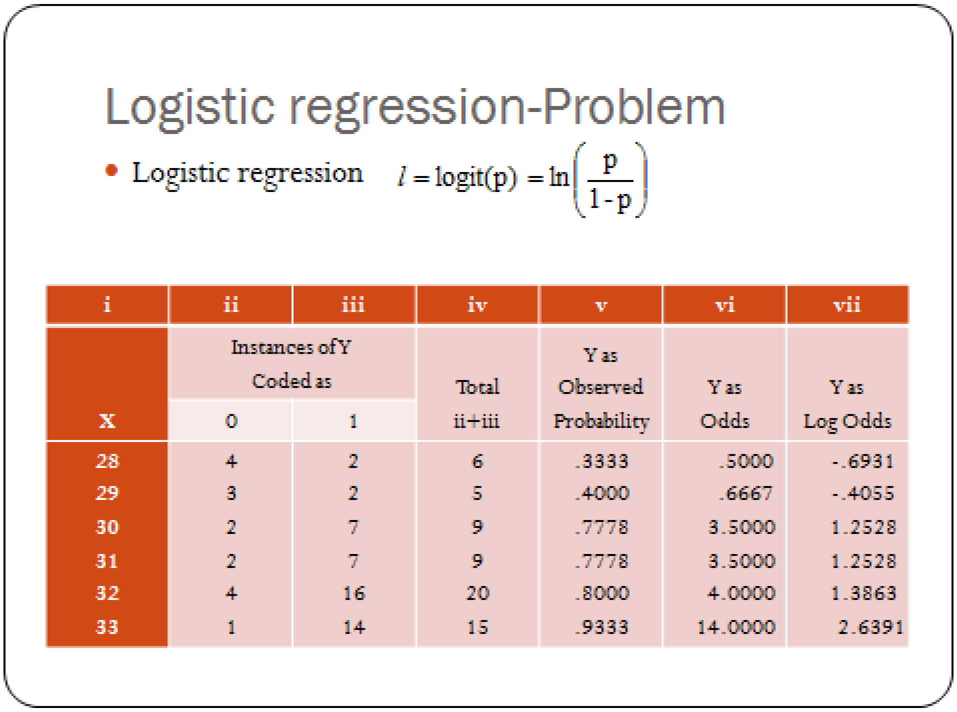
## Logistic regression

- **Logistic regression** measures the **relationship between** the categorical dependent variable and one or more independent variables by estimating probabilities using a **logistic function**,



## Logistic regression

- So, logistic regression is
- Log[odds(p)]=β0+β1x[In case of single explanatory variable]
- $\beta_0$ is the **log-odds** in favour of Y=1 when
- X1=X2=…=Xp=0.
- $\beta_0$ is such that $\exp(\beta_0)$ is the **odds** in favour of Y=1
- when X1=X2=…=Xp=0.
- $\beta_0$ is such that $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ is the **probability** that Y=1
- when X1=X2=…=Xp=0.



## Logistic regression-Problem

- Logistic regression $l = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$

| i | ii | iii | iv | v | vi | vii |
|---|---|---|---|---|---|---|
| | Instances of Y Coded as | | Total | Y as Observed Probability | Y as Odds | Y as Log Odds |
| X | 0 | 1 | ii+iii | | | |
| 28 | 4 | 2 | 6 | .3333 | .5000 | -.6931 |
| 29 | 3 | 2 | 5 | .4000 | .6667 | -.4055 |
| 30 | 2 | 7 | 9 | .7778 | 3.5000 | 1.2528 |
| 31 | 2 | 7 | 9 | .7778 | 3.5000 | 1.2528 |
| 32 | 4 | 16 | 20 | .8000 | 4.0000 | 1.3863 |
| 33 | 1 | 14 | 15 | .9333 | 14.0000 | 2.6391 |

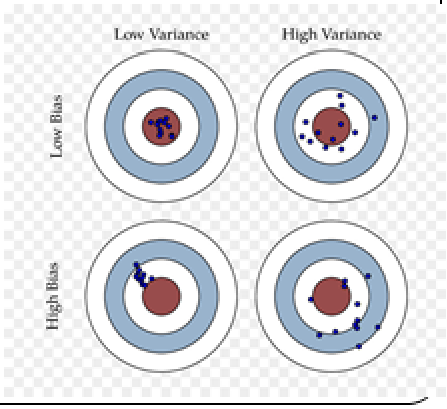
- Once this initial linear regression is obtained, the predicted log odds for any particular value of X can then be translated back into a predicted probability value. Thus, for X=31 in the present example, the predicted log odds would be
- log[odds] = -17.2086+(.5934x31) = 1.1868
- The corresponding predicted odds would be
- odds = exp(log[odds]) = exp(1.1868)=3.2766
- And the corresponding predicted probability would be
- probability = odds/(1+odds)=3.2766/(1+3.2766)
- = .7662

## Linear basis function models:

- LBF models
- These are generalizations of the linear regressions in that the coefficients of the predictors are not constants but functions.
- In general, we do linear regression of t on ϕ1(x), ϕ2(x), . . . , $\phi_{M-1}$(x), where the ϕj are basis functions, that we have selected to allow for a non-linear function of x. This gives the following model: $y(x, w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x) = w^T \phi(x)$
- where w is the vector of all M regression coefficients (including the intercept, w0) and ϕ(x) is the vector of all basis function values at input x, including ϕ0(x) = 1 for the intercept.

## Bias-Variance

- Bias is the difference between the average prediction of our model and the correct value
- Variance is the variability of model prediction for a given data point . High variance means tuned to training data and couldn't generalize.



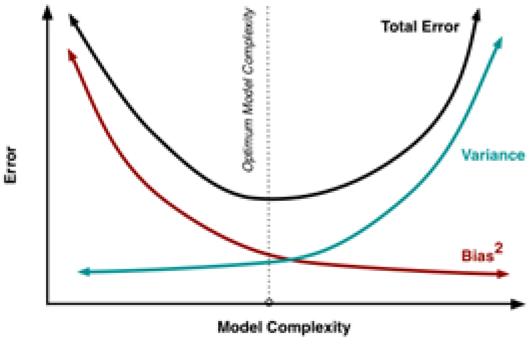
- BV-Decomposition

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$

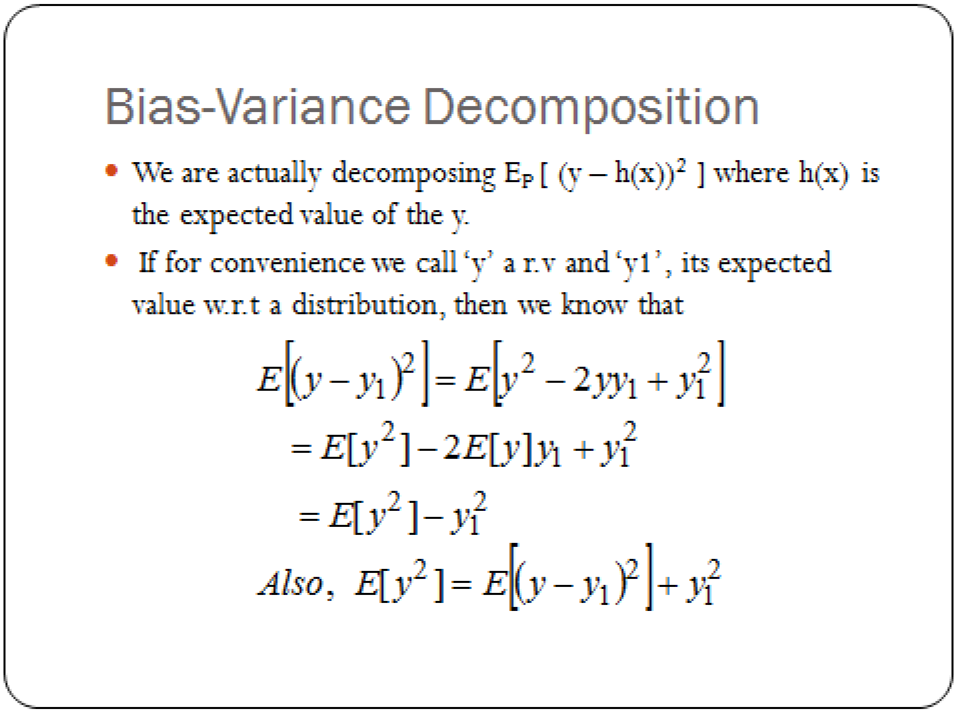
## Bias-Variance

- If our model is too simple and has very few parameters then it may have high bias and low variance.
- If our model has large number of parameters then it's going to have high variance and low bias.
- This tradeoff in complexity is why there is a tradeoff between bias and variance.



## Bias-Variance Decomposition:

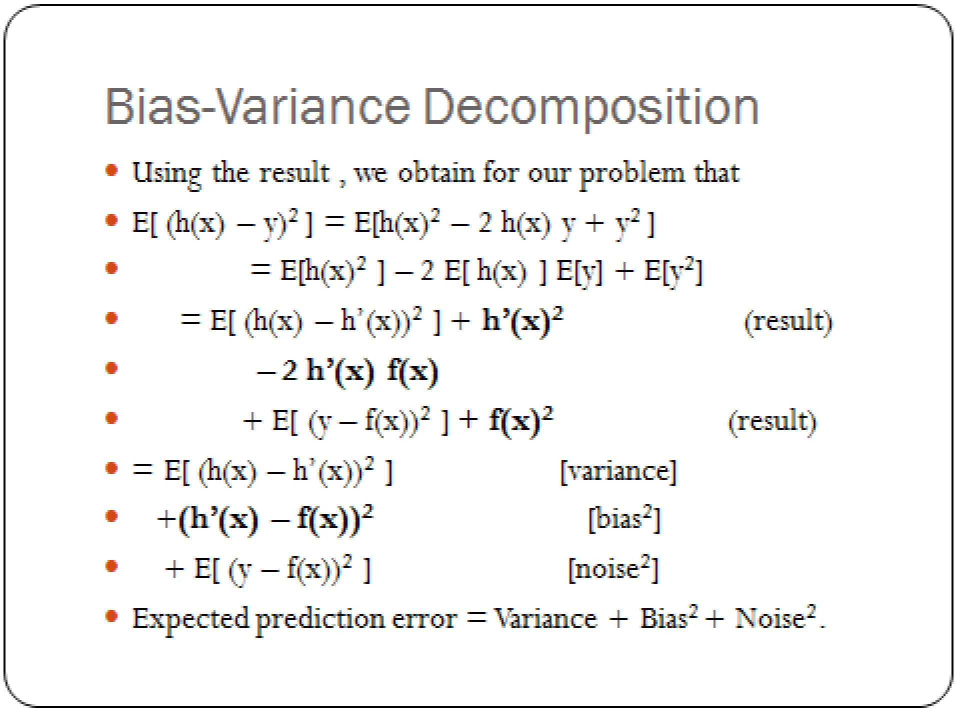
- In regression, we assume that the actual relation is y=f(x)+ε, where we assume that the error is normally distributed with zero mean and SD=σ. But the regressed line is our fit, a best line, based on the data , denoted by h(x)=w*x + b.
- The “best” tag for the line is because of the fact we minimize the squared error $\sum_i [y_i - h(x_i)]^2$
- For the data (yi,xi).
- Now we’ll decompose the expected value of the square for a random data point.



### Bias-Variance Decomposition

- We are actually decomposing E$_P$ [ (y – h(x))$^2$ ] where h(x) is the expected value of the y.
- If for convenience we call ‘y’ a r.v and ‘y1’, its expected value w.r.t a distribution, then we know that

$$E\left[(y-y_1)^2\right] = E\left[y^2 - 2yy_1 + y_1^2\right]$$
$$= E[y^2] - 2E[y]y_1 + y_1^2$$
$$= E[y^2] - y_1^2$$
$$Also,\ E[y^2] = E\left[(y-y_1)^2\right] + y_1^2$$



## Bias-Variance Decomposition

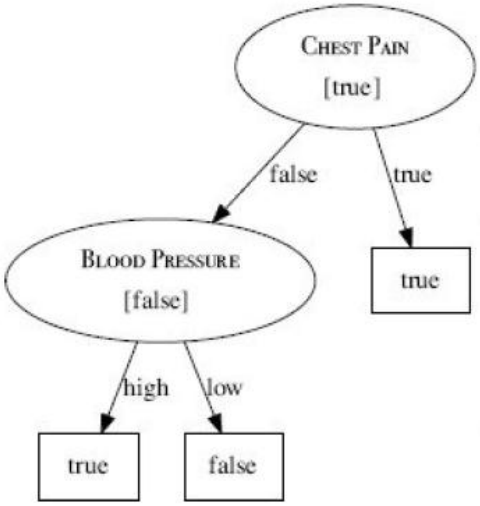
- Using the result , we obtain for our problem that
- $E[ (h(x) – y)^2 ] = E[h(x)^2 – 2 h(x) y + y^2 ]$
- $= E[h(x)^2 ] – 2 E[ h(x) ] E[y] + E[y^2]$
- $= E[ (h(x) – h'(x))^2 ] + \mathbf{h'(x)^2}$ (result)
- $\mathbf{– 2 h'(x) f(x)}$
- $+ E[ (y – f(x))^2 ] + \mathbf{f(x)^2}$ (result)
- $= E[ (h(x) – h'(x))^2 ]$ [variance]
- $\mathbf{+(h'(x) – f(x))^2}$ [bias$^2$]
- $+ E[ (y – f(x))^2 ]$ [noise$^2$]
- Expected prediction error = Variance + Bias$^2$ + Noise$^2$.

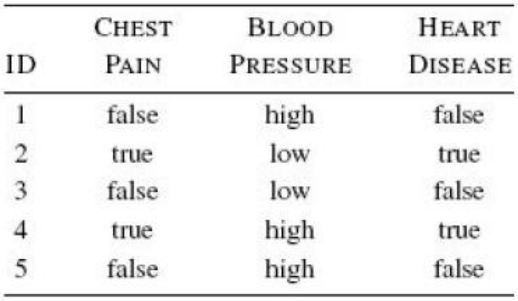
- Variance: $E[ (h(x) – h'(x))^2 ]$
- Describes how much h(x) varies from one training set S to another
- Bias: [h'(x) – f(x)]
- Describes the average error of h(x).
- Noise: $E[ (y – f(x))^2 ] = E[\varepsilon^2] = \sigma^2$
- Describes how much y varies from f(x)
- For each data point x, we will now have
- the observed corresponding value y and several predictions y1, …, yK.
- Compute the average prediction h'.
- Estimate bias as (h' – y)
- Estimate variance as $\Sigma k\ (yk – h')^2/(K – 1)$
- Assume noise is 0

- Assume noise is 0 ?
- Because in practice we usually bootstrap data and hence are not real.
- If we have multiple data points with the same x value, then we can estimate the noise.
- We can also estimate noise by pooling y values from nearby x values.
- Models that fit the data poorly have high bias: “inflexible models” such as linear regression, regression stumps
- Models that can fit the data very well have low bias but high variance: “flexible” models such as nearest neighbor regression, regression trees
- This suggests that bagging of a flexible model can reduce the variance while benefiting from the low bias

# Decision Trees

**decision trees**, the fundamental structure used in information-based machine learning, before presenting the fundamental
measures of information content that are used: **entropy** and **information gain**.

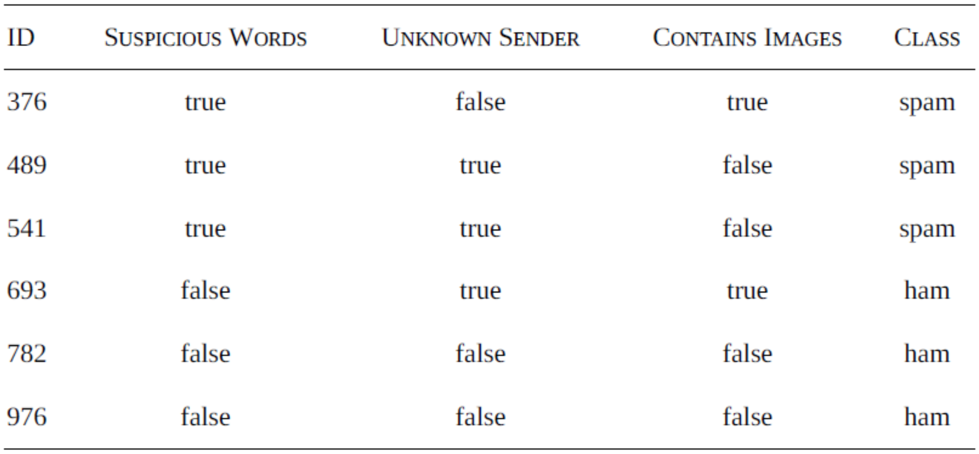




| ID | CHEST PAIN | BLOOD PRESSURE | HEART DISEASE |
|---|---|---|---|
| 1 | false | high | false |
| 2 | true | low | true |
| 3 | false | low | false |
| 4 | true | high | true |
| 5 | false | high | false |

Consider the spam data:

An email spam prediction dataset.



| ID | SUSPICIOUS WORDS | UNKNOWN SENDER | CONTAINS IMAGES | CLASS |
|---|---|---|---|---|
| 376 | true | false | true | spam |
| 489 | true | true | false | spam |
| 541 | true | true | false | spam |
| 693 | false | true | true | ham |
| 782 | false | false | false | ham |
| 976 | false | false | false | ham |

Entropy :

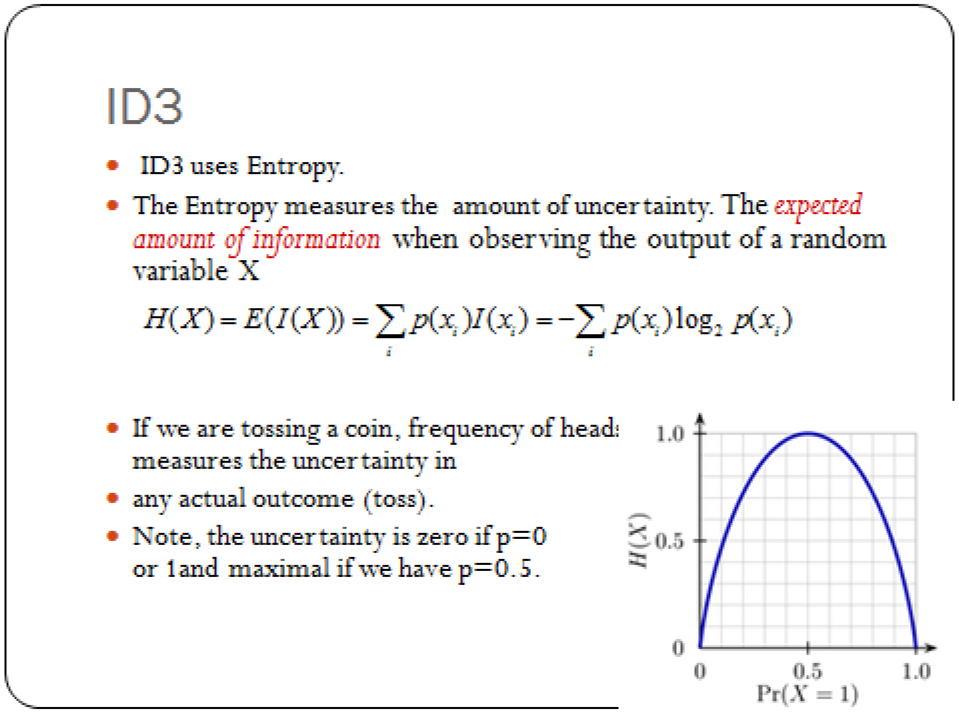
entropy model defines a computational measure of the impurity of the elements in a set. An easy way to understand the entropy of a
set is to think in terms of the uncertainty associated with guessing the result if you were to make a random selection from the set. For example, if you were to randomly select a card from the set in Figure4.5(a) you would have zero uncertainty, as you would know for
sure that you would select an ace of spades. So, this set has zero entropy. If, however, you were to randomly select an element from the set in Figure 4.5(f) you would be very uncertain about any prediction as there are twelve possible outcomes, each of which is equally likely. This is why this set has very high entropy. The other sets in Figure 4.5
have entropy values between these two extremes.

Shannon's model of entropy is a weighted sum of the logs of the probabilities of each possible outcome when we make a random selection from a set. The weights used in the sum are the probabilities of the outcomes themselves so that outcomes with high probabilities contribute more to the overall entropy of a set than outcomes with low probabilities. Shannon's model of entropy is defined as $H(t) = -\sum_{i=1}^{l} P(t = i) * \log(P(t = i))$

where *P*(*t* = *i*) is the probability that the outcome of randomly selecting an element *t* is the type *i*, *l* is the number of different types of things in the set, and *s* is an arbitrary logarithmic base. The minus sign at the beginning of the equation is simply added to convert the negative numbers returned by the log function to positive ones (as described above). We will always use 2 as the base, *s*, when we calculate entropy, which means that we measure entropy in **bits**. Equation  is the cornerstone of modern **information theory** and is an excellent measure of the impurity, **heterogeneity**, of a set.

To understand how Shannon's entropy model works, consider the example of a set of 52 different playing cards. The probability of randomly selecting any specific card *i* from this set, *P*(*card* = *i*), is quite low, just . The entropy of the set of 52 playing cards is calculated as

$$H(card) = -\sum_{i=1}^{52} P(card = i) * \log(P(card = i)) = 5.7\,bits$$



ID3

- ID3 uses Entropy.
- The Entropy measures the amount of uncertainty. The *expected amount of information* when observing the output of a random variable X

$$H(X) = E(I(X)) = \sum_i p(x_i) I(x_i) = -\sum_i p(x_i) \log_2 p(x_i)$$

- If we are tossing a coin, frequency of head: measures the uncertainty in
- any actual outcome (toss).
- Note, the uncertainty is zero if p=0 or 1and maximal if we have p=0.5.

1.0
0.5
0
$H(X)$
0
0.5
1.0
$\Pr(X = 1)$

## Information Gain:

**information gain** is a measure of the reduction in the overall entropy of a set of instances that is achieved by testing on a descriptive feature. Computing information gain is a three-step process:

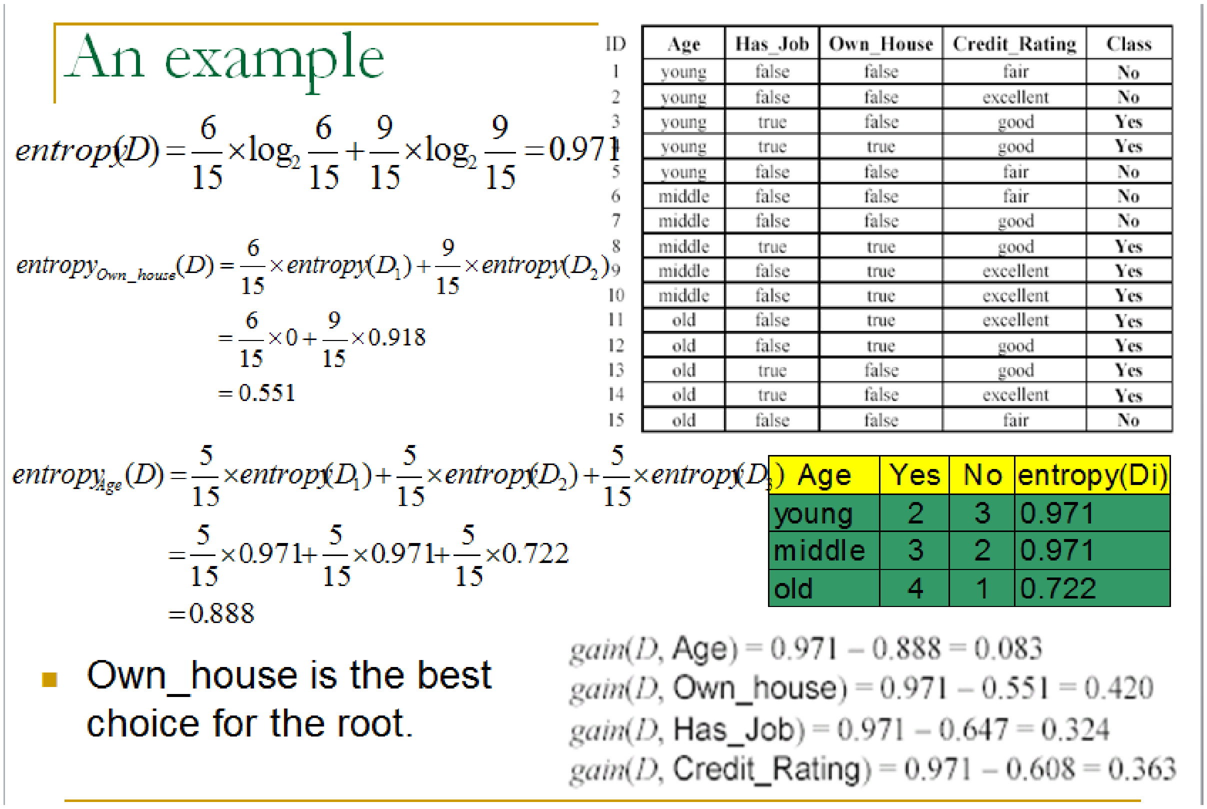
1. Compute the entropy of the original dataset with respect to the target feature. This gives us an measure of how much information is required in order to organize the dataset into pure sets.
2. For each descriptive feature, create the sets that result by partitioning the instances in the dataset using their feature values, and then sum the entropy scores of each of these sets. This gives a measure of the information that remains required to organize the instances into pure sets after we have split them using the descriptive feature.
3. Subtract the remaining entropy value (computed in step 2) from the original entropy value (computed in step 1) to give the information gain.

## Tree Pruning:

A predictive model **overfits** the training set when at least some of the predictions it returns are based on spurious patterns present in the training data used to induce the model.

Overfitting happens for a number of reasons, including **sampling variance**18 and noise in the training set. decision trees overfit by splitting the data on irrelevant features that only appear relevant due to noise or sampling variance in the training data. The likelihood of overfitting occurring increases as a tree gets deeper because the resulting predictions are based on smaller and smaller subsets as the dataset is partitioned after each feature test in the path.

**Tree pruning** identifies and removes subtrees within a decision tree that are likely to be due to noise and sample variance in the training set used to induce it. In cases where a subtree is deemed to be overfitting, pruning the subtree means replacing the subtree with a leaf node that makes a prediction based on the majority target feature level (or average target feature value) of the dataset created by merging the instances from all the leaf nodes in the subtree.



## An example

$$entropy(D) = \frac{6}{15} \times \log_2 \frac{6}{15} + \frac{9}{15} \times \log_2 \frac{9}{15} = 0.971$$

| ID | Age | Has_Job | Own_House | Credit_Rating | Class |
|---|---|---|---|---|---|
| 1 | young | false | false | fair | No |
| 2 | young | false | false | excellent | No |
| 3 | young | true | false | good | Yes |
| 4 | young | true | true | good | Yes |
| 5 | young | false | false | fair | No |
| 6 | middle | false | false | fair | No |
| 7 | middle | false | false | good | No |
| 8 | middle | true | true | good | Yes |
| 9 | middle | false | true | excellent | Yes |
| 10 | middle | false | true | excellent | Yes |
| 11 | old | false | true | excellent | Yes |
| 12 | old | false | true | good | Yes |
| 13 | old | true | false | good | Yes |
| 14 | old | true | false | excellent | Yes |
| 15 | old | false | false | fair | No |

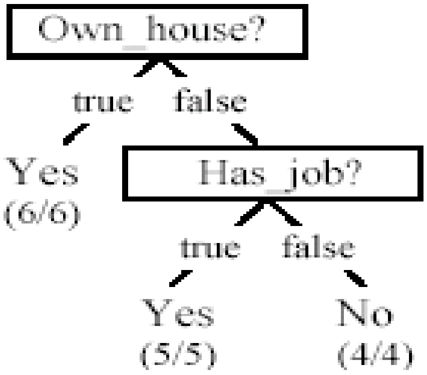
$$\begin{aligned} entropy_{Own\_house}(D) &= \frac{6}{15} \times entropy(D_1) + \frac{9}{15} \times entropy(D_2) \\ &= \frac{6}{15} \times 0 + \frac{9}{15} \times 0.918 \\ &= 0.551 \end{aligned}$$

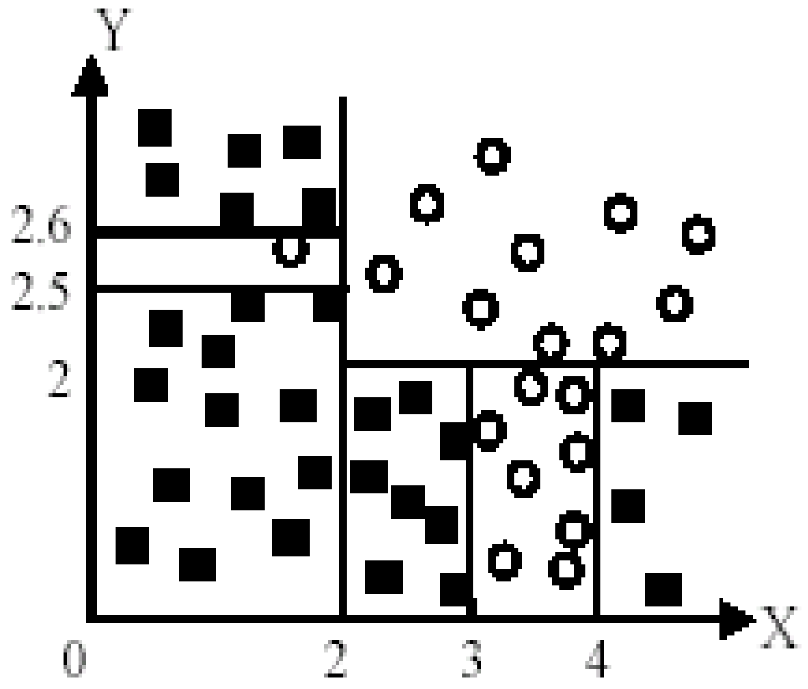
$$\begin{aligned} entropy_{Age}(D) &= \frac{5}{15} \times entropy(D_1) + \frac{5}{15} \times entropy(D_2) + \frac{5}{15} \times entropy(D_3) \\ &= \frac{5}{15} \times 0.971 + \frac{5}{15} \times 0.971 + \frac{5}{15} \times 0.722 \\ &= 0.888 \end{aligned}$$

| Age | Yes | No | entropy(Di) |
|---|---|---|---|
| young | 2 | 3 | 0.971 |
| middle | 3 | 2 | 0.971 |
| old | 4 | 1 | 0.722 |

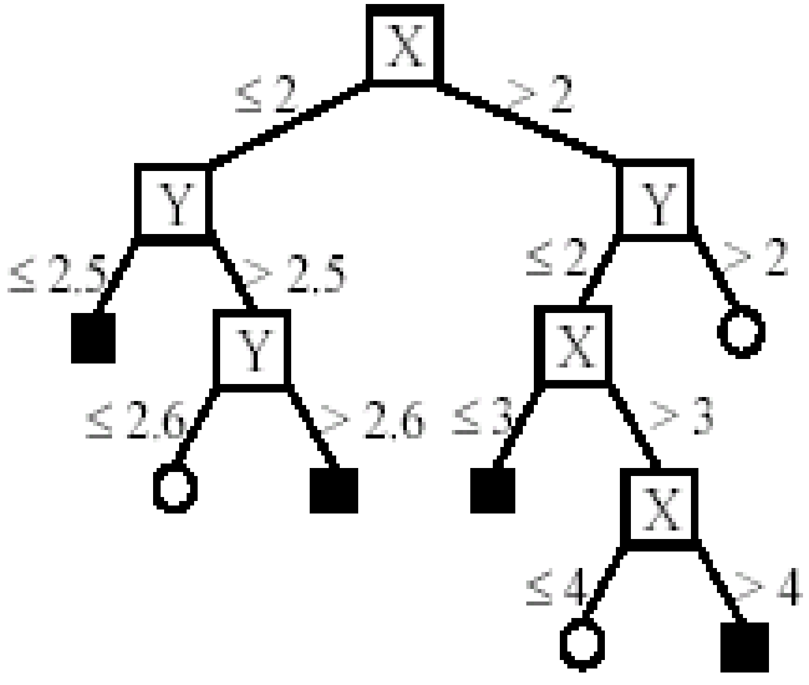
- Own_house is the best choice for the root.

$$gain(D, \text{Age}) = 0.971 - 0.888 = 0.083$$
$$gain(D, \text{Own\_house}) = 0.971 - 0.551 = 0.420$$
$$gain(D, \text{Has\_Job}) = 0.971 - 0.647 = 0.324$$
$$gain(D, \text{Credit\_Rating}) = 0.971 - 0.608 = 0.363$$





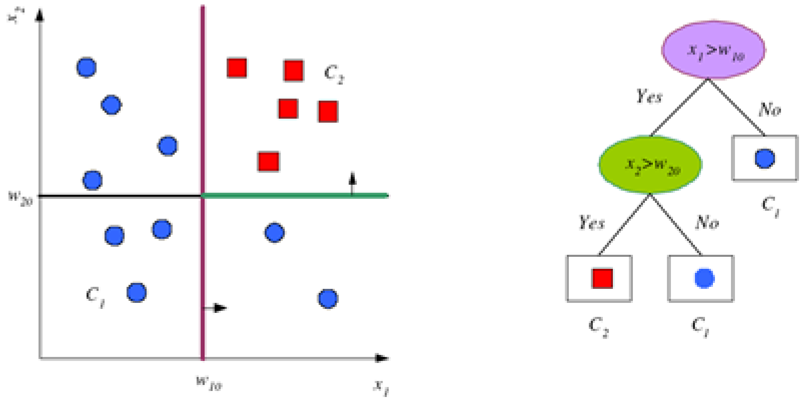
(A) A partition of the data space



(B). The decision tree

# Decision trees

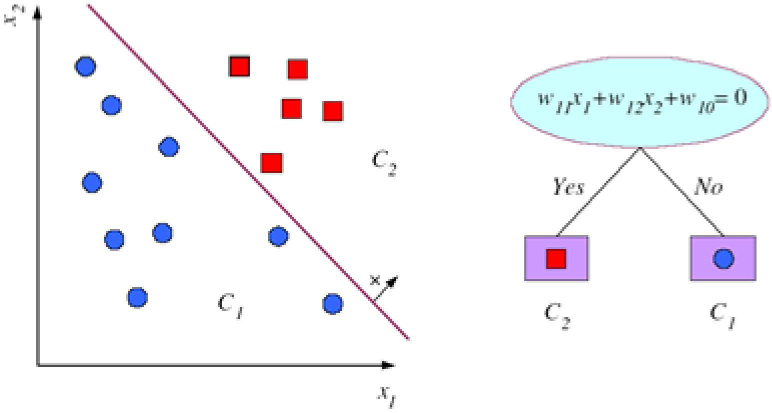
- Look at the picture



- Uni-variate, Binary split, Leaves are in classification.

# Decision Trees
# Multivariate Split

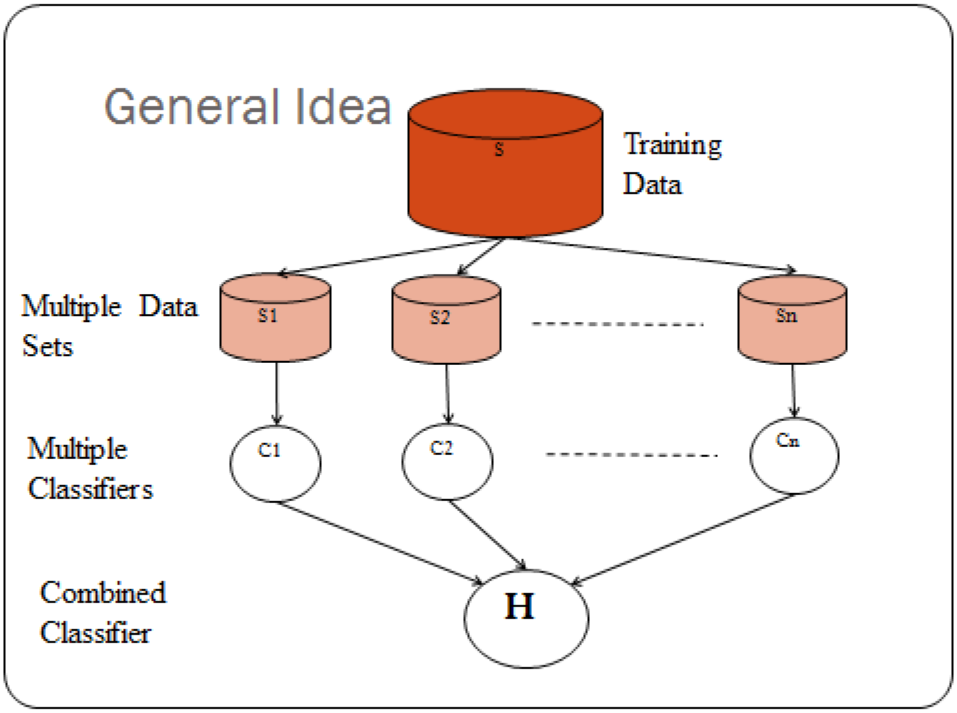
- Regression like lines based.



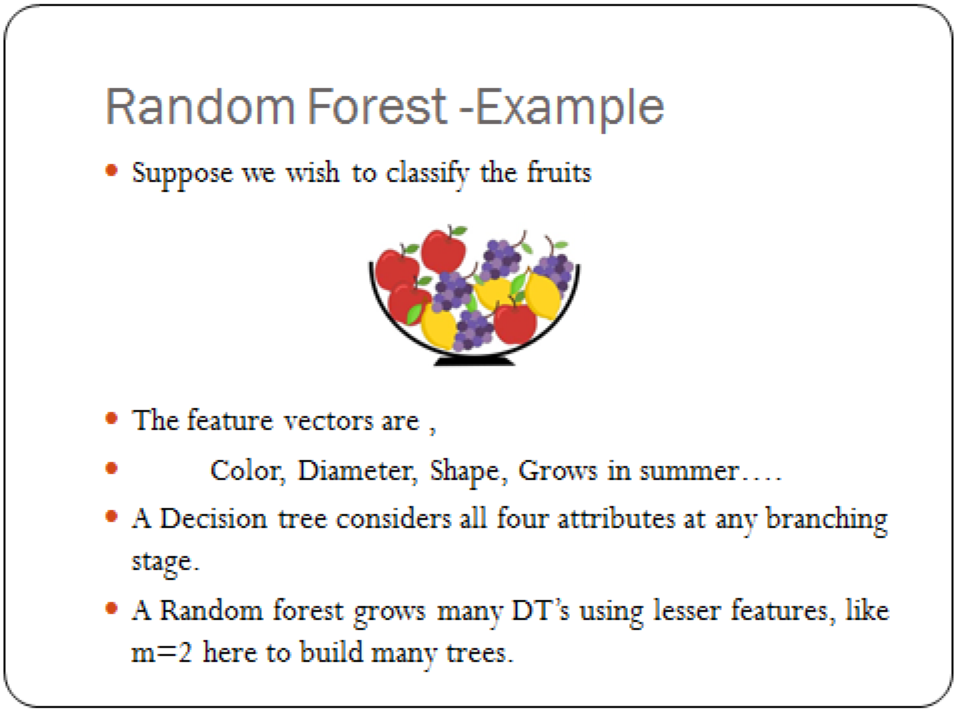
## Random Forest:

- Issues with a Decision tree
  - Easy to build but not flexible and efficient, Over-fitting
- Many trees, with a subset of features(RF)
- Aggregation of all trees.
- Bootstrapping, Bagging

- **Random forest** is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.
- For a given dataset(with p feature vectors), create many datasets from it using random sampling with replacement(Bootstrapping) , create one Decision tree using each generated dataset with m(<p) randomly selected feature vectors. Run a new case on each of the DT 's , then use mode of the classes(Classification) or average (Regression) to finalize.
- For classification, the default value for *m* is √*p* and the minimum node size is one.
- For regression, the default value for m is *p/3* and the minimum node size is five.

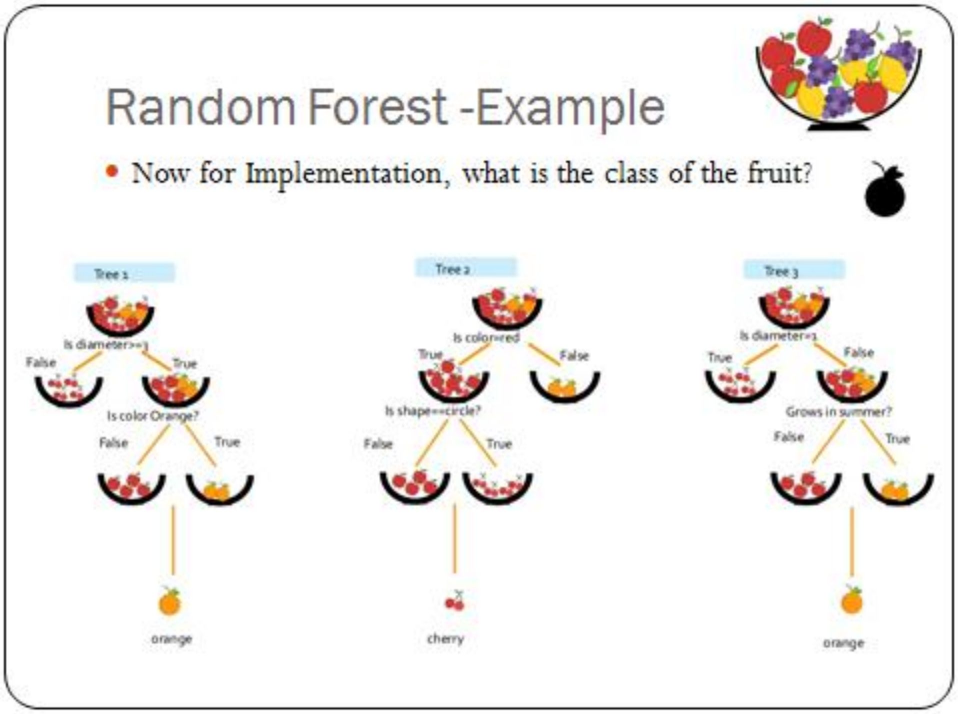


- We are creating a forest with many trees . Each DT may be simple model but together the ensemble is a better model as can be seen by the reduction in error.
- If for example we have 25 DT's, each independent but with identical error of 0.35which is either the product of all the errors or a Binomial distributed value, will surely be very less than 0.35.



## Random Forest -Example

- Suppose we wish to classify the fruits
- The feature vectors are ,
- Color, Diameter, Shape, Grows in summer….
- A Decision tree considers all four attributes at any branching stage.
- A Random forest grows many DT's using lesser features, like m=2 here to build many trees.



## Random Forest -Example

- Now for Implementation, what is the class of the fruit?

Random Forest

Advantages:

The advantages of random forest are:

- It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
- It runs efficiently on large databases.

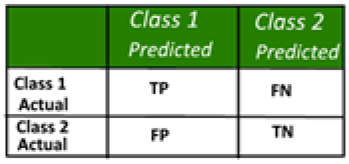
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing error in class population unbalanced data sets.
- Generated forests can be saved for future use on other data.
- Prototypes are computed that give information about the relation between the variables and the classification.
- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or (by scaling) give interesting views of the data.
- The capabilities of the above can be extended to unlabeled data, leading to unsupervised clustering, data views and outlier detection.
- It offers an experimental method for detecting variable interactions.

Random Forest
Disadvantages:

- Random forests have been observed to overfit for some datasets with noisy classification/regression tasks.
- For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.
- Note: For each tree grown, 33-36% of samples are not selected in bootstrap, called out of bootstrap (OOB) samples

# Confusion matrix example

- CM: Class 1 : Positive
- Class 2 : Negative



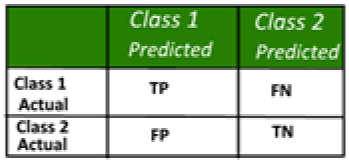
| | Class 1 Predicted | Class 2 Predicted |
|---|---|---|
| Class 1 Actual | TP | FN |
| Class 2 Actual | FP | TN |

- **Definition of the Terms:**

- Positive (P) : Observation is positive (for example: is an apple).
- Negative (N) : Observation is not positive (for example: is not an apple).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

# Confusion matrix example

- CM: Performance of a Classification model
- Class 1 : Positive; Class 2 : Negative



| | Class 1 Predicted | Class 2 Predicted |
|---|---|---|
| Class 1 Actual | TP | FN |
| Class 2 Actual | FP | TN |

- **Definition of the Terms:**

- Accuracy : [TP+TN]/[TP+FN+FP+TN]
- How many were correctly labeled.
- Specificity : TN/[TN+FP]
- Of all people from class2, how many were correctly labeled.
- Sensitivity/Recall : TP/[TP+FN]
  - High Recall indicates the class is correctly recognized.
  - Of all people from class 1, how many were correctly labeled.
- Precision : TP/[TP+FP]
  - High Precision indicates an example labeled as positive is indeed positive

## Confusion matrix example

- CM:

| N=165 | Pred No | Pred Yes | |
|---|---|---|---|
| Actual No | TN=50 | FP=10 | 60 |
| Actual Yes | FN=5 | TP=100 | 105 |
| | 55 | 110 | N=165 |

- Accuracy = (TP + TN) / (TP + TN + FP + FN)

  =(100 + 50) / (100 + 5 + 10 + 50) = 0.90
- Recall = TP / (TP + FN) = 100 / (100 + 5) = 0.95
- Precision = TP / (TP + FP)=100/ (100+10) = 0.91
- **F-measure: Harmonic mean b/w precision and recall.**

  Fmeasure = (2 * Recall * Precision) / (Recall + Precision)

  = (2 * 0.95 * 0.91) / (0.91 + 0.95) = 0.92

  High F score indicates balance between Precision and recall

## Confusion matrix example

- CM:
- Accuracy is not relevant for Imbalanced data(Total frequency of one class very high than the other) and in such cases we use precision and recall.

| Actual\Predicted | Positive | Negative | | |
|---|---|---|---|---|
| Positive | TP | FN | Non-Healthy/Disease | Sensitivity /Recall |
| Negative | FP | TN | Healthy | Specificity |
| | Precision | | N=Total | |

-

Accuracy