



INTRODUCTION TO DATA ANALYTICS

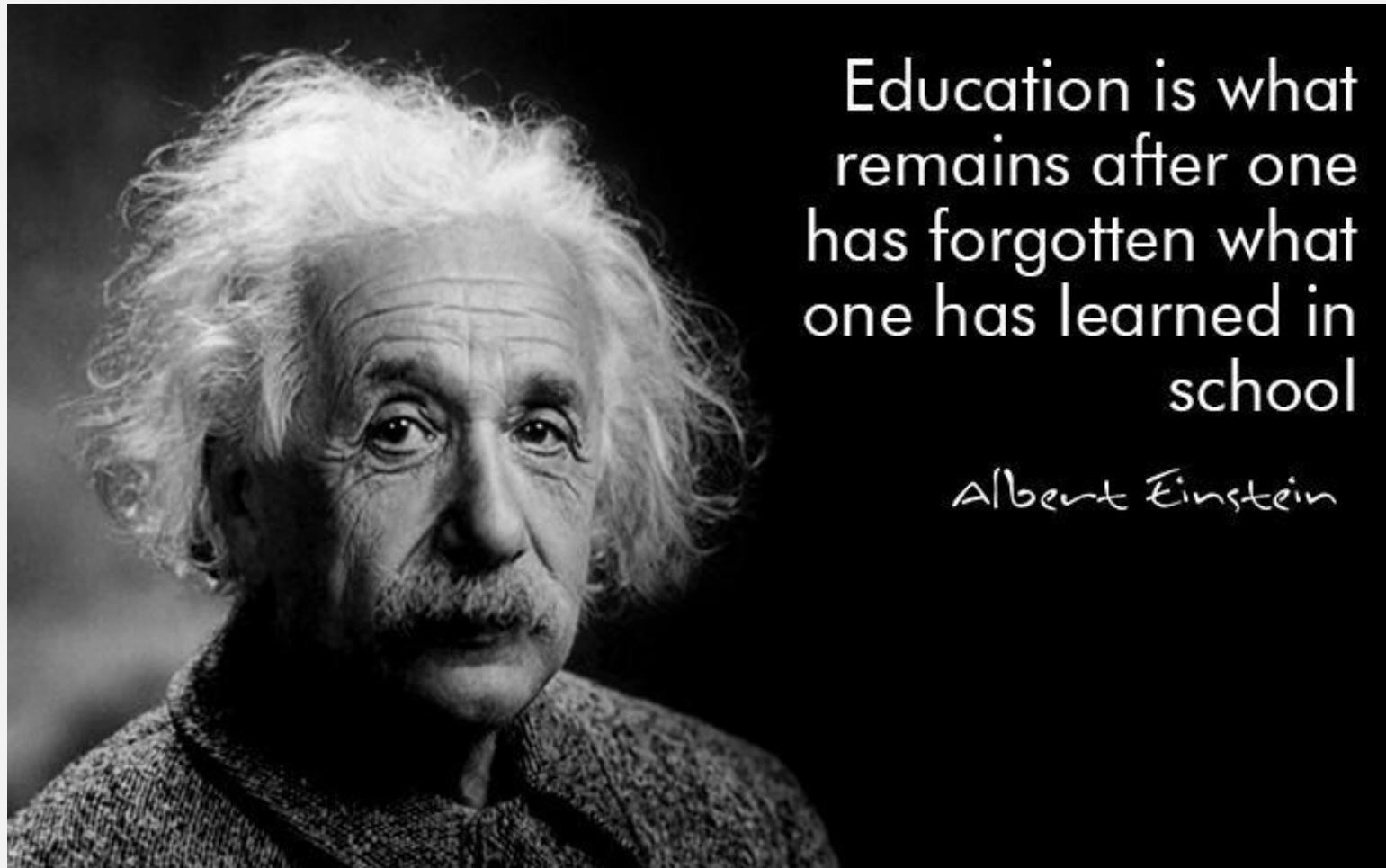
Class #12 **Estimation**

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

QUOTE OF THE DAY..



Education is what
remains after one
has forgotten what
one has learned in
school

Albert Einstein

IN THIS PRESENTATION...

- POINT ESTIMATION
- GENERAL CONCEPTS OF POINT ESTIMATION
 - **Unbiased Estimators**
 - **Variance of a Point Estimator**
 - **Standard Error: Reporting a Point Estimate**
 - **Mean Squared Error of an Estimator**
- METHODS OF POINT ESTIMATION
 - **Method of Moments**
 - **Method of Maximum Likelihood**

POINT ESTIMATION

- Statistical inference is always focused on drawing conclusions about one or more parameters of a population.
- An important part of this process is obtaining estimates of the parameters.
- Suppose that we want to obtain a **point estimate** (a reasonable value) of a population parameter.
- Any function of the observation, or any **statistic**, is also a random variable. For example, the sample mean and the sample variance are statistics and they are also random variables.
- Since a statistic is a random variable, it has a probability distribution. We call the probability distribution of a statistic a **sampling distribution**.

POINT ESTIMATION

- When discussing inference problems, it is convenient to have a general symbol to represent the parameter of interest.
- We will use the Greek symbol θ (theta) to represent the parameter. The symbol θ can represent the mean μ , the variance σ^2 , or any parameter of interest to us.
- The objective of point estimation is to select a single number, based on sample data, that is the most plausible value for θ .
- A numerical value of a sample statistic will be used as the point estimate. The statistic $\hat{\Theta}$ is called a **point estimator** of θ .

POINT ESTIMATOR

Definition 12.1: Point Estimator

A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\widehat{\Theta}$. The statistic $\widehat{\Theta}$ is called the **point estimator**.

- **Example:**
- Suppose that the random variable X is normally distributed with an unknown mean μ . The sample mean is a point estimator of the unknown population mean μ . That is, $\hat{\mu} = \bar{X}$.

POINT ESTIMATOR

Estimation problems occur frequently in engineering. We often need to estimate

- The mean μ of a single population
- The variance σ^2 (or standard deviation σ) of a single population
- The proportion p of items in a population that belong to a class of interest.
- The difference in means of two populations, $\mu_1 - \mu_2$
- The difference in two population proportions, $p_1 - p_2$

POINT ESTIMATOR

Reasonable point estimates of these parameters are as follows:

- For μ , the estimate is $\hat{\mu} = \bar{X}$, the sample mean.
- For σ^2 , the estimate is $\hat{\sigma}^2 = s^2$, the sample variance.
- For p , the estimate is $\hat{p} = x/n$, the sample proportion, where x is the number of items in a random sample of size n that belong to the class of interest.
- For $\mu_1 - \mu_2$, the estimate is $\hat{\mu}_1 - \hat{\mu}_2 = \hat{x}_1 - \hat{x}_2$, the difference between the sample means of two independent random samples.
- For $p_1 - p_2$, the estimate is $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions computed from two independent random samples.

CENTRAL LIMIT THEOREM

Definition 12.2: Approximate Sampling Distribution of a Difference in Sample Means

If we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , and if \bar{X}_1 and \bar{X}_2 are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is approximately standard normal, if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of Z is exactly standard normal.

UNBIASED ESTIMATORS

An estimator should be “close” in some sense to the true value of the unknown parameter. Formally, we say that $\widehat{\Theta}$ is an unbiased estimator of θ if the expected value of $\widehat{\Theta}$ is equal to θ . This is equivalent to saying that the mean of the probability distribution of $\widehat{\Theta}$ (or the mean of the sampling distribution of $\widehat{\Theta}$) is equal to θ .

Definition 12.3: Bias of an Estimator

The point estimator $\widehat{\Theta}$ is an **unbiased estimator** for the parameter θ if

$$E(\widehat{\Theta}) = \theta$$

If the estimator is not unbiased, then the difference

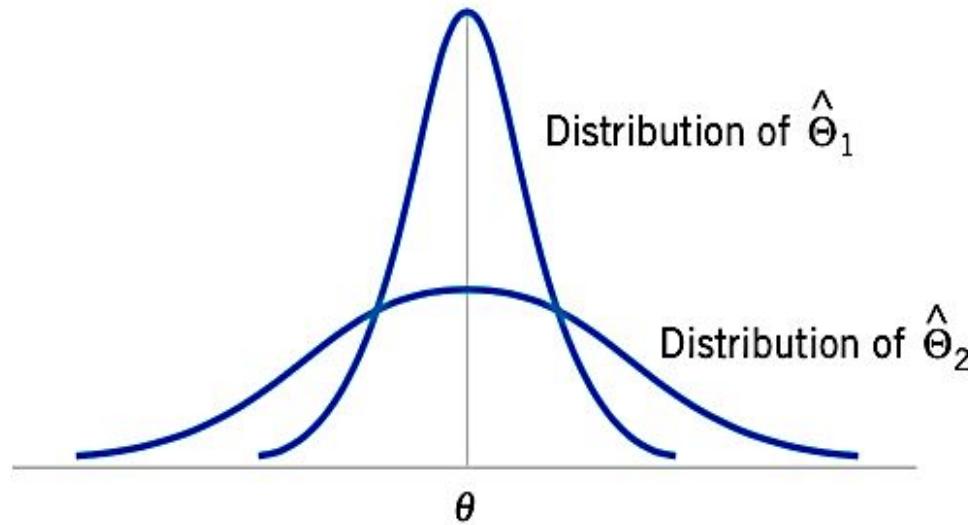
$$E(\widehat{\Theta}) - \theta$$

is called the **bias** of the estimator $\widehat{\Theta}$.

When an estimator is unbiased, the bias is zero; that is, $E(\widehat{\Theta}) - \theta = 0$.

VARIANCE OF A POINT ESTIMATOR

A logical principle of estimation, when selecting among several estimators, is to choose the estimator that has minimum variance.



If we consider all unbiased estimators of θ , the one with the smallest variance is called the **minimum variance unbiased estimator** (MVUE).

STANDARD ERROR

When the numerical value or point estimate of a parameter is reported, it is usually desirable to give some idea of the precision of estimation. The measure of precision usually employed is the standard error of the estimator that has been used.

The standard error of an estimator $\hat{\Theta}$ is its standard deviation, given by $\sigma_{\hat{\Theta}} = \sqrt{V(\hat{\Theta})}$. If the standard error involves unknown parameters that can be estimated, substitution of those values into $\sigma_{\hat{\Theta}}$ produces an **estimated standard error**, denoted by $\hat{\sigma}_{\hat{\Theta}}$.

Sometimes the estimated standard error is denoted by $s_{\hat{\Theta}}$ or $se(\hat{\Theta})$.

MEAN SQUARED ERROR OF AN ESTIMATOR

Sometimes it is necessary to use a biased estimator. In such cases, the mean squared error of the estimator can be important. The **mean squared error** of an estimator $\widehat{\Theta}$ is the expected squared difference between $\widehat{\Theta}$ and θ .

The mean squared error of an estimator $\widehat{\Theta}$ of the parameter θ is defined as

$$MSE(\widehat{\Theta}) = E(\widehat{\Theta} - \theta)^2$$

The mean squared error can be rewritten as follows:

$$\begin{aligned} MSE(\widehat{\Theta}) &= E[\widehat{\Theta} - E(\widehat{\Theta})]^2 + [\theta - E(\widehat{\Theta})]^2 \\ &= V(\widehat{\Theta}) + (bias)^2 \end{aligned}$$

That is, the mean squared error of $\widehat{\Theta}$ is equal to the variance of the estimator plus the squared bias. If $\widehat{\Theta}$ is an unbiased estimator of θ , the mean squared error of $\widehat{\Theta}$ is equal to the variance of $\widehat{\Theta}$.

MEAN SQUARED ERROR OF AN ESTIMATOR

The mean squared error is an important criterion for comparing two estimators. Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two estimators of the parameter θ , and let $MSE(\widehat{\theta}_1)$ and $MSE(\widehat{\theta}_2)$ be the mean squared errors of $\widehat{\theta}_1$ and $\widehat{\theta}_2$. Then the **relative efficiency** of $\widehat{\theta}_2$ to $\widehat{\theta}_1$ is defined as

$$\frac{MSE(\widehat{\theta}_1)}{MSE(\widehat{\theta}_2)}$$

If this relative efficiency is less than 1, we would conclude that $\widehat{\theta}_1$ is a more efficient estimator of θ than $\widehat{\theta}_2$, in the sense that it has a smaller mean squared error.

METHODS OF POINT ESTIMATION

The definitions of unbiasedness and other properties of estimators do not provide any guidance about how good estimators can be obtained.

Methods for obtaining point estimators:

- The method of moments
- The method of maximum likelihood.

METHOD OF MOMENTS

The general idea behind the method of moments is to equate **population moments**, which are defined in terms of expected values, to the corresponding **sample moments**. The population moments will be functions of the unknown parameters. Then these equations are solved to yield estimators of the unknown parameters.

Definition 12.4: Moments

Let X_1, X_2, \dots, X_n be a random sample from the probability distribution $f(x)$, where $f(x)$ can be a discrete probability mass function or a continuous probability density function. The k th **population moment (or distribution moment)** is $E(X^k)$, $k = 1, 2, \dots$. The corresponding k th sample moment is $\left(\frac{1}{n}\right) \sum_{i=1}^n X_i^k$, $k = 1, 2, \dots$.

METHOD OF MOMENTS

To illustrate, the first population moment is $E(X) = \mu$, and the first sample moment is $\left(\frac{1}{n}\right)\sum_{i=1}^n X_i = \bar{X}$. Thus by equating the population and sample moments, we find that $\hat{\mu} = \bar{X}$. That is, the sample mean is the **moment estimator** of the population mean. In the general case, the population moments will be functions of the unknown parameters of the distribution, say, $\theta_1, \theta_2, \dots, \theta_m$.

Definition 12.5: Moment Estimators

Let X_1, X_2, \dots, X_n be a random sample from either a probability mass function or a probability density function with m unknown parameters $\theta_1, \theta_2, \dots, \theta_m$. The **moment estimators** $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ are found by equating the first m population moments to the first m sample moments and solving the resulting equations for the unknown parameters.

METHOD OF MAXIMUM LIKELIHOOD

One of the best methods of obtaining a point estimator of a parameter is the method of maximum likelihood. This technique was developed in the 1920s by a famous British statistician, Sir R. A. Fisher. As the name implies, the estimator will be the value of the parameter that maximizes the likelihood function.

Definition 12.6: Maximum Likelihood Estimator

Suppose that X is a random variable with probability distribution $f(x; \theta)$, where θ is a single unknown parameter. Let x_1, x_2, \dots, x_n be the observed values in a random sample of size n . Then the **likelihood function** of the sample is

$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdots \cdot f(x_n; \theta)$$

Note that the likelihood function is now a function of only the unknown parameter θ . The **maximum likelihood estimator (MLE)** of θ is the value of θ that maximizes the likelihood function $L(\theta)$.

METHOD OF MAXIMUM LIKELIHOOD

In the case of a discrete random variable, the interpretation of the likelihood function is simple. The likelihood function of the sample $L(\theta)$ is just the probability

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

That is, $L(\theta)$ is just the probability of obtaining the sample values x_1, x_2, \dots, x_n . Therefore, in the discrete case, the maximum likelihood estimator is an estimator that maximizes the probability of occurrence of the sample values.

PROPERTIES OF THE MAXIMUM LIKELIHOOD ESTIMATOR

The method of maximum likelihood is often the estimation method that mathematical statisticians prefer, because it produces estimators with good statistical properties.

Under very general and not restrictive conditions, when the sample size n is large and if $\hat{\theta}$ is the maximum likelihood estimator of the parameter θ ,

1. $\hat{\theta}$ is an approximately unbiased estimator for θ , $[E(\hat{\theta}) \approx \theta]$
2. the variance of $\hat{\theta}$ is nearly as small as the variance that could be obtained with any other estimator, and
3. $\hat{\theta}$ has an approximate normal distribution.

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 13

Non-parametric tests

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

QUOTE OF THE DAY..

Try not to become a person of success, but rather try to become a person of value.

ALBERT EINSTEIN,Theoretical physicist

INTRODUCTION

- All of the tests presented in hypothesis testing are called **parametric tests** and are based on certain assumptions.
- For example, when running tests of hypothesis for means of continuous outcomes, all parametric tests assume that the outcome is approximately normally distributed in the population. This does not mean that the data in the observed sample follows a normal distribution, but rather that the outcome follows a normal distribution in the full population which is not observed.
- Many statistical tests are **robust**, which means that they maintain their statistical properties even when assumptions are not entirely met. Tests are robust in the presence of violations of the normality assumption when the sample size is large based on the Central Limit Theorem.
- When the sample size is small and the distribution of the outcome is not known and cannot be assumed to be approximately normally distributed, then alternative tests called **nonparametric tests** are appropriate.

WHEN TO USE A NONPARAMETRIC TEST

- Nonparametric tests are called **distribution-free tests** because they are based on fewer assumptions (e.g., they do not assume that the outcome is approximately normally distributed).
- Parametric tests involve specific probability distributions (e.g., the normal distribution) and the tests involve estimation of the key parameters of that distribution (e.g., the mean or difference in means) from the sample data. The cost of fewer assumptions is that nonparametric tests are generally less powerful than their parametric counterparts (i.e., when the alternative is true, they may be less likely to reject H_0).
- It can sometimes be difficult to assess whether a continuous outcome follows a normal distribution and, thus, whether a parametric or nonparametric test is appropriate.
- There are several statistical tests that can be used to assess whether data are likely from a normal distribution. The most popular are **the Anderson-Darling test, and the Shapiro-Wilk test**.

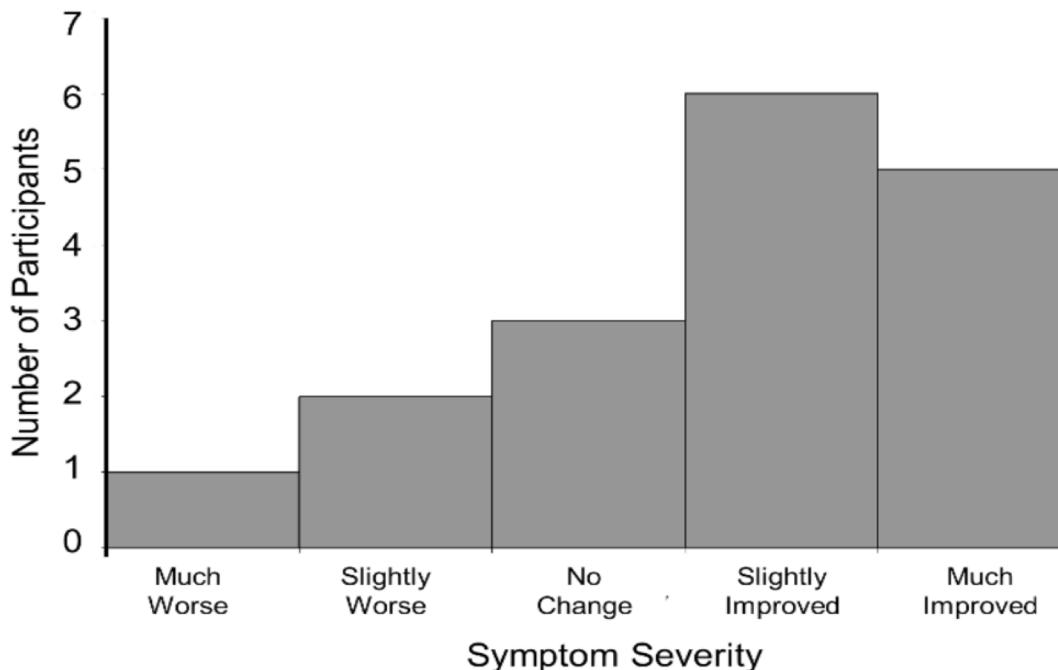
WHEN TO USE A NONPARAMETRIC TEST

- There are some situations when it is clear that the outcome does not follow a normal distribution. These include situations:
 - when the outcome is an **ordinal variable or a rank**,
 - when there are **definite outliers** or
 - when the outcome has **clear limits of detection**.

USING AN ORDINAL SCALE

Consider a clinical trial where study participants are asked to rate their symptom severity following 6 weeks on the assigned treatment. Symptom severity might be measured on a 5 point ordinal scale with response options: Symptoms got much worse, slightly worse, no change, slightly improved, or much improved. Suppose there are a total of n=20 participants in the trial, randomized to an experimental treatment or placebo, and the outcome data are distributed as shown in the figure below.

Distribution of Symptom Severity in Total Sample



The distribution of the outcome (symptom severity) does not appear to be normal as more participants report improvement in symptoms as opposed to worsening of symptoms.

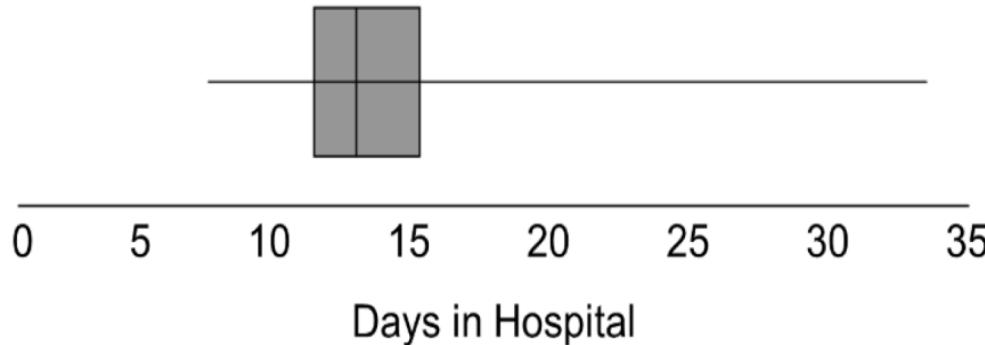
WHEN THE OUTCOME IS A RANK

In some studies, the outcome is a rank. For example, in new-born health studies an APGAR score is often used to assess the health of a new-born. The score, which ranges from 1-10, is the sum of five component scores based on the infant's condition at birth. APGAR scores generally do not follow a normal distribution, since most new-borns have scores of 7 or higher (normal range).

WHEN THERE ARE OUTLIERS

In some studies, the outcome is continuous but subject to outliers or extreme values. For example, days in the hospital following a particular surgical procedure is an outcome that is often subject to outliers. Suppose in an observational study investigators wish to assess whether there is a difference in the days patients spend in the hospital following liver transplant in for-profit versus nonprofit hospitals. Suppose we measure days in the hospital following transplant in n=100 participants, 50 from for-profit and 50 from non-profit hospitals. The number of days in the hospital are summarized by the box-whisker plot below.

Distribution of Days in the Hospital Following Transplant



- $Q_1 - 1.5(Q_3 - Q_1)$ as a lower limit and $Q_3 + 1.5(Q_3 - Q_1)$ as an upper limit to detect outliers.
- In the box-whisker plot above, $Q_1 = 12$ and $Q_3 = 16$, thus outliers are values below $12 - 1.5(16 - 12) = 6$ or above $16 + 1.5(16 - 12) = 22$.

Note that 75% of the participants stay at most 16 days in the hospital following transplant, while at least 1 stays 35 days which would be considered an outlier.

LIMITS OF DETECTION

In some studies, the outcome is a continuous variable that is measured with some imprecision (e.g., with clear limits of detection). For example, some instruments or evaluation cannot measure presence of specific quantities above or below certain limits.

HIV viral load is a measure of the amount of virus in the body and is measured as the amount of virus per a certain volume of blood. It can range from "not detected" or "below the limit of detection" to hundreds of millions of copies. Thus, in a sample some participants may have measures like 1,254,000 or 874,050 copies and others are measured as "not detected." If a substantial number of participants have undetectable levels, the distribution of viral load is not normally distributed.

ADVANTAGES OF NONPARAMETRIC TESTS

Nonparametric tests have some distinct advantages. With outcomes such as those described before, nonparametric tests may be the only way to analyse these data. Outcomes that are ordinal, ranked, subject to outliers or measured imprecisely are difficult to analyse with parametric methods without making major assumptions about their distributions as well as decisions about coding some values (e.g., "not detected"). Nonparametric tests can be relatively simple to conduct.

Hypothesis Testing with Nonparametric Tests

In nonparametric tests, the hypotheses are not about population parameters (e.g., $\mu=50$ or $\mu_1=\mu_2$). Instead, the null hypothesis is more general. For example, when comparing two independent groups in terms of a continuous outcome, the null hypothesis in a parametric test is $H_0: \mu_1 = \mu_2$. In a nonparametric test the null hypothesis is that the two populations are equal, often this is interpreted as the two populations are **equal in terms of their central tendency**.

NONPARAMETRIC TESTING

Assigning Ranks

- The outcome variable (ordinal, interval or continuous) is ranked from lowest to highest and the analysis focuses on the ranks as opposed to the measured or raw values. For example, suppose we measure self-reported pain using a visual analog scale with anchors at 0 (no pain) and 10 (agonizing pain) and record the following in a sample of n=6 participants:

7 5 9 3 0 2

- The ranks, which are used to perform a nonparametric test, are assigned as follows: First, the data are ordered from smallest to largest. The lowest value is then assigned a rank of 1, the next lowest a rank of 2 and so on. The largest value is assigned a rank of n (in this example, n=6). The observed data and corresponding ranks are shown below:

Ordered Observed Data:

| | | | | | |
|---|---|---|---|---|---|
| 0 | 2 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|

Ranks:

| | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|

NONPARAMETRIC TESTING

- A complicating issue that arises when assigning ranks occurs when there are ties in the sample (i.e., the same values are measured in two or more participants). For example, suppose that the following data are observed in our sample of $n=6$:

Observed Data: 7 7 9 3 0 2

- The 4th and 5th ordered values are both equal to 7. When assigning ranks, the recommended procedure is to assign the mean rank of 4.5 to each (i.e. the mean of 4 and 5), as follows:

| | | | | | | |
|-------------------------------|---|---|---|-----|-----|---|
| Ordered Observed Data: | 0 | 2 | 3 | 7 | 7 | 9 |
| Ranks: | 1 | 2 | 3 | 4.5 | 4.5 | 6 |

- Suppose that there are three values of 7. In this case, we assign a rank of 5 (the mean of 4, 5 and 6) to the 4th, 5th and 6th values, as follows:

| | | | | | | |
|-------------------------------|---|---|---|---|---|---|
| Ordered Observed Data: | 0 | 2 | 3 | 7 | 7 | 7 |
| Ranks: | 1 | 2 | 3 | 5 | 5 | 5 |

NONPARAMETRIC TESTING

Note:

Using this approach of assigning the mean rank when there are ties ensures that the sum of the ranks is the same in each sample (for example, $1+2+3+4+5+6=21$, $1+2+3+4.5+4.5+6=21$ and $1+2+3+5+5+5=21$). Using this approach, the sum of the ranks will always equal $n(n+1)/2$. When conducting nonparametric tests, it is useful to check the sum of the ranks before proceeding with the analysis.

NONPARAMETRIC TESTING

To conduct nonparametric tests, we again follow the five-step approach outlined in the hypothesis testing.

- Set up hypotheses and select the level of significance α . Analogous to parametric testing, the research hypothesis can be one- or two- sided (one- or two-tailed), depending on the research question of interest.
- Select the appropriate test statistic. The test statistic is a single number that summarizes the sample information. In nonparametric tests, the observed data is converted into ranks and then the ranks are summarized into a test statistic.
- Set up decision rule. The decision rule is a statement that tells under what circumstances to reject the null hypothesis. Note that in some nonparametric tests we reject H_0 if the test statistic is large, while in others we reject H_0 if the test statistic is small. We make the distinction as we describe the different tests.
- Compute the test statistic. Here we compute the test statistic by summarizing the ranks into the test statistic identified in Step 2.
- Conclusion. The final conclusion is made by comparing the test statistic (which is a summary of the information observed in the sample) to the decision rule. The final conclusion is either to reject the null hypothesis (because it is very unlikely to observe the sample data if the null hypothesis is true) or not to reject the null hypothesis (because the sample data are not very unlikely if the null hypothesis is true).

MANN WHITNEY U TEST (WILCOXON RANK SUM TEST)

- A popular nonparametric test to compare outcomes between two independent groups is the Mann Whitney U test.
- The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape). Some investigators interpret this test as comparing the medians between the two populations.
- In contrast, the null and two-sided research hypotheses for the nonparametric test are stated as follows:

H₀: The two populations are equal

H₁: The two populations are not equal.

- This test is often performed as a two-sided test and, thus, the research hypothesis indicates that the populations are not equal as opposed to specifying directionality. A one-sided research hypothesis is used if interest lies in detecting a positive or negative shift in one population as compared to the other. The procedure for the test involves pooling the observations from the two samples into one combined sample, keeping track of which sample each observation comes from, and then ranking lowest to highest from 1 to n₁+n₂, respectively.

MANN WHITNEY U TEST (WILCOXON RANK SUM TEST)

- **Example:**
- Consider a Phase II clinical trial designed to investigate the effectiveness of a new drug to reduce symptoms of asthma in children. A total of n=10 participants are randomized to receive either the new drug or a placebo. Participants are asked to record the number of episodes of shortness of breath over a 1 week period following receipt of the assigned treatment. The data are shown below.

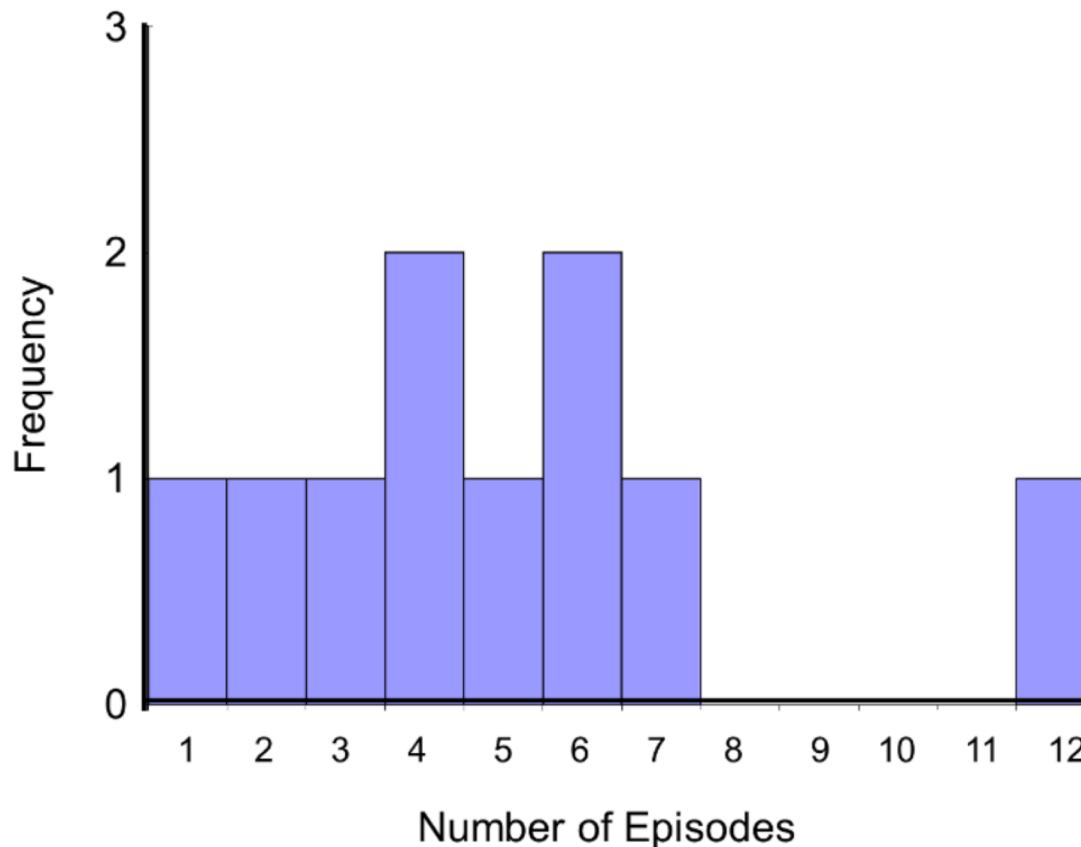
| | | | | | |
|----------|---|---|---|---|----|
| Placebo | 7 | 5 | 6 | 4 | 12 |
| New Drug | 3 | 6 | 4 | 2 | 1 |

Is there a difference in the number of episodes of shortness of breath over a 1 week period in participants receiving the new drug as compared to those receiving the placebo? By inspection, it appears that participants receiving the placebo have more episodes of shortness of breath, but is this statistically significant?

EXAMPLE 1

- In this example, the outcome is a count and in this sample the data do not follow a normal distribution.

Frequency Histogram of Number of Episodes of Shortness of Breath



EXAMPLE 1

- In addition, the sample size is small ($n_1=n_2=5$), so a nonparametric test is appropriate. The hypothesis is given below, and we run the test at the 5% level of significance (i.e., $\alpha=0.05$).

H_0 : The two populations are equal

H_1 : The two populations are not equal.

- Note that if the null hypothesis is true (i.e., the two populations are equal), we expect to see similar numbers of episodes of shortness of breath in each of the two treatment groups, and we would expect to see some participants reporting few episodes and some reporting more episodes in each group. This does not appear to be the case with the observed data. A test of hypothesis is needed to determine whether the observed data is evidence of a statistically significant difference in populations.
- The first step is to assign ranks and to do so we order the data from smallest to largest. This is done on the combined or total sample (i.e., pooling the data from the two treatment groups ($n=10$)), and assigning ranks from 1 to 10, as follows. We also need to keep track of the group assignments in the total sample.

EXAMPLE 1

| | | Total Sample (Ordered Smallest to Largest) | | | | Ranks | |
|---------|----------|---|----------|---------|----------|-------|-----|
| Placebo | New Drug | Placebo | New Drug | Placebo | New Drug | | |
| 7 | 3 | | | 1 | | | 1 |
| 5 | 6 | | | 2 | | | 2 |
| 6 | 4 | | | 3 | | | 3 |
| 4 | 2 | 4 | | 4 | | 4.5 | 4.5 |
| 12 | 1 | 5 | | | | 6 | |
| | | 6 | | 6 | | 7.5 | 7.5 |
| | | 7 | | | | 9 | |
| | | 12 | | | | 10 | |

The goal of the test is to determine whether the observed data support a difference in the populations of responses. First, we sum the ranks in each group. In the placebo group, the sum of the ranks is 37; in the new drug group, the sum of the ranks is 18. As a check on our assignment of ranks, we have $n(n+1)/2 = 10(11)/2=55$ which is equal to $37+18 = 55$.

For the test, we call the placebo group 1 and the new drug group 2. $R_1=37$ and $R_2=18$. If the null hypothesis is true (i.e., if the two populations are equal), we expect R_1 and R_2 to be similar. In this example, the lower values (lower ranks) are clustered in the new drug group (group 2), while the higher values (higher ranks) are clustered in the placebo group (group 1). This is suggestive, but is the observed difference in the sums of the ranks simply due to chance? To answer this we will compute a test statistic to summarize the sample information and look up the corresponding value in a probability distribution.

EXAMPLE 1

Test Statistic for the Mann Whitney U Test

The test statistic for the Mann Whitney U Test is denoted **U** and is the *smaller* of U_1 and U_2

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where R_1 = sum of the ranks for group 1 and R_2 = sum of the ranks for group 2.

$$U_1 = 5(5) + \frac{5(6)}{2} - 37 = 3$$

$$U_2 = 5(5) + \frac{5(6)}{2} - 18 = 22$$

U=3. Is this evidence in support of the null or research hypothesis?

Smaller values of U support the research hypothesis, and larger values of U support the null hypothesis.

Key Concept:

- For any Mann-Whitney U test, the theoretical range of U is from 0 (complete separation between groups, H_0 most likely false and H_1 most likely true) to $n_1 * n_2$ (little evidence in support of H_1).
- In every test, $U_1 + U_2$ is always equal to $n_1 * n_2$. In the example above, U can range from 0 to 25 and smaller values of U support the research hypothesis (i.e., we reject H_0 if U is small).

EXAMPLE 1

- In every test, we must determine whether the observed U supports the null or research hypothesis. Specifically, we determine a critical value of U such that **if the observed value of U is less than or equal to the critical value, we reject H_0 in favor of H_1 and if the observed value of U exceeds the critical value we do not reject H_0 .**
- The critical value of U can be found from the table. To determine the appropriate critical value we need sample sizes (for Example: $n_1=n_2=5$) and our two-sided level of significance ($\alpha=0.05$). For the above Example, the critical value is 2, and the decision rule is to reject H_0 if $U \leq 2$. **We do not reject H_0 because $3 > 2$.** We do not have statistically significant evidence at $\alpha =0.05$, to show that the two populations of numbers of episodes of shortness of breath are not equal. However, in this example, the failure to reach statistical significance may be due to low power. **The sample data suggest a difference, but the sample sizes are too small to conclude that there is a statistically significant difference.**

EXAMPLE 2

- A new approach to prenatal care is proposed for pregnant women living in a rural community. The new program involves in-home visits during the course of pregnancy in addition to the usual or regularly scheduled visits. A pilot randomized trial with 15 pregnant women is designed to evaluate whether women who participate in the program deliver healthier babies than women receiving usual care. The outcome is the APGAR score measured 5 minutes after birth. APGAR scores range from 0 to 10 with scores of 7 or higher considered normal (healthy), 4-6 low and 0-3 critically low. The data are shown below.

| | | | | | | | | |
|-------------|---|---|---|---|----|---|---|---|
| Usual Care | 8 | 7 | 6 | 2 | 5 | 8 | 7 | 3 |
| New Program | 9 | 9 | 7 | 8 | 10 | 9 | 6 | |

Is there statistical evidence of a difference in APGAR scores in women receiving the new and enhanced versus usual prenatal care?

EXAMPLE 2

- **Step 1.** Set up hypotheses and determine level of significance.

H_0 : The two populations are equal

H_1 : The two populations are not equal. $\alpha = 0.05$

- **Step 2.** Select the appropriate test statistic.

- Because APGAR scores are not normally distributed and the samples are small ($n_1=8$ and $n_2=7$), we can use the Mann Whitney U test.
- The test statistic is U, the smaller of U_1 and U_2

- **Step 3.** Set up decision rule.

- The appropriate critical value can be found from the table. To determine the appropriate critical value we need sample sizes ($n_1=8$ and $n_2=7$) and our two-sided level of significance ($\alpha=0.05$).
- The critical value for this test with $n_1=8$, $n_2=7$ and $\alpha = 0.05$ is 10 and the decision rule is as follows: **Reject H_0 if $U \leq 10$.**

EXAMPLE 2

- **Step 4.** Compute the test statistic.
- The first step is to assign ranks of 1 through 15 to the smallest through largest values in the total sample, as follows:

| | | Total Sample (Ordered Smallest to Largest) | | Ranks | |
|------------|-------------|---|-------------|----------------------|----------------------|
| Usual Care | New Program | Usual Care | New Program | Usual Care | New Program |
| 8 | 9 | 2 | | 1 | |
| 7 | 8 | 3 | | 2 | |
| 6 | 7 | 5 | | 3 | |
| 2 | 8 | 6 | 6 | 4.5 | 4.5 |
| 5 | 10 | 7 | 7 | 7 | 7 |
| 8 | 9 | 7 | | 7 | |
| 7 | 6 | 8 | 8 | 10.5 | 10.5 |
| 3 | | 8 | 8 | 10.5 | 10.5 |
| | | | 9 | | 13.5 |
| | | | 9 | | 13.5 |
| | | | 10 | | 15 |
| | | | | R ₁ =45.5 | R ₂ =74.5 |

EXAMPLE 2

- Next, we sum the ranks in each group. In the usual care group, the sum of the ranks is $R_1=45.5$ and in the new program group, the sum of the ranks is $R_2=74.5$. Recall that the sum of the ranks will always equal $n(n+1)/2$. As a check on our assignment of ranks, we have $n(n+1)/2 = 15(16)/2=120$ which is equal to $45.5+74.5 = 120$.
- We now compute U_1 and U_2 , as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 8(7) + \frac{8(9)}{2} - 45.5 = 46.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 8(7) + \frac{7(8)}{2} - 74.5 = 9.5$$

Thus, the test statistic is $U=9.5$.

•**Step 5.** Conclusion:

We reject H_0 because $9.5 \leq 10$. We have statistically significant evidence at $\alpha =0.05$ to show that the populations of APGAR scores are not equal in women receiving usual prenatal care as compared to the new program of prenatal care.

Any question?

REFERENCE

- The detail material related to this lecture can be found in
 - D'Agostino RB and Stevens MA. Goodness of Fit Techniques.
 - Apgar, Virginia (1953). "A proposal for a new method of evaluation of the newborn infant ". Curr. Res. Anesth. Analg. 32 (4): 260-267.
 - Conover WJ. Practical Nonparametric Statistics, 2nd edition, New York: John Wiley and Sons.
 - Siegel and Castellan. (1988). "Nonparametric Statistics for the Behavioral Sciences," 2nd edition, New York: McGraw-Hill.



INTRODUCTION TO DATA ANALYTICS

Class # 14

Non-parametric tests

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

QUOTE OF THE DAY..

Try not to become a person of success, but rather try to become a person of value.

ALBERT EINSTEIN,Theoretical physicist

TESTS WITH MATCHED SAMPLES

- This section describes nonparametric tests to compare two groups with respect to a continuous outcome when the data are collected on matched or paired samples.
- This section describes procedures that should be used when the outcome cannot be assumed to follow a normal distribution. There are two popular nonparametric tests to compare outcomes between two matched or paired groups. The first is called the **Sign Test** and the second the **Wilcoxon Signed Rank Test**.
- When data are matched or paired, we compute difference scores for each individual and analyze difference scores. The same approach is followed in nonparametric tests. In parametric tests, the null hypothesis is that the mean difference (μ_d) is zero. In nonparametric tests, the null hypothesis is that the median difference is zero.

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

- The two comparison groups are said to be **dependent**, and the data can arise from a single sample of participants where each participant is measured twice (possibly before and after an intervention) or from two samples that are matched on specific characteristics (e.g., siblings).
- When the samples are dependent, we focus on **difference scores** in each participant or between members of a pair and the test of hypothesis is based on the mean difference, μ_d . The null hypothesis again reflects "no difference" and is stated as $H_0: \mu_d = 0$.
- Note that there are some instances where it is of interest to test whether there is a difference of a particular magnitude (e.g., $\mu_d = 5$) but in most instances the null hypothesis reflects no difference (i.e., $\mu_d = 0$).

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

Example:

- A new drug is proposed to lower total cholesterol and a study is designed to evaluate the efficacy of the drug in lowering cholesterol. Fifteen patients agree to participate in the study and each is asked to take the new drug for 6 weeks. However, before starting the treatment, each patient's total cholesterol level is measured. The initial measurement is a pre-treatment or baseline value. After taking the drug for 6 weeks, each patient's total cholesterol level is measured again and the data are shown below. The rightmost column contains difference scores for each patient, computed by subtracting the 6 week cholesterol level from the baseline level. The differences represent the reduction in total cholesterol over 4 weeks. (The differences could have been computed by subtracting the baseline total cholesterol level from the level measured at 6 weeks. The way in which the differences are computed does not affect the outcome of the analysis only the interpretation.)

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

| Subject Identification Number | Baseline | 6 Weeks | Difference |
|-------------------------------|----------|---------|------------|
| 1 | 215 | 205 | 10 |
| 2 | 190 | 156 | 34 |
| 3 | 230 | 190 | 40 |
| 4 | 220 | 180 | 40 |
| 5 | 214 | 201 | 13 |
| 6 | 240 | 227 | 13 |
| 7 | 210 | 197 | 13 |
| 8 | 193 | 173 | 20 |
| 9 | 210 | 204 | 6 |
| 10 | 230 | 217 | 13 |
| 11 | 180 | 142 | 38 |
| 12 | 260 | 262 | -2 |
| 13 | 210 | 207 | 3 |
| 14 | 190 | 184 | 6 |
| 15 | 200 | 193 | 7 |

Because the differences are computed by subtracting the cholesterols measured at 6 weeks from the baseline values, positive differences indicate reductions and negative differences indicate increases (e.g., participant 12 increases by 2 units over 6 weeks). The goal here is to test whether there is a statistically significant reduction in cholesterol.

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

In order to conduct the test, we need to summarize the differences. In this sample, we have

| Subject Identification Number | Difference | Difference ² |
|-------------------------------|------------|-------------------------|
| 1 | 10 | 100 |
| 2 | 34 | 1156 |
| 3 | 40 | 1600 |
| 4 | 40 | 1600 |
| 5 | 13 | 169 |
| 6 | 13 | 169 |
| 7 | 13 | 169 |
| 8 | 20 | 400 |
| 9 | 6 | 36 |
| 10 | 13 | 169 |
| 11 | 38 | 1444 |
| 12 | -2 | 4 |
| 13 | 3 | 9 |
| 14 | 6 | 36 |
| 15 | 7 | 49 |
| Totals | 254 | 7110 |

$$s_d = \sqrt{\frac{\sum \text{Differences}^2 - (\sum \text{Differences})^2 / n}{n-1}}$$

$$s_d = \sqrt{\frac{7110 - (254)^2 / 15}{14}} = \sqrt{\frac{2808.93}{14}} = \sqrt{200.64} = 14.2$$

$$n = 15, \\ \bar{x}_d = 16.9, \\ S_d = 14.2$$

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

Is there statistical evidence of a reduction in mean total cholesterol in patients after using the new medication for 6 weeks? We will run the test using the five-step approach.

- **Step 1.** Set up hypotheses and determine level of significance

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0 \quad \alpha=0.05$$

NOTE: If we had computed differences by subtracting the baseline level from the level measured at 6 weeks then negative differences would have reflected reductions and the research hypothesis would have been $H_1: \mu_d < 0$.

- **Step 2.** Select the appropriate test statistic.

Because the sample size is small ($n<30$) the appropriate test statistic is

$$t = \frac{\bar{X}_d - \mu_d}{s_d / \sqrt{n}}$$

- **Step 3.** Set up decision rule.

This is an one-tailed test, using a t statistic and a 5% level of significance. The appropriate critical value can be found in the t Table at the right, with $df=15-1=14$. The critical value for an upper-tailed test with $df=14$ and $\alpha=0.05$ is 2.145 and the decision rule is Reject H_0 if $t \geq 2.145$.

TESTS WITH MATCHED SAMPLES – PARAMETRIC TESTS

- **Step 4.** Compute the test statistic.
- We now substitute the sample data into the formula for the test statistic identified in Step 2.

$$t = \frac{\bar{X}_d - \mu_d}{s_d / \sqrt{n}} = \frac{16.9 - 0}{14.2 / \sqrt{15}} = 4.61$$

- **Step 5.** Conclusion.
- We reject H_0 because $4.61 \geq 2.145$. We have statistically significant evidence at $\alpha=0.05$ to show that there is a reduction in cholesterol levels over 6 weeks.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

Example:

- Consider a clinical investigation to assess the effectiveness of a new drug designed to reduce repetitive behaviors in children affected with autism. If the drug is effective, children will exhibit fewer repetitive behaviors on treatment as compared to when they are untreated. A total of 8 children with autism enroll in the study. Each child is observed by the study psychologist for a period of 3 hours both before treatment and then again after taking the new drug for 1 week. The time that each child is engaged in repetitive behavior during each 3 hour observation period is measured. Repetitive behavior is scored on a scale of 0 to 100 and scores represent the percent of the observation time in which the child is engaged in repetitive behavior. For example, a score of 0 indicates that during the entire observation period the child did not engage in repetitive behavior while a score of 100 indicates that the child was constantly engaged in repetitive behavior. The data are shown below.

| Child | Before Treatment | After 1 Week of Treatment |
|-------|------------------|---------------------------|
| 1 | 85 | 75 |
| 2 | 70 | 50 |
| 3 | 40 | 50 |
| 4 | 65 | 40 |
| 5 | 80 | 20 |
| 6 | 75 | 65 |
| 7 | 55 | 40 |
| 8 | 20 | 25 |

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

- Looking at the data, it appears that some children improve (e.g., Child 5 scored 80 before treatment and 20 after treatment), but some got worse (e.g., Child 3 scored 40 before treatment and 50 after treatment). Is there statistically significant improvement in repetitive behavior after 1 week of treatment?.
- Because the before and after treatment measures are paired, we compute difference scores for each child. In this example, we subtract the assessment of repetitive behaviors after treatment from that measured before treatment so that difference scores represent improvement in repetitive behavior. The question of interest is whether there is significant improvement after treatment.

| Child | Before Treatment | After 1 Week of Treatment | Difference (Before-After) |
|-------|------------------|---------------------------|------------------------------|
| 1 | 85 | 75 | 10 |
| 2 | 70 | 50 | 20 |
| 3 | 40 | 50 | -10 |
| 4 | 65 | 40 | 25 |
| 5 | 80 | 20 | 60 |
| 6 | 75 | 65 | 10 |
| 7 | 55 | 40 | 15 |
| 8 | 20 | 25 | -5 |

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

- In this small sample, the observed difference (or improvement) scores vary widely and are subject to extremes (e.g., the observed difference of 60 is an outlier). Thus, a nonparametric test is appropriate to test whether there is significant improvement in repetitive behavior before versus after treatment. The hypotheses are given below.

H_0 : The median difference is zero

H_1 : The median difference is positive

- In this example, the null hypothesis is that there is no difference in scores before versus after treatment. If the null hypothesis is true, we expect to see some positive differences (improvement) and some negative differences (worsening). If the research hypothesis is true, we expect to see more positive differences after treatment as compared to before.

How to solve this: Sign test and Wilcoxon signed Rank test

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

SIGN TEST:

- The Sign Test is the simplest nonparametric test for matched or paired data. The approach is to analyze only the signs of the difference scores, as shown below:

| Child | Before Treatment | After 1 Week of Treatment | Difference (Before-After) | Sign |
|-------|------------------|---------------------------|---------------------------|------|
| 1 | 85 | 75 | 10 | + |
| 2 | 70 | 50 | 20 | + |
| 3 | 40 | 50 | -10 | - |
| 4 | 65 | 40 | 25 | + |
| 5 | 80 | 20 | 60 | + |
| 6 | 75 | 65 | 10 | + |
| 7 | 55 | 40 | 15 | + |
| 8 | 20 | 25 | -5 | - |

If the null hypothesis is true (i.e., if the median difference is zero) then we expect to see approximately half of the differences as positive and half of the differences as negative. If the research hypothesis is true, we expect to see more positive differences.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

Test Statistic for the Sign Test

- The test statistic for the Sign Test is the number of positive signs or number of negative signs, whichever is smaller. In this example, we observe 2 negative and 6 positive signs. Is this evidence of significant improvement or simply due to chance?
- Determining whether the observed test statistic supports the null or research hypothesis is done following the same approach used in parametric testing. Specifically, we determine a critical value such that if the smaller of the number of positive or negative signs is less than or equal to that critical value, then we reject H_0 in favor of H_1 and if the smaller of the number of positive or negative signs is greater than the critical value, then we do not reject H_0 . Notice that this is a one-sided decision rule corresponding to our one-sided research hypothesis.

Critical Values for the Sign Test

- To determine the appropriate critical value we need the sample size, which is equal to the number of matched pairs ($n=8$) and our one-sided level of significance $\alpha=0.05$. For this example, the critical value is 5, and the decision rule is to reject H_0 if the smaller of the number of positive or negative signs < 5 .

Conclusion:

- **We reject H_0 because $2 < 5$.** We have sufficient evidence at $\alpha=0.05$ to show that there is improvement in repetitive behavior after taking the drug as compared to before. In essence, we could use the critical value to decide whether to reject the null hypothesis.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

When Difference Scores are Zero

- There is a special circumstance that needs attention when implementing the Sign Test which arises when one or more participants have difference scores of zero (i.e., their paired measurements are identical). If there is just one difference score of zero, some investigators drop that observation and reduce the sample size by 1 (i.e., the sample size would be $n-1$). This is a reasonable approach if there is just one zero. However, if there are two or more zeros, an alternative approach is preferred.
- If there is an even number of zeros, we randomly assign them positive or negative signs.
- If there is an odd number of zeros, we randomly drop one and reduce the sample size by 1, and then randomly assign the remaining observations positive or negative signs.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

Example:

- A new chemotherapy treatment is proposed for patients with breast cancer. Investigators are concerned with patient's ability to tolerate the treatment and assess their quality of life both before and after receiving the new chemotherapy treatment. Quality of life (QOL) is measured on an ordinal scale and for analysis purposes, numbers are assigned to each response category as follows: 1=Poor, 2= Fair, 3=Good, 4= Very Good, 5 = Excellent. The data are shown below.

| Patient | QOL Before | | QOL After | |
|---------|------------------------|------------------------|------------------------|------------------------|
| | Chemotherapy Treatment | Chemotherapy Treatment | Chemotherapy Treatment | Chemotherapy Treatment |
| 1 | | 3 | | 2 |
| 2 | | 2 | | 3 |
| 3 | | 3 | | 4 |
| 4 | | 2 | | 4 |
| 5 | | 1 | | 1 |
| 6 | | 3 | | 4 |
| 7 | | 2 | | 4 |
| 8 | | 3 | | 3 |
| 9 | | 2 | | 1 |
| 10 | | 1 | | 3 |
| 11 | | 3 | | 4 |
| 12 | | 2 | | 3 |

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

The question of interest is whether there is a difference in QOL after chemotherapy treatment as compared to before.

- **Step 1.** Set up hypotheses and determine level of significance.

H_0 : The median difference is zero

H_1 : The median difference is not zero ($\alpha=0.05$)

- **Step 2.** Select the appropriate test statistic.

The test statistic for the Sign Test is the smaller of the number of positive or negative signs.

- **Step 3.** Set up the decision rule.

The appropriate critical value for the Sign Test can be found in the table of critical values for the Sign Test. To determine the appropriate critical value we need the sample size (or number of matched pairs, $n=12$), and our two-sided level of significance $\alpha=0.05$.

The critical value for this two-sided test with $n=12$ and $\alpha=0.05$ is 13, and the decision rule is as follows: Reject H_0 if the smaller of the number of positive or negative signs ≤ 13 .

- **Step 4.** Compute the test statistic.

Because the before and after treatment measures are paired, we compute difference scores for each patient. In this example, we subtract the QOL measured before treatment from that measured after.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TESTS

We now capture the signs of the difference scores and because there are two zeros, we randomly assign one negative sign (i.e., "-" to patient 5) and one positive sign (i.e., "+" to patient 8), as follows:

| Patient | QOL Before Chemotherapy Treatment | QOL After Chemotherapy Treatment | Difference (After-Before) | Sign |
|---------|-----------------------------------|----------------------------------|---------------------------|------|
| 1 | 3 | 2 | -1 | - |
| 2 | 2 | 3 | 1 | + |
| 3 | 3 | 4 | 1 | + |
| 4 | 2 | 4 | 2 | + |
| 5 | 1 | 1 | 0 | - |
| 6 | 3 | 4 | 1 | + |
| 7 | 2 | 4 | 2 | + |
| 8 | 3 | 3 | 0 | + |
| 9 | 2 | 1 | -1 | - |
| 10 | 1 | 3 | 2 | + |
| 11 | 3 | 4 | 1 | + |
| 12 | 2 | 3 | 1 | + |

The test statistic is the number of negative signs which is equal to 3.

Step 5. Conclusion.

We reject H_0 because $3 < 13$. We have statistically significant evidence at $\alpha=0.05$ to show that there is a difference in QOL after chemotherapy treatment as compared to before.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST- WILCOXON SIGNED RANK TEST

Another popular nonparametric test for matched or paired data is called the Wilcoxon Signed Rank Test. Like the Sign Test, it is based on difference scores, but in addition to analyzing the signs of the differences, it also takes into account the magnitude of the observed differences.

| Child | Before Treatment | After 1 Week of Treatment |
|-------|------------------|---------------------------|
| 1 | 85 | 75 |
| 2 | 70 | 50 |
| 3 | 40 | 50 |
| 4 | 65 | 40 |
| 5 | 80 | 20 |
| 6 | 75 | 65 |
| 7 | 55 | 40 |
| 8 | 20 | 25 |

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST- WILCOXON SIGNED RANK TEST

| Child | Before Treatment | After 1 Week of Treatment | Difference (Before-After) |
|-------|---------------------|------------------------------|------------------------------|
| 1 | 85 | 75 | 10 |
| 2 | 70 | 50 | 20 |
| 3 | 40 | 50 | -10 |
| 4 | 65 | 40 | 25 |
| 5 | 80 | 20 | 60 |
| 6 | 75 | 65 | 10 |
| 7 | 55 | 40 | 15 |
| 8 | 20 | 25 | -5 |

The next step is to rank the difference scores. We first order the *absolute values of the difference scores* and assign rank from 1 through n to the smallest through largest absolute values of the difference scores, and assign the mean rank when there are ties in the absolute values of the difference scores.

| Observed Differences | Ordered Absolute Values of Differences | Ranks |
|----------------------|--|-------|
| 10 | -5 | 1 |
| 20 | 10 | 3 |
| -10 | -10 | 3 |
| 25 | 10 | 3 |
| 60 | 15 | 5 |
| 10 | 20 | 6 |
| 15 | 25 | 7 |
| -5 | 60 | 8 |

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST-WILCOXON SIGNED RANK TEST

The final step is to attach the signs ("+" or "-") of the observed differences to each rank as shown below.

| Observed Differences | Ordered Absolute Values of Difference Scores | Ranks | Signed Ranks |
|----------------------|--|-------|--------------|
| 10 | -5 | 1 | -1 |
| 20 | 10 | 3 | 3 |
| -10 | -10 | 3 | -3 |
| 25 | 10 | 3 | 3 |
| 60 | 15 | 5 | 5 |
| 10 | 20 | 6 | 6 |
| 15 | 25 | 7 | 7 |
| -5 | 60 | 8 | 8 |

Similar to the Sign Test, hypotheses for the Wilcoxon Signed Rank Test concern the population median of the difference scores. The research hypothesis can be one- or two-sided. Here we consider a one-sided test.

H_0 : The median difference is zero

H_1 : The median difference is positive

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST-WILCOXON SIGNED RANK TEST

Test Statistic for the Wilcoxon Signed Rank Test

The test statistic for the Wilcoxon Signed Rank Test is W , defined as the smaller of $W+$ (sum of the positive ranks) and $W-$ (sum of the negative ranks). If the null hypothesis is true, we expect to see similar numbers of lower and higher ranks that are both positive and negative (i.e., $W+$ and $W-$ would be similar). If the research hypothesis is true we expect to see more higher and positive ranks (in this example, more children with substantial improvement in repetitive behavior after treatment as compared to before, i.e., $W+$ much larger than $W-$).

In this example, $W+ = 32$ and $W- = 4$. Recall that the sum of the ranks (ignoring the signs) will always equal $n(n+1)/2$. As a check on our assignment of ranks, we have $n(n+1)/2 = 8(9)/2 = 36$ which is equal to $32+4$. The test statistic is $W = 4$.

TESTS WITH MATCHED SAMPLES – NON-PARAMETRIC TEST-WILCOXON SIGNED RANK TEST

Critical Values of W

To determine the appropriate one-sided critical value we need sample size ($n=8$) and our one-sided level of significance ($\alpha=0.05$). For this example, the critical value of W is 5 and the decision rule is to reject H_0 if $W \leq 5$. Thus, we reject H_0 , because $4 \leq 5$.

Conclusion:

We have statistically significant evidence at $\alpha = 0.05$, to show that the median difference is positive (i.e., that repetitive behavior improves.)

Any question?

REFERENCE

- The detail material related to this lecture can be found in
 - D'Agostino RB and Stevens MA. Goodness of Fit Techniques.
 - Apgar, Virginia (1953). "A proposal for a new method of evaluation of the newborn infant ". Curr. Res. Anesth. Analg. 32 (4): 260-267.
 - Conover WJ. Practical Nonparametric Statistics, 2nd edition, New York: John Wiley and Sons.
 - Siegel and Castellan. (1988). "Nonparametric Statistics for the Behavioral Sciences," 2nd edition, New York: McGraw-Hill.



INTRODUCTION TO DATA ANALYTICS

Class # 15

Relation Analysis

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

QUOTE OF THE DAY..

Nothing great was ever achieved without enthusiasm.

- RALPH WALDO EMERSON, American philosopher

THIS TOPIC INCLUDES...

- Introduction
- Measures of Relationship
- Correlation Analysis
 - χ^2 - Test
 - Spearman's Correlation Analysis
 - Pearson's Correlation Analysis
- Regression Analysis
- Auto-Regression Analysis

RELATIONSHIP ANALYSIS

- **Example: Wage Data**

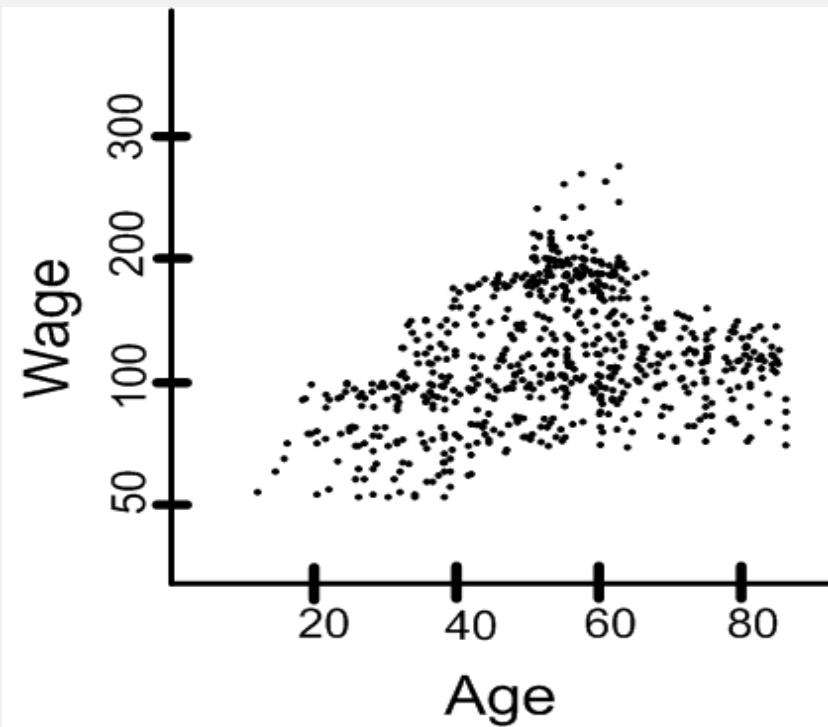
A large data regarding the wages for a group of employees from the eastern region of India is given.

In particular, we wish to understand the following relationships:

- *Employee's age and wage:* How wages vary with ages?
- *Calendar year and wage:* How wages vary with time?
- *Employee's age and education:* Whether wages are anyway related with employees' education levels?

RELATIONSHIP ANALYSIS

- Example: Wage Data
 - Case I. Wage versus Age
 - From the data set, we have a graphical representations, which is as follows:

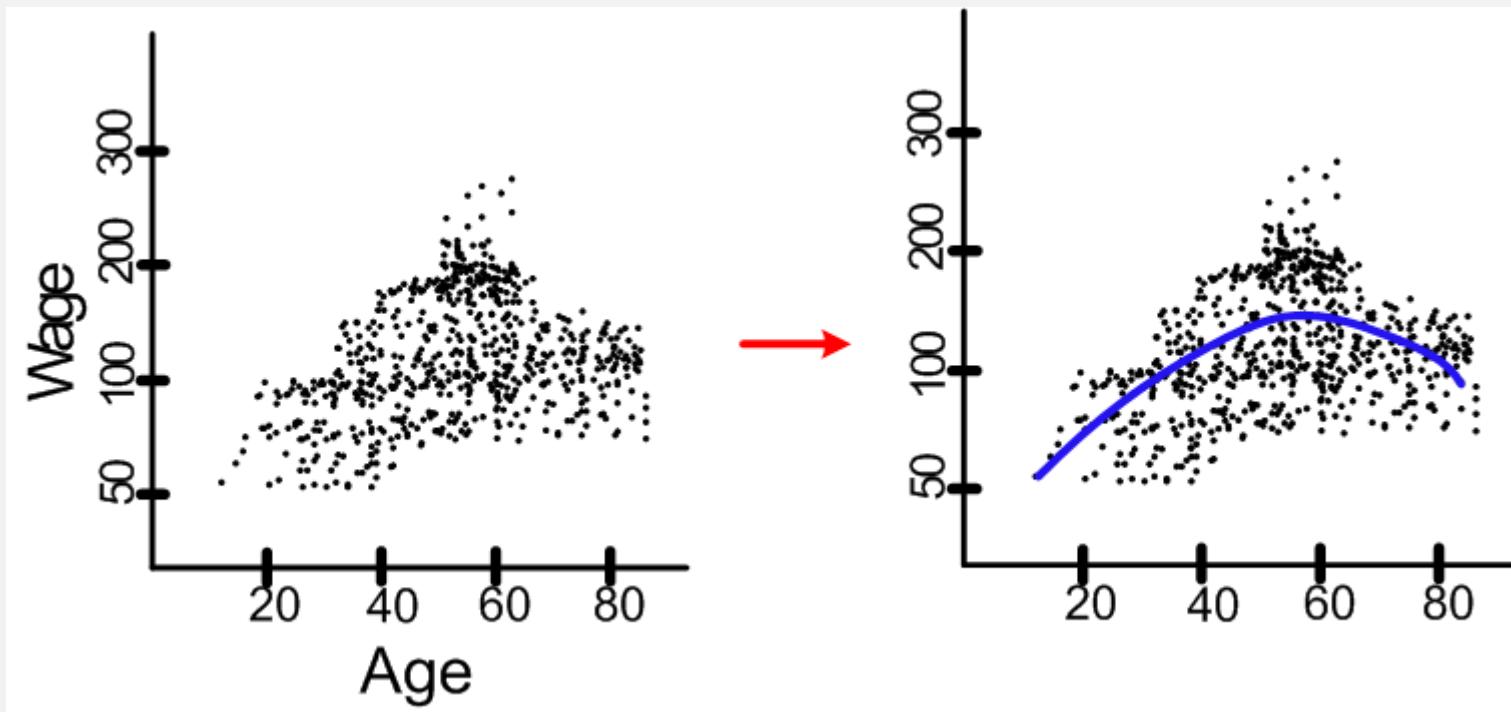


?

How wages vary with ages?

RELATIONSHIP ANALYSIS

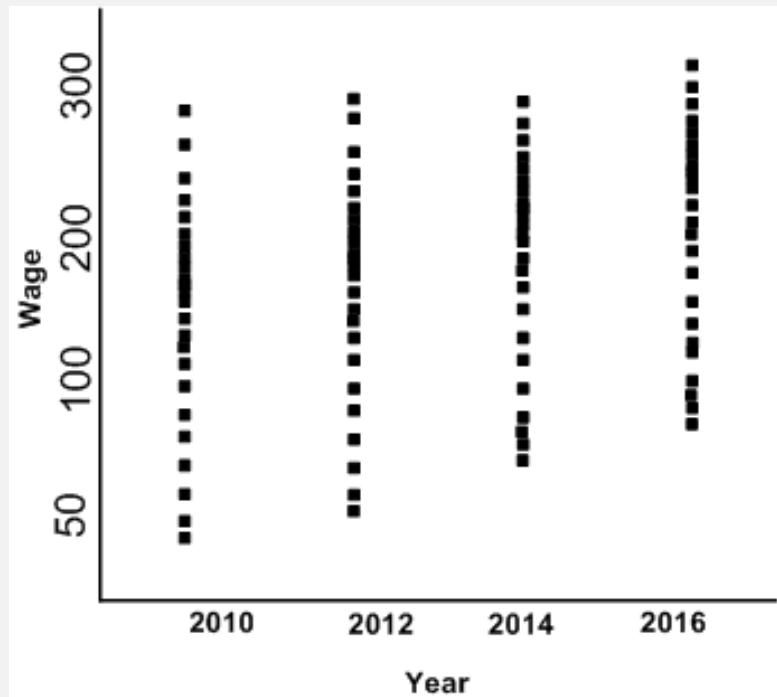
- Example: Wage Data
 - *Employee's age and wage:* How wages vary with ages?



Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

RELATIONSHIP ANALYSIS

- Example: Wage Data
 - Case II. Wage versus Year
 - From the data set, we have a graphical representations, which is as follows:

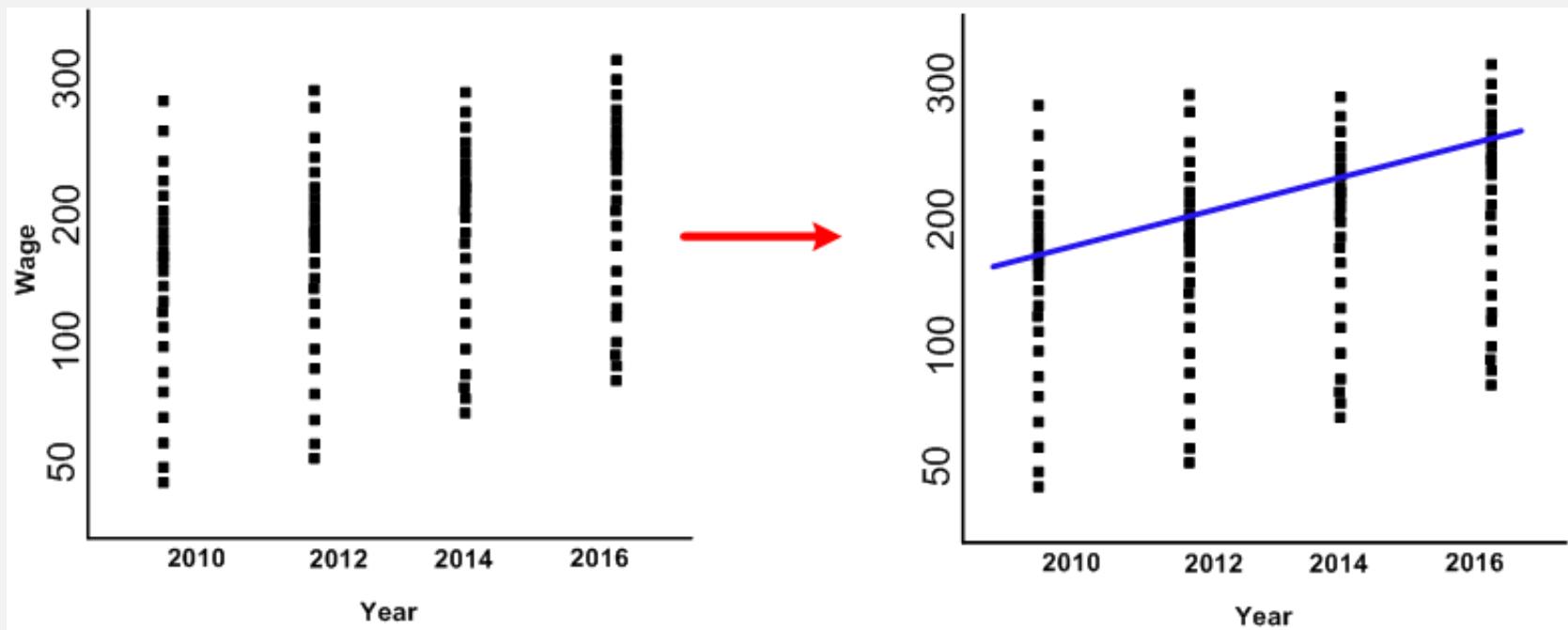


?

How wages vary with time?

RELATIONSHIP ANALYSIS

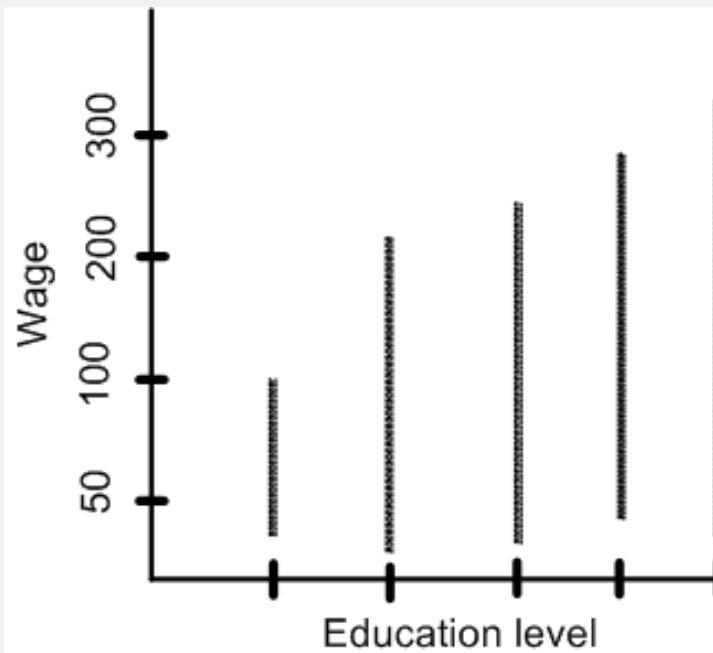
- Example: Wage Data
 - *Wage and calendar year:* How wages vary with years?



Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

RELATIONSHIP ANALYSIS

- Example: Wage Data
 - Case III. Wage versus Education
 - From the data set, we have a graphical representations, which is as follows:

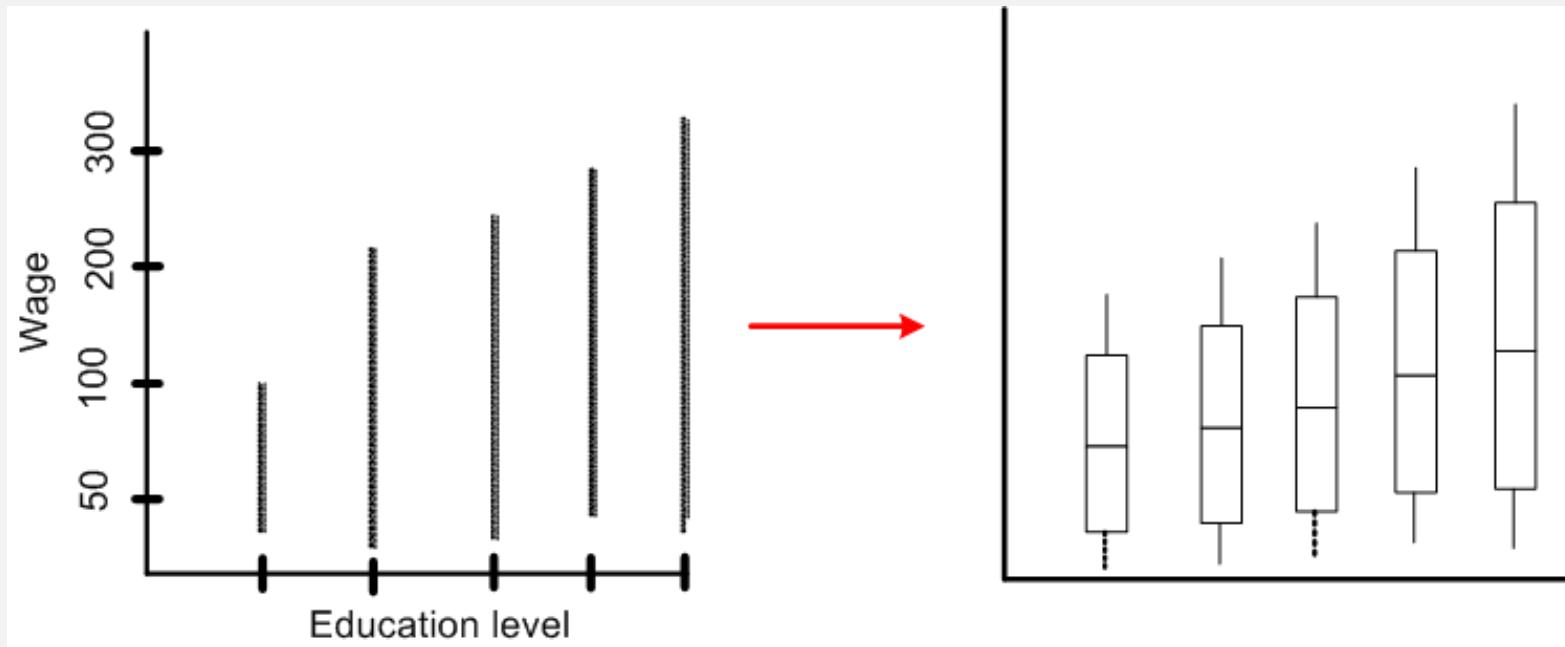


?

Whether wages are related with education?

RELATIONSHIP ANALYSIS

- Example: Wage Data
 - *Wage and education level:* Whether wages vary with employees' education levels?



Interpretation: On the average, wage increases with the level of education.

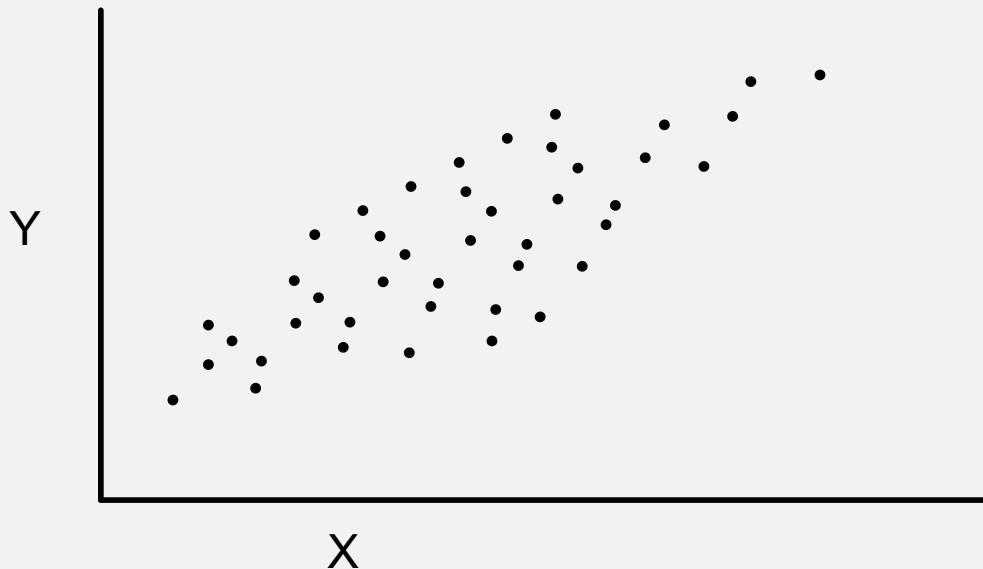
RELATIONSHIP ANALYSIS

Given an employee's wage can we predict his education level?

Whether wage has any association with both age and education level?

etc....

AN OPEN CHALLENGE!

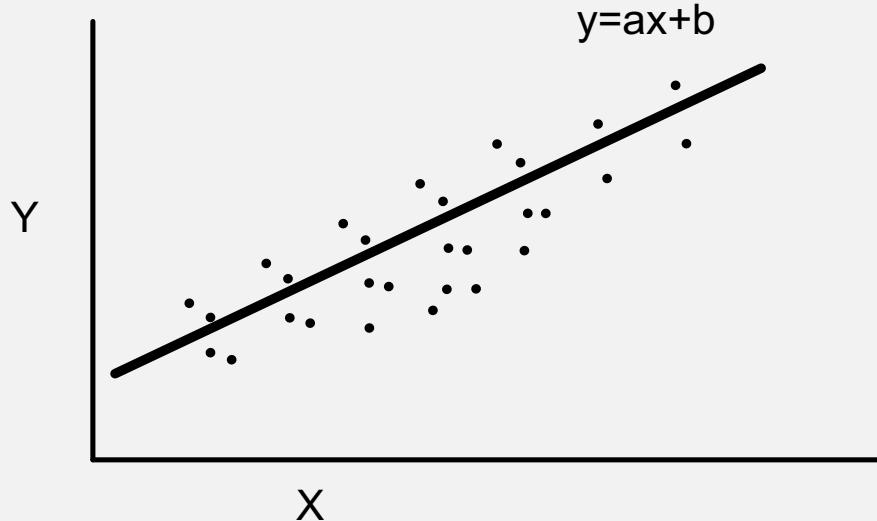


Suppose there are countably infinite points in the XY plane. We need a huge memory to store all such points.

Is there any way out to store this information with a least amount of memory?

Say, with two values only.

YAHOO!



Just decide the values of **a** and **b**
(as if storing one point's data only!)

Note: Here, tricks was to find a relationship among all the points.

MEASURES OF RELATIONSHIP

- *Univariate population:* The population consisting of only one variable.

| | | | | | | | |
|-------------|----|----|----|----|----|----|----|
| Temperature | 20 | 30 | 21 | 18 | 23 | 45 | 52 |
|-------------|----|----|----|----|----|----|----|

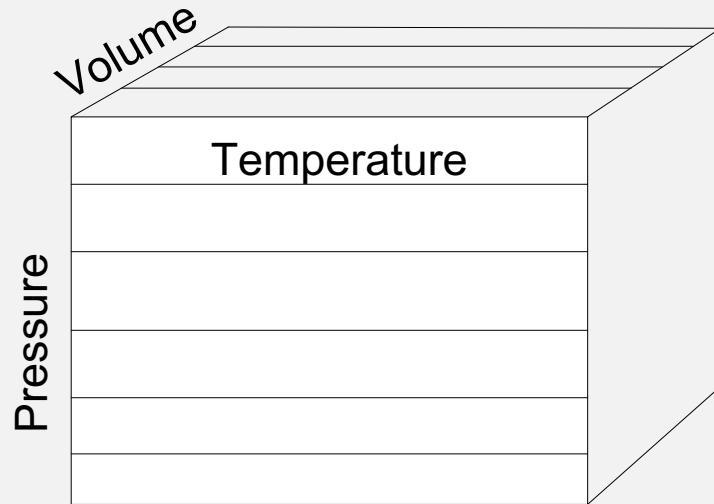
Here, statistical measures are suffice to find a relationship.

- *Bivariate population:* Here, the data happen to be on two variables.

| | | | | |
|-------------|----|-----|------|-----|
| Pressure | 1 | 1.1 | | 0.8 |
| Temperature | 35 | 41 | | 29 |

MEASURES OF RELATIONSHIP

- *Multivariate population:* If the data happen to be more than two variable.



If we add another variable say viscosity in addition to Pressure, Volume or Temperature?

MEASURES OF RELATIONSHIP

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist **correlation** (i.e., association) between two (or more) variables?

If yes, of **what degree**?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?

If yes, of **what degree** and in **which direction**?

To find solutions to the above questions, two approaches are known.

- **Correlation Analysis**
- **Regression Analysis**

CORRELATION ANALYSIS

CORRELATION ANALYSIS

- In statistics, the word **correlation** is used to denote some form of association between two variables.
 - Example: **Weight** is correlated with **height**

Example:

| | | | | | | | |
|-----|-------|-------|-------|-------|-------|-------|-------|
| A | a_1 | a_2 | a_3 | a_4 | a_5 | a_6 | |
| B | b_1 | b_2 | b_3 | b_4 | b_5 | b_6 | |

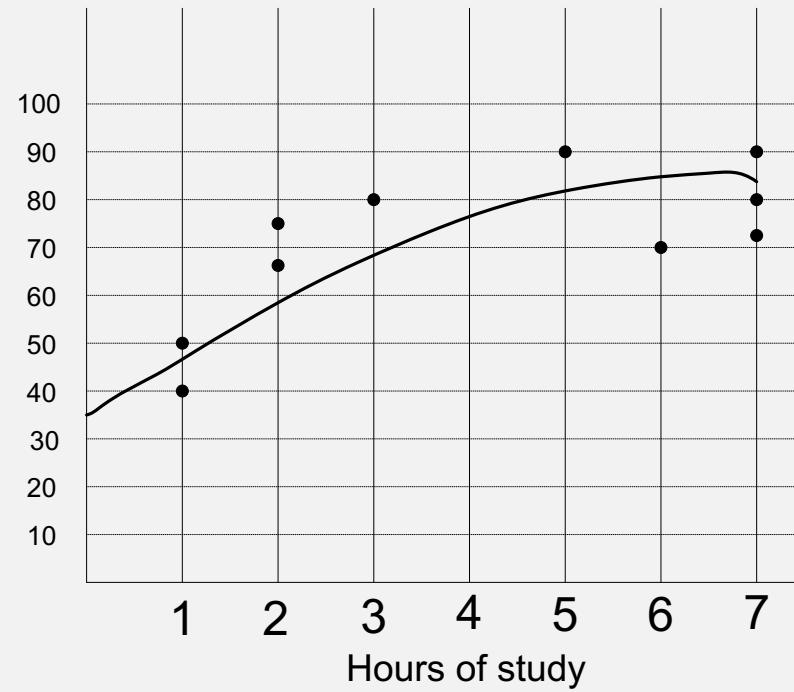
The correlation may be positive, negative or zero.

- **Positive correlation:** If the value of the attribute A **increases with the increase** in the value of the attribute B and vice-versa.
- **Negative correlation:** If the value of the attribute A **decreases with the increase** in the value of the attribute B and vice-versa.
- **Zero correlation:** When the values of attribute A **varies at random** with B and vice-versa.

CORRELATION ANALYSIS

- In order to measure the degree of correlation between two attributes.

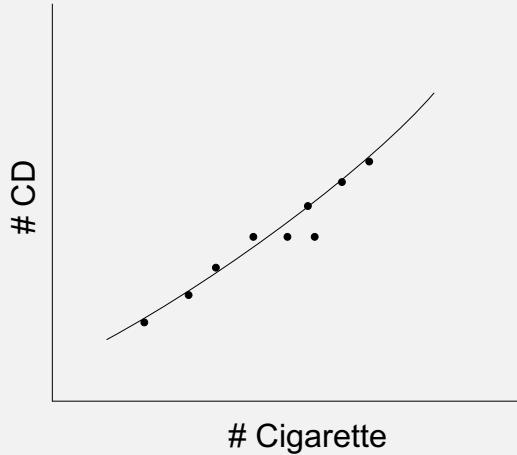
| Hours Study | Exam Score |
|-------------|------------|
| 3 | 80 |
| 5 | 90 |
| 2 | 75 |
| 6 | 80 |
| 7 | 90 |
| 1 | 50 |
| 2 | 65 |
| 7 | 85 |
| 1 | 40 |
| 7 | 100 |



CORRELATION ANALYSIS

- Do you find any correlation between X and Y as shown in the table?.

| | | | | | | | |
|-----------------------------------|----|----|----|----|----|----|----|
| <i>No. of CD's sold in shop X</i> | 25 | 30 | 35 | 42 | 48 | 52 | 56 |
| <i>No. of cigarette sold in Y</i> | 5 | 7 | 9 | 10 | 11 | 11 | 12 |



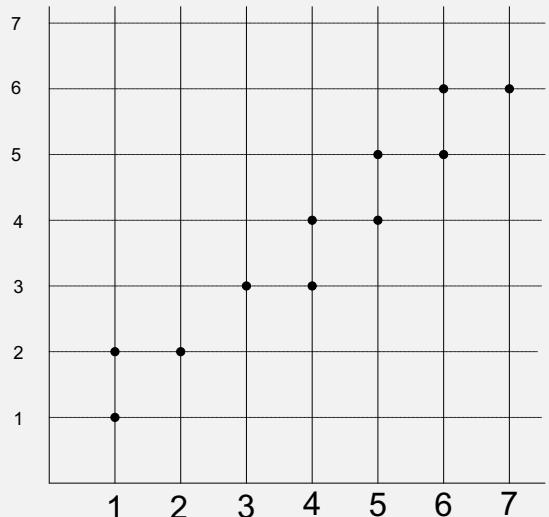
Note:

In data analytics, correlation analysis make sense only when relationship make sense.

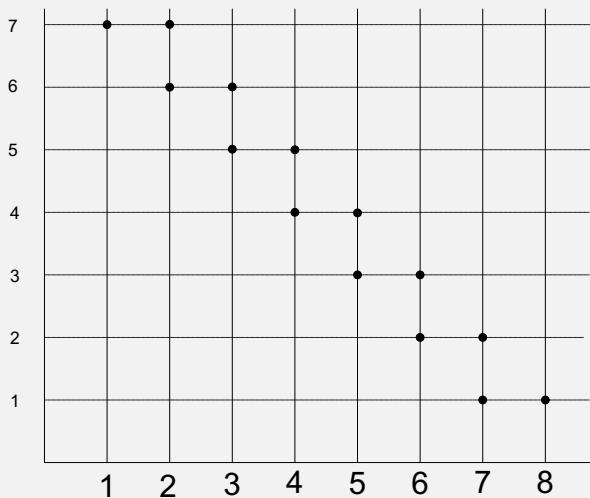
There should be a cause-effect relationship.

CORRELATION ANALYSIS

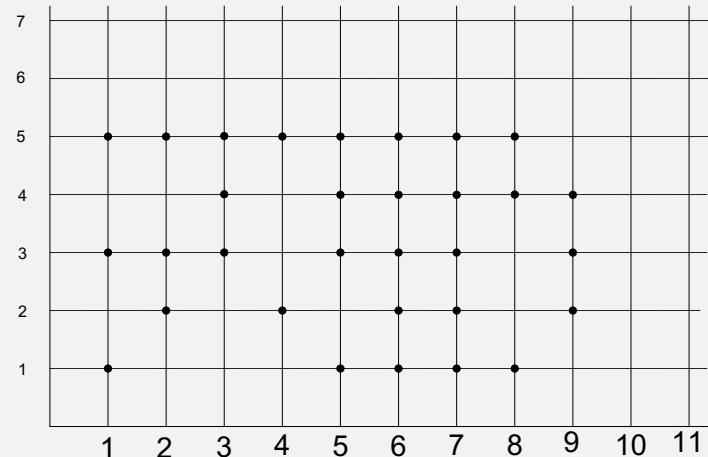
Positive correlation



Negative correlation



Zero correlation

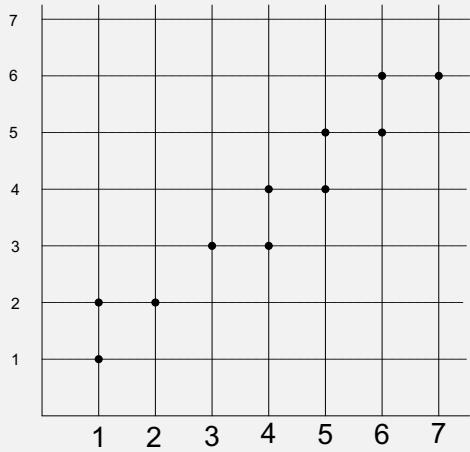


CORRELATION COEFFICIENT

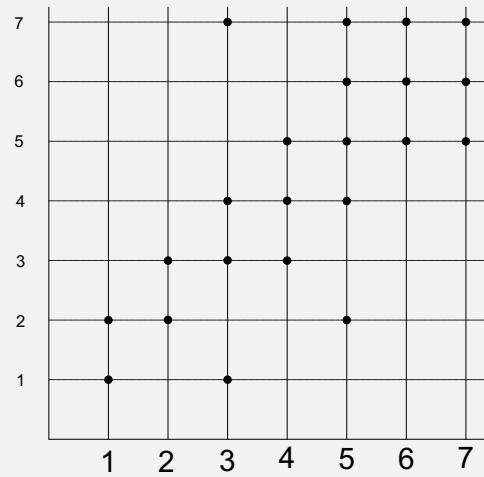
- Correlation coefficient is used to measure the **degree of association**.
- It is usually denoted by r .
- The value of r lies between +1 and -1.
- Positive values of r indicates positive correlation between two variables, whereas, negative values of r indicate negative correlation.
- $r = +1$ implies **perfect positive correlation**, and otherwise.
- The value of r nearer to +1 or -1 indicates **high degree of correlation** between the two variables.
- $r = 0$ implies, there is no correlation

CORRELATION COEFFICIENT

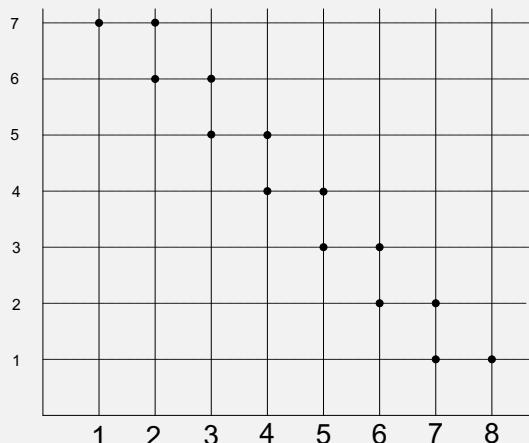
High Positive Correlation



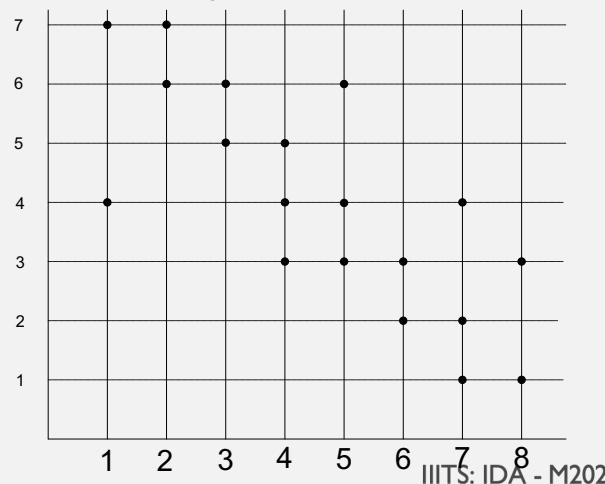
Low Positive Correlation



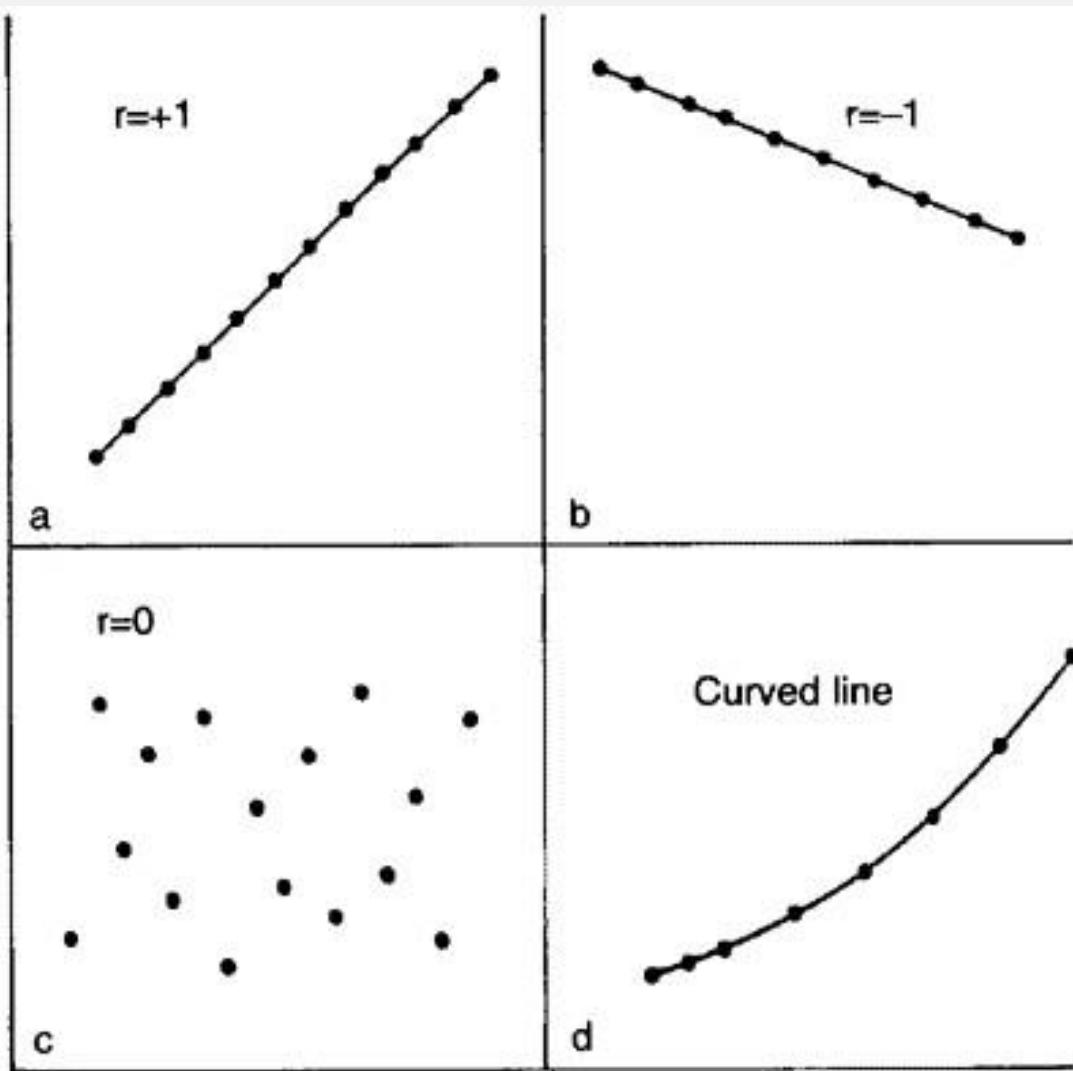
High Negative Correlation



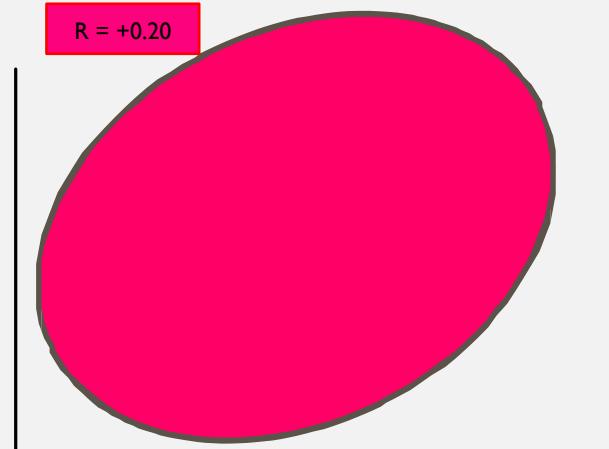
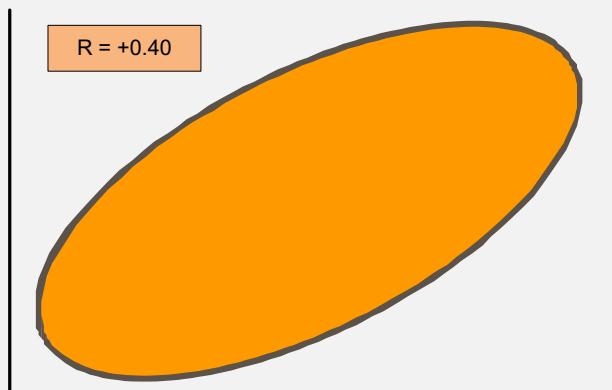
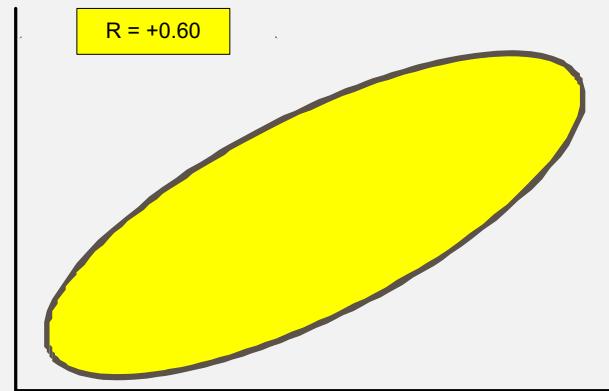
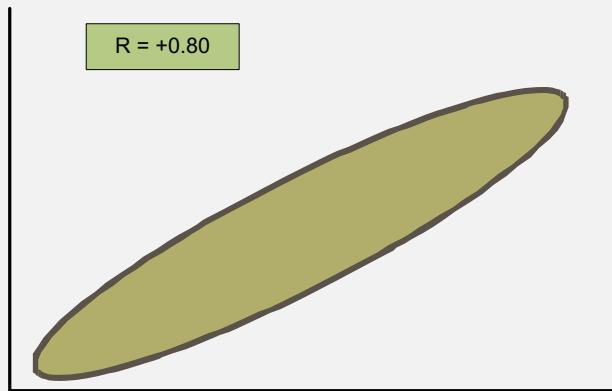
Low Negative Correlation



CORRELATION COEFFICIENT



CORRELATION COEFFICIENT



MEASURING CORRELATION COEFFICIENTS

- There are three methods known to measure the correlation coefficients
 - Karl Pearson's coefficient of correlation
 - This method is applicable to find correlation coefficient between two **numerical** attributes
 - Charles Spearman's coefficient of correlation
 - This method is applicable to find correlation coefficient between two **ordinal** attributes
 - Chi-square coefficient of correlation
 - This method is applicable to find correlation coefficient between two **categorical** attributes

PEARSON'S CORRELATION COEFFICIENT

KARL PEARSON'S CORRELATION COEFFICIENT

- This is also called **Pearson's Product Moment Correlation**

Definition 7.1: Karl Pearson's correlation coefficient

Let us consider two attributes are X and Y .

The Karl Pearson's coefficient of correlation is denoted by r^* and is defined as

$$r^* = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X \cdot \sigma_Y}$$

where X_i = i – th value of X – variable

\bar{X} = mean of X

Y_i = i – th value of Y – variable

\bar{Y} = mean of Y

n = number of pairs of observation of X and Y

σ_X = standard deviations of X

σ_Y = standard deviation of Y

KARL PEARSON'S COEFFICIENT OF CORRELATION

Example 7.1: Correlation of Gestational Age and Birth Weight

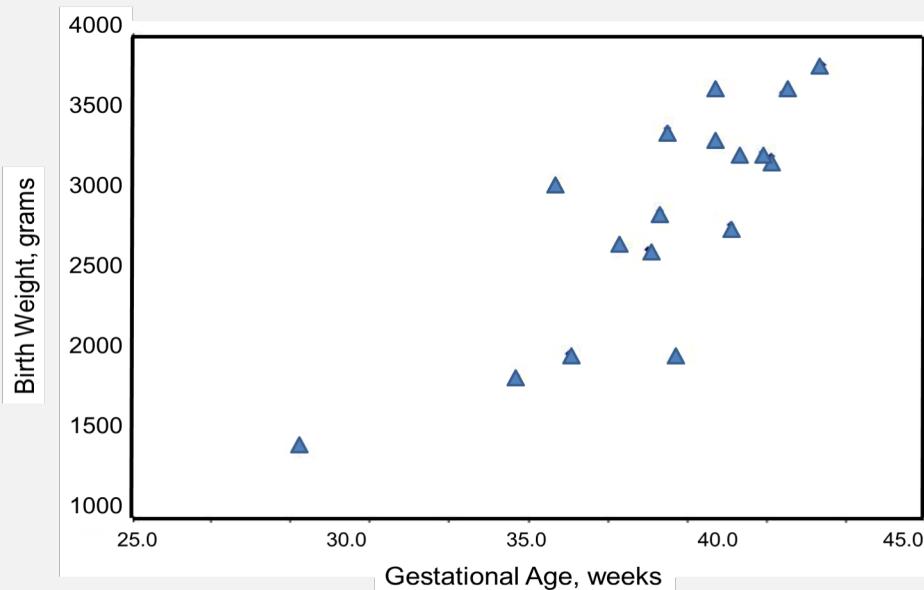
- A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

| Infant ID # | Gestational Age (wks) | Birth Weight (gm) |
|-------------|-----------------------|-------------------|
| 1 | 34.7 | 1895 |
| 2 | 36.0 | 2030 |
| 3 | 29.3 | 1440 |
| 4 | 40.1 | 2835 |
| 5 | 35.7 | 3090 |
| 6 | 42.4 | 3827 |
| 7 | 40.3 | 3260 |
| 8 | 37.3 | 2690 |
| 9 | 40.9 | 3285 |
| 10 | 38.3 | 2920 |
| 11 | 38.5 | 3430 |
| 12 | 41.4 | 3657 |
| 13 | 39.7 | 3685 |
| 14 | 39.7 | 3345 |
| 15 | 41.1 | 3260 |
| 16 | 38.0 | 2680 |
| 17 | 38.7 | 2005 |

KARL PEARSON'S COEFFICIENT OF CORRELATION

Example 7.1: Correlation of Gestational Age and Birth Weight

- We wish to estimate the association between gestational age and infant birth weight.
- In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus Y = birth weight and X = gestational age.
- The data are displayed in a [scatter diagram](#) in the figure below.



KARL PEARSON'S COEFFICIENT OF CORRELATION

Example 7.1: Correlation of Gestational Age and Birth Weight

- For the given data, it can be shown the following

$$\bar{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{49,334}{17} = 2902.$$

$$s_x^2 = \frac{\Sigma (X - \bar{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

$$s_y^2 = \frac{\Sigma (Y - \bar{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,578.8.$$

$$r^* = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{s_x \cdot s_y} = 0.82$$

Conclusion: The sample's correlation coefficient indicates a strong positive correlation between Gestational Age and Birth Weight.

KARL PEARSON'S COEFFICIENT OF CORRELATION

Example 7.1: Correlation of Gestational Age and Birth Weight

- **Significance Test**

- To test whether the association is merely apparent, and might have arisen by chance use the ***t* test** in the following calculation

$$t = r \sqrt{\frac{n - 2}{1 - r^2}}$$

- Number of pair of observation is 17. Hence,

$$t = 0.82 \sqrt{\frac{17 - 2}{1 - 0.82^2}} = 1.44$$

- Consulting the t-test table, at **degrees of freedom 15** and for $\alpha = 0.05$, we find that $t = 1.753$. Thus, the value of Pearson's correlation coefficient in this case **may be regarded as highly significant**.

RANK CORRELATION COEFFICIENT

CHARLES SPEARMAN'S CORRELATION COEFFICIENT

- This correlation measurement is also called **Rank correlation**.
- This technique is applicable to determine the degree of correlation between two variables in case of **ordinal data**.
- We can assign rank to the different values of a variable with ordinal data type.

Example:

| Height: [VS S L T VT] | 1 2 3 4 5 | Rank assigned |
|----------------------------|----------------|---------------|
| T – shirt: [XS S L XL XXL] | 11 12 13 14 15 | |
| | | Rank assigned |

CHARLES SPEARMAN'S CORRELATION COEFFICIENT

Definition 7.2: Charles Spearman's correlation coefficient

The rank correlation can be defined as

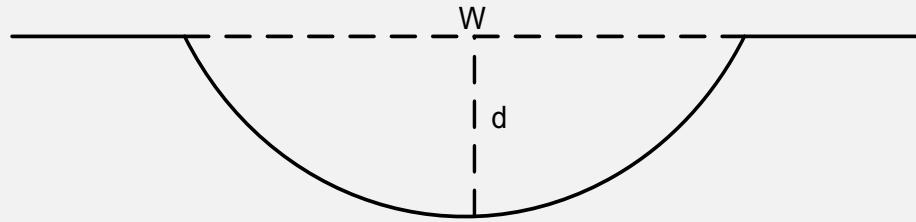
$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i = Difference between ranks of i^{th} pair of the two variables
 n = Number of pairs of observations

- The Spearman's coefficient is often used as a statistical methods to aid either proving or disproving a hypothesis.

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

Example 7.2: The hypothesis that the depth of a river **does not progressively increase** with the width of the river.



A sample of size 10 is collected to test the hypothesis, using Spearman's correlation coefficient.

| Sample# | Width in m | Depth in m |
|---------|------------|------------|
| 1 | 0 | 0 |
| 2 | 50 | 10 |
| 3 | 150 | 28 |
| 4 | 200 | 42 |
| 5 | 250 | 59 |
| 6 | 300 | 51 |
| 7 | 350 | 73 |
| 8 | 400 | 85 |
| 9 | 450 | 104 |
| 10 | 500 | 96 |

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

Step 1: Assign rank to each data. It is customary to assign rank 1 to the largest data, and 2 to next largest and so on.

Note: If there are two or more samples with the same value, the mean rank should be used.

| <i>Data</i> | 20 | 25 | 25 | 25 | 30 |
|--------------------|----|----|----|----|----|
| <i>Assign rank</i> | 5 | 4 | 3 | 2 | 1 |
| <i>Final rank</i> | 5 | 3 | 3 | 3 | 1 |

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

Step 2: The contingency table will look like

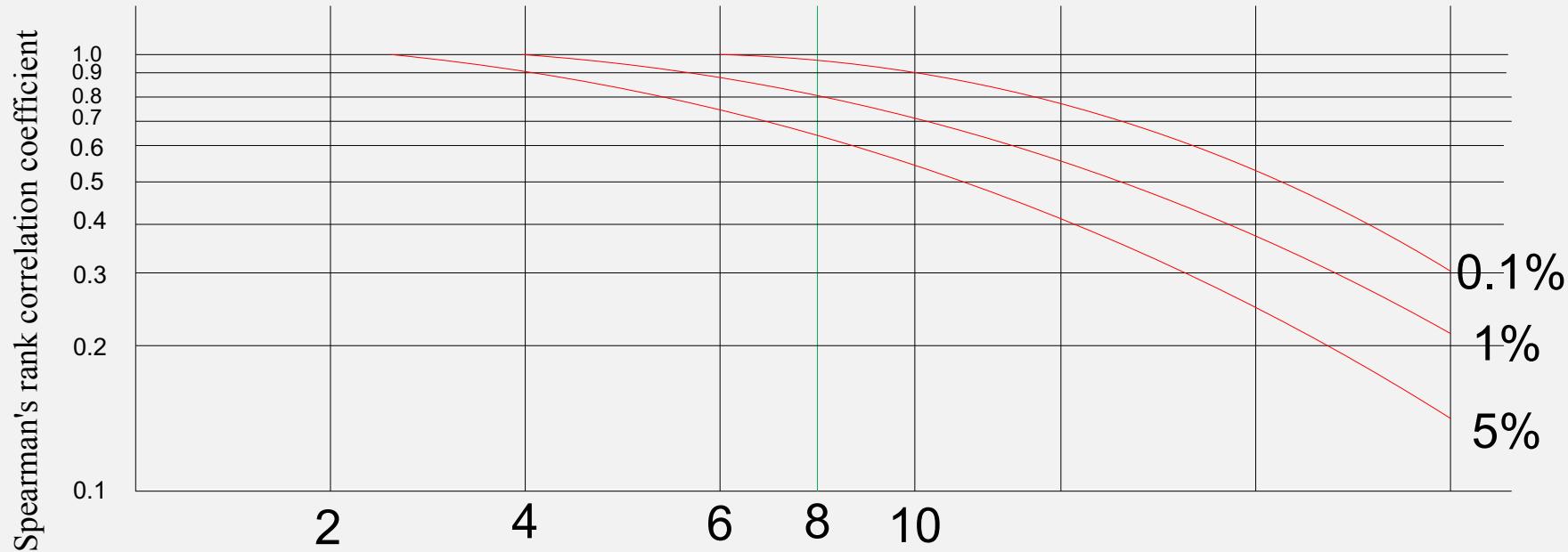
| Sample# | Width | Width rank | Depth | Depth rank | d | d^2 |
|---|-------|------------|-------|------------|----|--------------------|
| 1 | 0 | 10 | 0 | 10 | 0 | 0 |
| 2 | 50 | 9 | 10 | 9 | 0 | 0 |
| 3 | 150 | 8 | 28 | 8 | 0 | 0 |
| 4 | 200 | 7 | 42 | 7 | 0 | 0 |
| 5 | 250 | 6 | 59 | 5 | 1 | 1 |
| 6 | 300 | 5 | 51 | 6 | -1 | 1 |
| 7 | 350 | 4 | 73 | 4 | 0 | 0 |
| 8 | 400 | 3 | 85 | 3 | 0 | 0 |
| 9 | 450 | 2 | 104 | 1 | 1 | 1 |
| 10 | 500 | 1 | 96 | 2 | -1 | 1 |
| $r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 4}{10 \times 99}$ | | | | | | $\sum d^2 = 4$ |
| $r_s = 0.9757$ | | | | | | IIITS: IDA - M2021 |

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

Step 3: To see, if this r_s value is significant, the Spearman's rank significance table (or graph) must be consulted.

Note: The degrees of freedom for the sample = $n - 2 = 8$

Assume, the significance level = 0.01



CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

Step 4: Final conclusion

From the graph, we see that $r_s = 0.9757$ lies above the line at 8 and 0.01 significance level. Hence, there is a greater than 99% chance that the relationship is significant (i.e., not random) and hence the hypothesis should be rejected.

Thus, we can reject the hypothesis and conclude that in this case, depth of a river **progressively increases** the further with the width of the river.

X²-CORRELATION ANALYSIS

CHI-SQUARED TEST OF CORRELATION

- This method is also alternatively termed as Pearson's χ^2 -test or simply χ^2 -test
- This method is applicable to categorical (discrete) data only.
- Suppose, two attributes A and B with categorical values

$$A = a_1, a_2, a_3, \dots, a_m \quad \text{and}$$

$$B = b_1, b_2, b_3, \dots, b_n$$

having m and n distinct values.

| | | | | | | | |
|-----|-------|-------|-------|-------|-------|-------|------|
| A | a_1 | a_2 | a_3 | a_1 | a_5 | a_1 | |
| B | b_1 | b_2 | b_3 | b_1 | b_5 | b_1 | |

Between whom we are to find the correlation relationship.

χ^2 – TEST METHODOLOGY

Contingency Table

Given a data set, it is customary to draw a contingency table, whose structure is given below.

| | b_1 | b_2 | ----- | b_j | ----- | b_n | Row Total |
|---------------------|-------|-------|-------|-------|-------|-------|--------------------|
| a_1 | | | | | | | |
| a_2 | | | | | | | |
| ⋮ | | | | | | | |
| a_i | | | | | | | |
| ⋮ | | | | | | | |
| a_m | | | | | | | |
| Column Total | | | | | | | Grand Total |

χ^2 – TEST METHODOLOGY

Entry into Contingency Table: Observed Frequency

In contingency table, an entry O_{ij} denotes the event that attribute A takes on value a_i and attribute B takes on value b_j (i.e., $A = a_i, B = b_j$).

| | | | | | | | |
|-----|-------|-------|-------|-------|-------|-------|-------|
| A | a_1 | a_2 | a_3 | a_i | a_5 | a_i | |
| B | b_j | b_2 | b_3 | b_j | b_5 | b_j | |

| | b_1 | b_2 | | b_j | | b_n | Row Total |
|---------------------|-------|-------|-------|----------|-------|-------|--------------------|
| a_1 | | | | | | | |
| a_2 | | | | | | | |
| ⋮ | | | | | | | |
| a_i | | | | O_{ij} | | | |
| ⋮ | | | | | | | |
| a_m | | | | | | | |
| Column Total | | | | | | | Grand Total |

χ^2 – TEST METHODOLOGY

Entry into Contingency Table: Expected Frequency

In contingency table, an entry e_{ij} denotes the expected frequency, which can be calculated as

$$e_{ij} = \frac{\text{Count}(A = a_i) \times \text{Count}(B = b_j)}{\text{Grand Total}} = \frac{A_i \times B_j}{N}$$

| | b ₁ | b ₂ | | b _j | | b _n | Row Total |
|---------------------|----------------|----------------|-------|----------------|-------|----------------|------------------|
| a ₁ | | | | | | | |
| a ₂ | | | | | | | |
| ⋮ | | | | | | | |
| a _i | | | | e_{ij} | | | A _i |
| ⋮ | | | | | | | |
| a _m | | | | | | | |
| Column Total | | | | B _j | | | N |

X² – TEST

Definition 7.3: χ^2 -Value

The χ^2 value (also known as the Pearson's χ^2 test) can be computes as

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency

e_{ij} is the expected frequency

χ^2 – TEST

- The cell that contribute the most to the χ^2 value are those whose actual count is very different from the expected.
- The χ^2 statistics tests the hypothesis that A and B are independent. The test is based on a significance level, with $(n-1) \times (m-1)$ degrees of freedom., with a contingency table of size $n \times m$
- If the hypothesis can be rejected, then we say that A and B are statistically related or associated.

χ^2 – TEST

Example 7.3: Survey on Gender versus Hobby.

- Suppose, a survey was conducted among a population of size 1500. In this survey, gender of each person and their hobby as either “book” or “computer” was noted. The survey result obtained in a table like the following.

| GENDER | HOBBY |
|--------|----------|
| | |
| | |
| M | Book |
| F | Computer |
| | |
| | |
| | |

- We have to find if there is any association between **Gender** and **Hobby** of a people, that is, we are to test whether “gender” and “hobby” are correlated.

χ^2 – TEST

Example 7.3: Survey on Gender versus Hobby.

- From the survey table, the observed frequency are counted and entered into the contingency table, which is shown below.

| | | GENDER | | Total |
|-------|----------|--------|--------|-------|
| HOBBY | | Male | Female | |
| | Book | 250 | 200 | 450 |
| | Computer | 50 | 1000 | 1050 |
| Total | | 300 | 1200 | 1500 |

χ^2 – TEST

Example 7.3: Survey on Gender versus Hobby.

- From the survey table, the **expected frequency** are counted and entered into the contingency table, which is shown below.

| | | GENDER | | |
|-------|----------|--------|--------|-------|
| | | Male | Female | Total |
| HOBBY | Book | 90 | 360 | 450 |
| | Computer | 210 | 840 | 1050 |
| Total | | 300 | 1200 | 1500 |

χ^2 – TEST

- Using equation for χ^2 computation, we get

$$\begin{aligned}\chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} \\ &= 507.93\end{aligned}$$

- This value needs to be compared with the tabulated value of χ^2 (available in any standard book on statistics) with 1 degree of freedom (for a table of $m \times n$, the degrees of freedom is $(m - 1) \times (n - 1)$; here $m = 2, n = 2$).
- For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.01 significance level is 10.828. Since our computed value is above this, we reject the hypothesis that “Gender” and “Hobby” are independent and hence, conclude that the two attributes are *strongly correlated* for the given group of people.

χ^2 – TEST

Example 7.4: Hypothesis on “accident proneness” versus “driver’s handedness”.

- Consider the following contingency table on car accidents among left and right-handed drivers’ of sample size 175.
- Hypothesis is that “*fatality of accidents is independent of driver’s handedness*”

| | | HANDEDNESS | | Total |
|----------|-----------|-------------|--------------|-------|
| FATALITY | Non-Fatal | Left-Handed | Right-Handed | |
| | | 8 | 141 | 149 |
| | Fatal | 3 | 23 | 26 |
| Total | | 11 | 164 | 175 |

- Find the correlation between Fatality and Handedness and test the significance of the correlation with significance level 0.1%.

REFERENCE

- The detail material related to this lecture can be found in

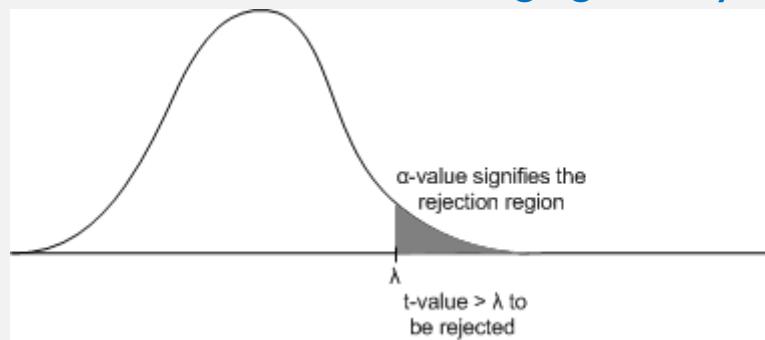
The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.

Any question?

QUESTIONS OF THE DAY...

1. For a given sample data the correlation coefficient according to the Karl Pearson's correlation analysis is found to be $r = 0.79$ with degree of freedom 69. Further, with significant test , the t-value is calculated as $t = 2.36$. From the t-test table, it is found that with degree of freedom 69, the t-value at 5% confidence level is 3.61. What is the inference that you can have in this case?

2. For a given degree of freedom, if α , the value of confidence level increases, then t-value increases. Is the statement correct? If not, what is the correct statement? Justify your answer. You can refer the following figure in your explanation.



QUESTIONS OF THE DAY...

3. Whether the Spearman's correlation coefficient analysis is applicable to the numeric data? If so, how?

4. Can χ^2 -analysis be applied to ordinal data or numeric data? Justify your answer.

5. Briefly explain the following with reference to the χ^2 correlation analysis.
 - a) Contingency table
 - b) Observed frequency
 - c) Expected frequency
 - d) Expression for -vale calculation
 - e) Hypothesis to be tested
 - f) Degree of freedom of sample data



INTRODUCTION TO DATA ANALYTICS

Class # 16

Relation Analysis

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

THIS TOPIC INCLUDES...

- Regression Analysis
 - Simple Linear Regression
 - Multiple Linear Regression
 - Non-Linear Regression Analysis
- Auto-Regression Analysis

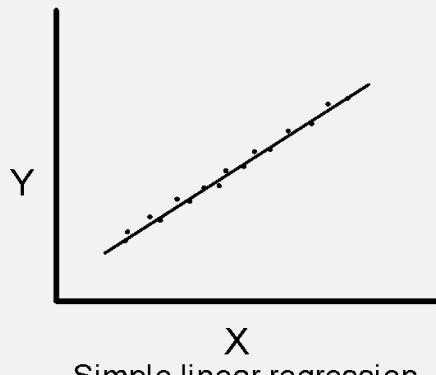
REGRESSION ANALYSIS

REGRESSION ANALYSIS

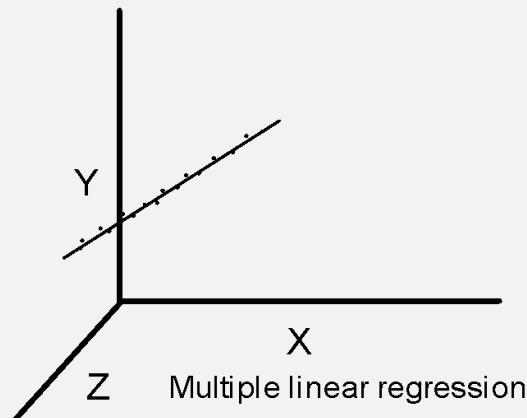
- The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

Classification of Regression Analysis Models

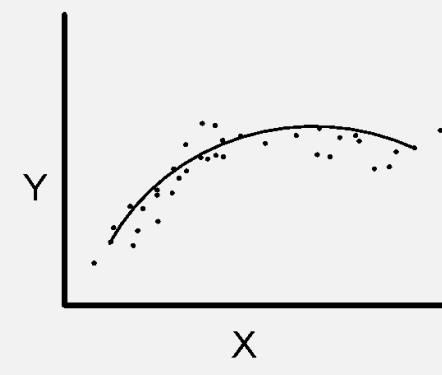
- Linear regression models
 - 1. Simple linear regression
 - 2. Multiple linear regression
- Non-linear regression models



Simple linear regression



Multiple linear regression



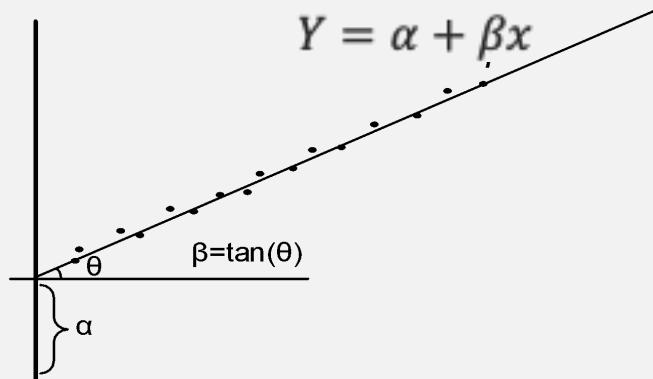
Non-linear regression

IIITS: IDA - M2021

SIMPLE LINEAR REGRESSION MODEL

In simple linear regression, we have only two variables:

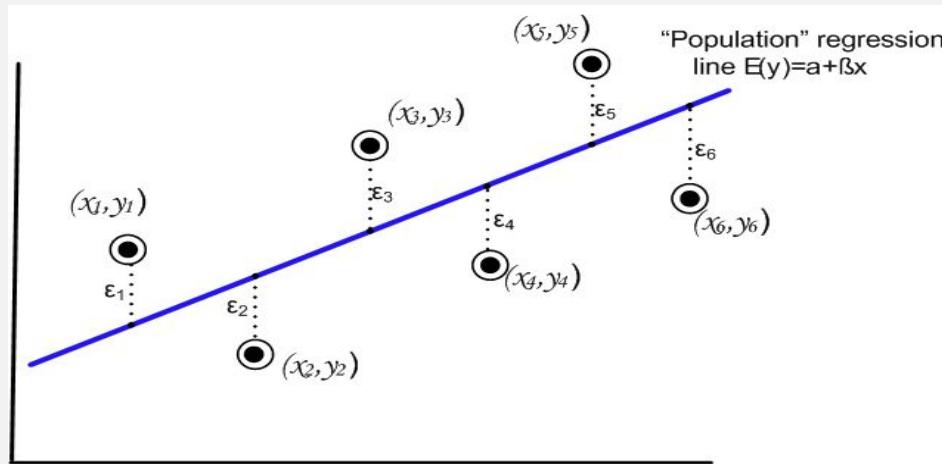
- Dependent variable (also called **Response**), usually denoted as Y .
- Independent variable (alternatively called **Regressor**), usually denoted as x .
- A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$



Note:

- There are infinite number of lines (and hence α_s and β_s)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α and β) quantifying the strength of that relationship.

REGRESSION ANALYSIS



Given the set $[(x_i, y_i), i = 1, 2, \dots, n]$ of data involving n pairs of (x, y) values, our objective is to find “true” or population regression line such that $Y = \alpha + \beta x + \epsilon$

Here, ϵ is a random variable with $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$. The quantity σ^2 is often called the **error variance**.

Note:

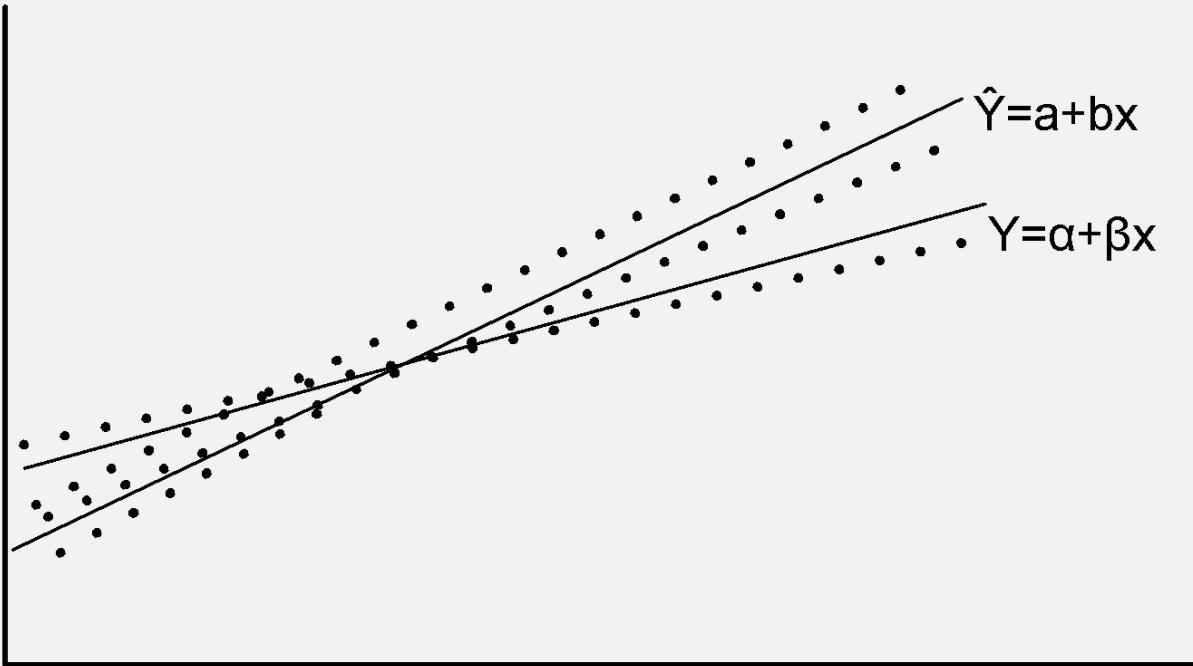
- $E(\epsilon) = 0$ implies that at a specific x , the y values are distributed around the “true” regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- α and β are called **regression coefficients**.
- α and β values are to be estimated from the data.

TRUE VERSUS FITTED REGRESSION LINE

- The task in regression analysis is to estimate the regression coefficients α and β .
- Suppose, we denote the estimates a for α and b for β . Then the fitted regression line is

$$\hat{Y} = a + bx$$

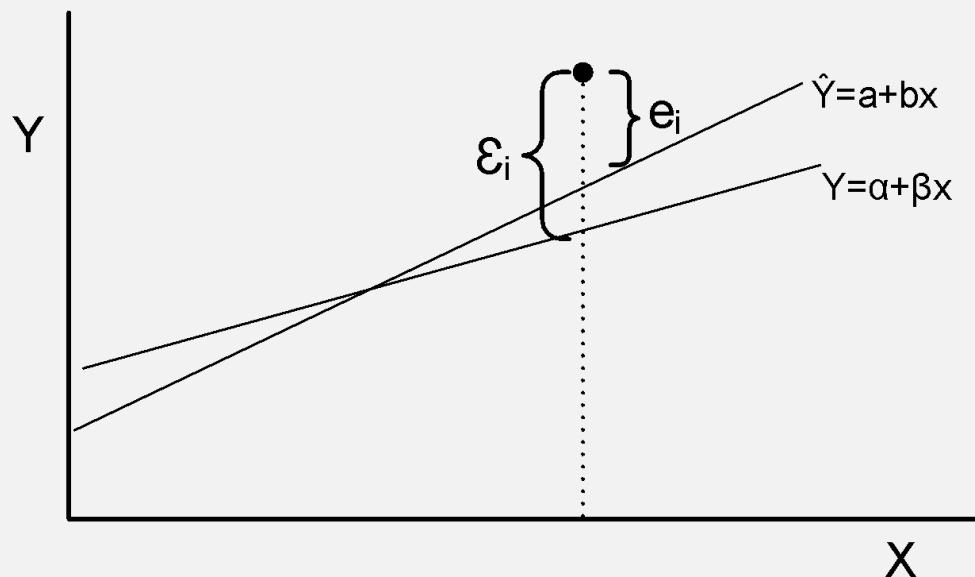
where \hat{Y} is the predicted or fitted value.



LEAST SQUARE METHOD TO ESTIMATE α AND β

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\hat{Y} = a + bx$. Thus, i^{th} residual is

$$e_i = Y_i - \hat{Y}_i, i = 1, 2, 3, \dots, n$$



LEAST SQUARE METHOD

- The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as SSE

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

- We are to minimize the value of SSE and hence to determine the parameters of a and b .
- Differentiating SSE with respect to a and b , we have

$$\frac{\partial(SSE)}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i)$$

$$\frac{\partial(SSE)}{\partial b} = -2 \sum_{i=1}^n (y_i - a - bx_i) \cdot x_i$$

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$

$$\frac{\partial(SSE)}{\partial b} = 0$$

LEAST SQUARE METHOD TO ESTIMATE α AND β

Thus we set

$$na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

These two equations can be solved to determine the values of a and b , and it can be calculated that

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

R^2 : MEASURE OF QUALITY OF FIT

- A quantity R^2 , is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.
- We have $SSE = \sum_{i=1}^n (y_i - \hat{y})^2$
- It signifies the **variability due to error**.
- Now, let us define the **total corrected sum of squares**, defined as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

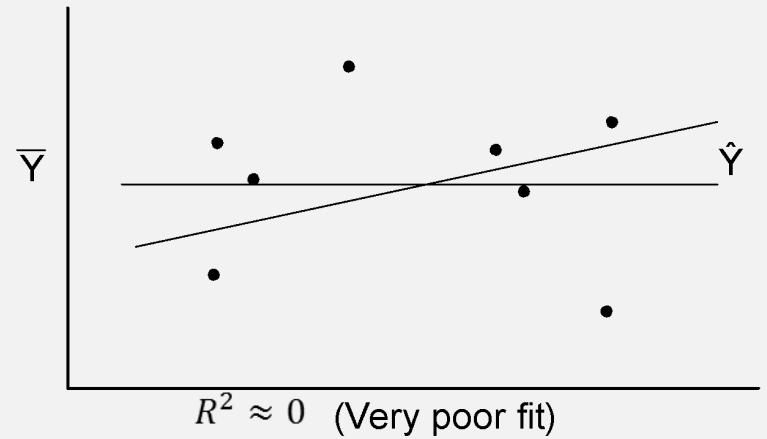
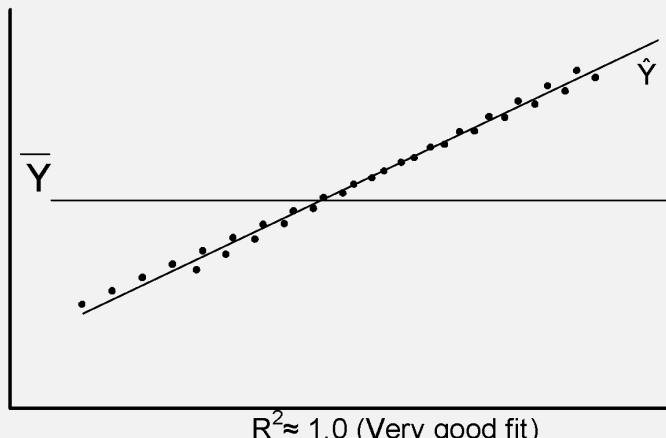
- SST represents the variation in the response values. The R^2 is

$$R^2 = 1 - \frac{SSE}{SST}$$

Note:

- If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)

R^2 : MEASURE OF QUALITY OF FIT



MULTIPLE LINEAR REGRESSION

- When more than one variable are independent variable, then the regression can be estimated as a **multiple regression model**
- When this model is linear in coefficients, it is called **multiple linear regression model**
- If k -independent variables $x_1, x_2, x_3, \dots, x_k$ are associated, the multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \epsilon$$

- And the estimated response is obtained as

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + b_k x_k$$

MULTIPLE LINEAR REGRESSION

Estimating the coefficients

Let the data points given to us is

$$(x_{1i}, x_{2i}, x_{3i}, \dots, \dots, \dots, x_{ki}, y_i) \quad i = 1, 2, \dots, n, \quad n > k$$

where y_i is the observed response to the values $x_{1i}, x_{2i}, x_{3i}, \dots, \dots, \dots, x_{ki}$ of k independent variables $x_1, x_2, x_3, \dots, \dots, \dots, x_k$.

Thus,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + \epsilon_i$$

$$\text{and} \quad \hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_k x_{ki} + e_i$$

where ϵ_i and e_i are the random error and residual error, respectively associated with true response y_i and fitted response \hat{y}_i .

Using the concept of **Least Square Method** to estimate $b_0, b_1, b_2, \dots, b_k$, we minimize the expression

$$\text{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

MULTIPLE LINEAR REGRESSION

- Differentiating SSE in turn with respect to $b_0, b_1, b_2, \dots, b_k$ and equating to zero, we generate the set of $(k+1)$ normal **estimation equations for multiple linear regression**.

$$nb_0 + b_1 \sum_{i=1}^n x_{1i} + b_2 \sum_{i=1}^n x_{2i} + \dots + b_k \sum_{i=1}^n x_{ki} = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n x_{1i}^2 + b_2 \sum_{i=1}^n x_{1i} \cdot x_{2i} + \dots + b_k \sum_{i=1}^n x_{1i} \cdot x_{ki} = \sum_{i=1}^n x_i \cdot y_i$$

...

...

$$b_0 \sum_{i=1}^n x_{ki} + b_1 \sum_{i=1}^n x_{ki} \cdot x_{1i} + b_2 \sum_{i=1}^n x_{ki} \cdot x_{2i} + \dots + b_k \sum_{i=1}^n x_{ki}^2 = \sum_{i=1}^n x_i \cdot y_i$$

- The system of linear equations can be solved for b_0, b_1, \dots, b_k by any appropriate method for solving system of linear equations.
- Hence, the multiple linear regression model can be built.

NON LINEAR REGRESSION MODEL

- When the regression equation is in terms of r -degree, $r>1$, then it is called nonlinear regression model. When more than one independent variables are there, then it is called Multiple Non linear Regression model. Also, alternatively termed as polynomial regression model. In general, it takes the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_r x^r + \epsilon$$

- The estimated response is obtained as

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \dots + b_r x^r$$

SOLVING FOR POLYNOMIAL REGRESSION MODEL

Given that $(x_i, y_i); i = 1, 2, \dots, n$ are n pairs of observations. Each observations would satisfy the equations:

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_r x^r + \epsilon_i$$

and $\hat{y}_i = b_0 + b_1 x + b_2 x^2 + \dots + b_r x^r + e_i$

where, r is the degree of polynomial

ϵ_i = is the i^{th} random error

e_i = is the i^{th} residual error

Note: The number of observations, n , must be at least as large as $r+1$, the number of parameters to be estimated.

The polynomial model can be transformed into a general linear regression model setting $x_1 = x, x_2 = x^2, \dots, x_n = x^r$. Thus, the equation assumes the form:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x^r + \epsilon_i$$

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_r x_r + e_i$$

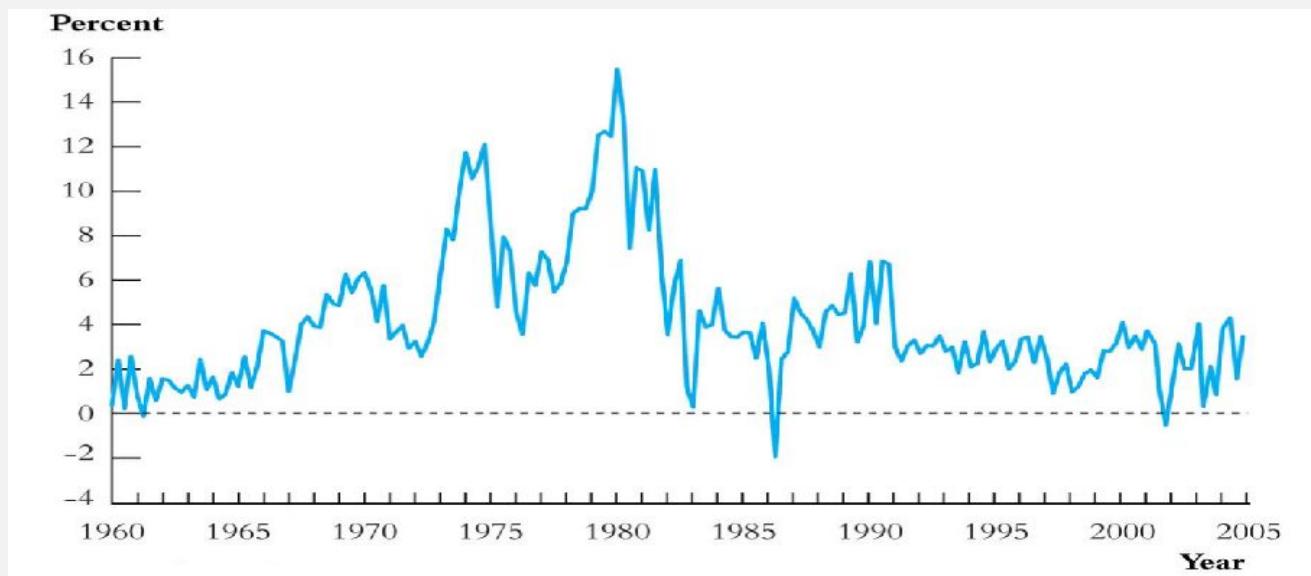
This model then can be solved using the procedure followed for multiple linear regression model.

AUTO-REGRESSION ANALYSIS

AUTO REGRESSION ANALYSIS

- Regression analysis for time-ordered data is known as Auto-Regression Analysis
- Time series data are data collected on the same observational unit at multiple time periods

Example: Indian rate of price inflation

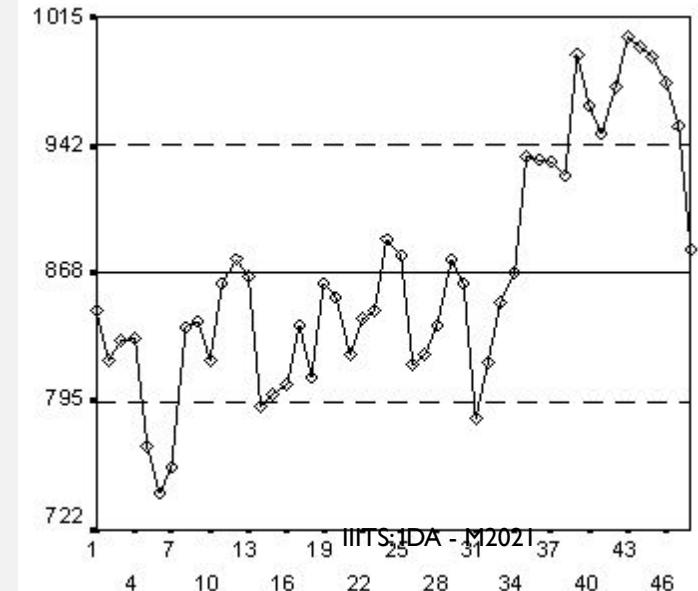
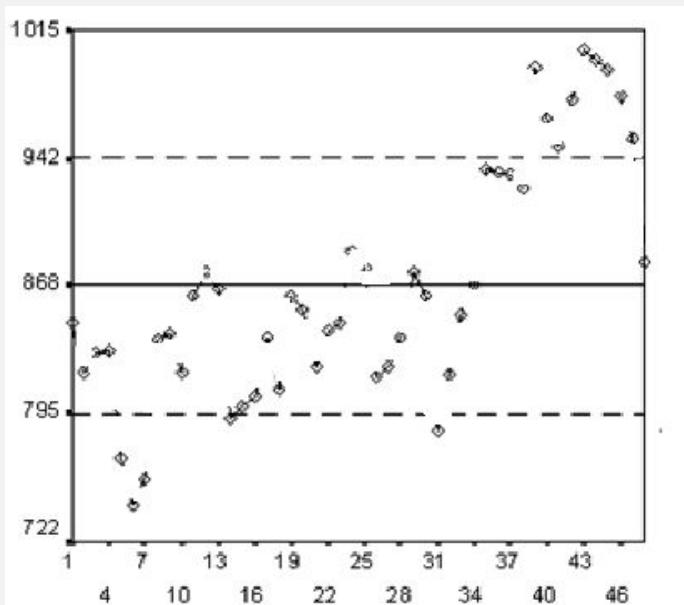
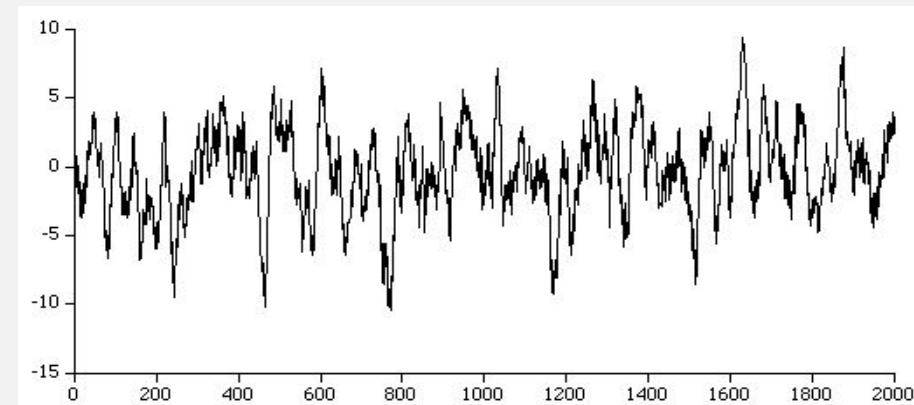
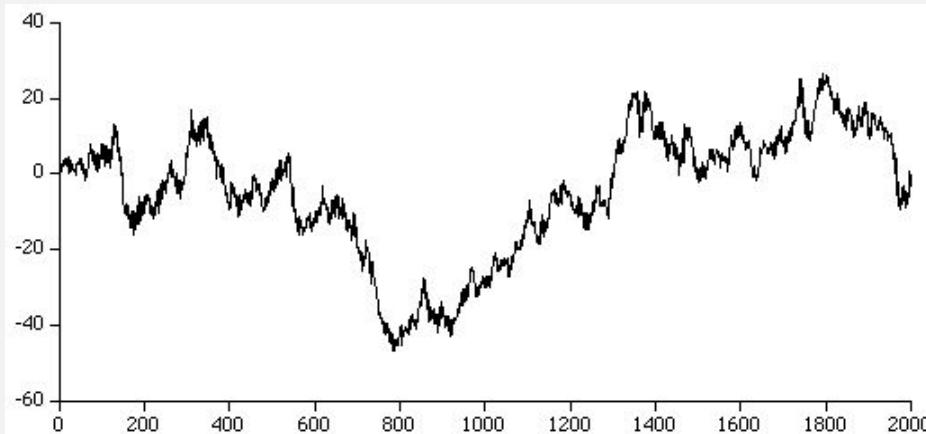


AUTO REGRESSION ANALYSIS

- **Examples:** Which of the following is a time-series data?
 - Aggregate consumption and GDP for a country (for example, 20 years of quarterly observations = 80 observations)
 - Yen/\$, pound/\$ and Euro/\$ exchange rates (daily data for 1 year = 365 observations)
 - Cigarette consumption per capita in a state, by years
 - Rainfall data over a year
 - Sales of tea from a tea shop in a season

AUTO REGRESSION ANALYSIS

- **Examples:** Which of the following graph is due to time-series data?



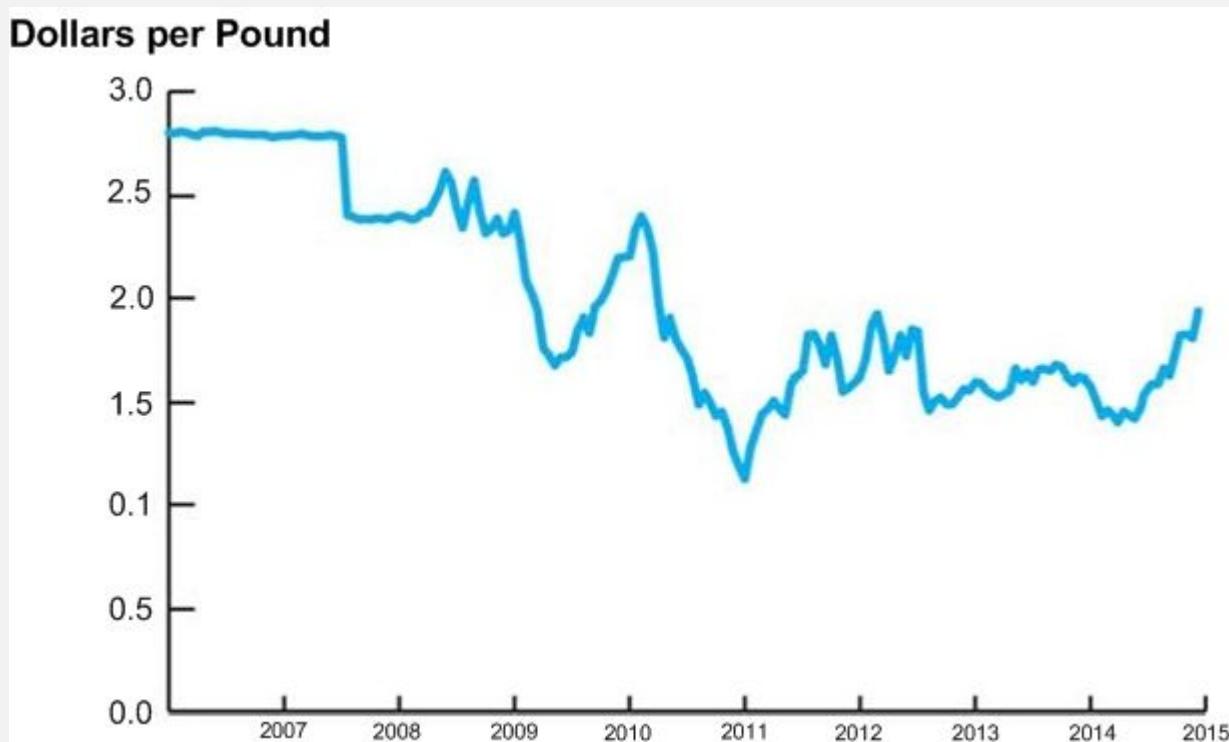
USE OF TIME SERIES DATA

- To develop forecast model
 - What will the rate of inflation by next year?
- To estimate dynamic causal effects
 - If the rate of interest increases the interest rate now, what will be the effect on the rates of inflation and unemployment in 3 months? in 12 months?
 - What is the effect over time on electronics good consumption of a hike in the excise duty?
- Time dependent analysis
 - Rates of inflation and unemployment in the country can be observed only over time!

MODELING WITH TIME SERIES DATA

- Correlation over time
 - Serial correlation, also called autocorrelation
 - Calculating standard error
- To estimate dynamic causal effects
 - Under which dynamic effects can be estimated?
 - How to estimate?
- Forecasting model
 - Forecasting model build on regression model

AUTO-REGRESSION MODEL FOR FORECASTING



- Can we predict the trend at a time say 2022?

SOME NOTATIONS AND CONCEPTS

- Y_t = Value of Y in a period t
- Data set $[Y_1, Y_2, \dots, Y_{T-1}, Y_T]$: T observations on the time series random variable Y
- **Assumptions**
 - We consider only consecutive, evenly spaced observations
 - For example, monthly, 2000-2015, no missing months
 - A time series Y_t is **stationary** if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{i+1}, Y_{i+2}, \dots, Y_{i+T})$ does not depend on i .
 - Stationary property implies that history is relevant. In other words, Stationary requires the future to be like the past (in a probabilistic sense).
 - Auto Regression analysis assumes that Y_t is stationary.

SOME NOTATIONS AND CONCEPTS

- There are four ways to have the time series data for AutoRegression analysis
 - **Lag:** The first lag of Y_t is Y_{t-1} , its j -th lag is Y_{t-j}
 - **Difference:** The first difference of a series, Y_t is its change between period t and $t-1$, that is, $y_t = Y_t - Y_{t-1}$
 - **Log difference:** $y_t = \log(Y_t) - \log(Y_{t-1})$
 - **Percentage:** $y_t = \frac{Y_{t-1}}{Y_t} \times 100$

SOME NOTATIONS AND CONCEPTS

- **Autocorrelation**

- The correlation of a series with its own lagged values is called autocorrelation (also called serial correlation)

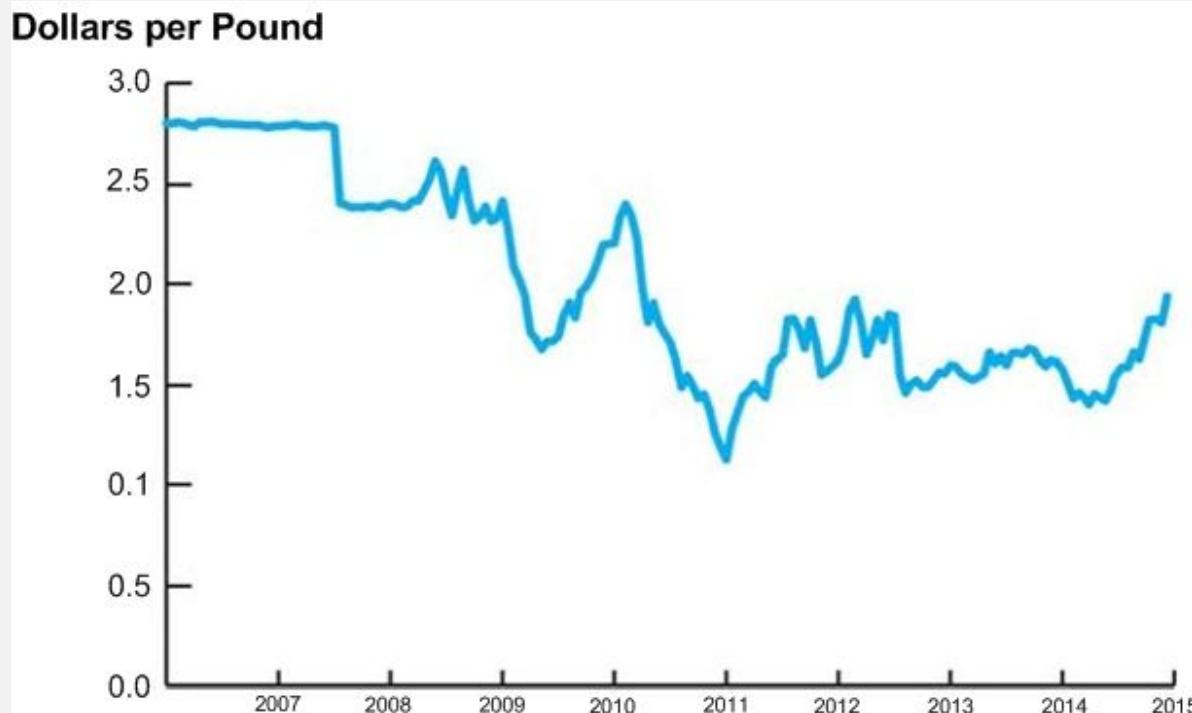
Definition: *j*-th Autocorrelation

The *j*-th autocorrelation, denoted by ρ_j is defined as

$$\rho_j = \frac{COV(Y_t, Y_{t-j})}{\sqrt{\sigma_{Y_t} \sigma_{Y_{t-j}}}}$$

where, $COV(Y_t, Y_{t-j})$ is the ***j*-th autocovariance**

SOME NOTATIONS AND CONCEPTS



- For the given data, say $\rho_1 = 0.84$
 - This implies that the Dollars per Pound is highly serially correlated
- Similarly, we can determine ρ_2, ρ_3, \dots etc., and hence different regression analyses

AUTO-REGRESSION MODEL FOR FORECASTING

- A natural starting point for forecasting model is to use past values of Y , that is, Y_{t-1}, Y_{t-2}, \dots to predict Y_t
- An autoregression is a regression model in which Y_t is regressed against its own lagged values.
- The number of lags used as regressors is called the **order of autoregression**
 - In first order autoregression (denoted as AR(1)), Y_t is regressed against Y_{t-1}
 - In p -th order autoregression (denoted as AR(p)), Y_t is regressed against, $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$

P-TH ORDER AUTO-REGRESSION MODEL

Definition: *p*-th AutoRegression Model

In general, the *p*-th order autoregression model is defined as

$$Y_t = \beta_0 + \sum_{i=1}^p \beta_i Y_{t-i} + \varepsilon_t$$

where, $\beta_0, \beta_1, \dots, \beta_p$ is called autoregression coefficients and ε_t is the noise term or residue and in practice it is assumed to Gaussian white noise

- For example, AR(1) is $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$
- The task in AR analysis is to derive the "best" values for β_i $i = 0, 1, \dots, p$ given a time series Y_t .

COMPUTING AR COEFFICIENTS

- A number of techniques known for computing the AR coefficients
- The most common method is called **Least Squares Method** (LSM)
- The LSM is based upon the **Yule-Walker equations**

$$\begin{bmatrix} 1 & r_1 & r_2 & r_3 & r_4 & \dots & r_{p-2} & r_{p-1} \\ r_1 & 1 & r_1 & r_2 & r_3 & \dots & r_{p-3} & r_{p-2} \\ r_2 & r_1 & 1 & r_1 & r_2 & \dots & r_{p-4} & r_{p-3} \\ r_3 & r_2 & r_1 & 1 & r_2 & \dots & r_{p-5} & r_{p-4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{p-1} & r_{p-2} & r_{p-3} & r_{p-4} & r_{p-5} & \dots & r_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ \vdots \\ \vdots \\ r_{p-1} \\ r_p \end{bmatrix}$$

- Here, r_i ($i = 1, 2, 3, \dots, p-1$) denotes the i -th auto correlation coefficient.
- β_0 can be chosen empirically, usually taken as zero.

REFERENCE

- The detail material related to this lecture can be found in
The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 17

Classification: Naïve Bayes' Classifier

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

AN INTERESTING FACT..



No number before 1,000 contains the letter A.

But there are plenty of E's, I's, O's, U's, and Y's.

TODAY'S DISCUSSION...

- Introduction to Classification
- Classification Techniques
 - Supervised and unsupervised classification
- Formal statement of supervised classification technique
- Bayesian Classifier
 - Principle of Bayesian classifier
 - Bayes' theorem of probability
- Naïve Bayesian Classifier

Identify the objects



INTRODUCTION TO CLASSIFICATION

Example 17.1

- Teacher classify students as A, B, C, D and F based on their marks. The following is one simple classification rule:

| | | |
|---|---|----------|
| Mark ≥ 90 | : | A |
| $90 > \text{Mark} \geq 80$ | : | B |
| $80 > \text{Mark} \geq 70$ | : | C |
| $70 > \text{Mark} \geq 60$ | : | D |
| $60 > \text{Mark}$ | : | F |

Note:

Here, we apply the above rule to a specific data
(in this case a table of marks).

EXAMPLES OF CLASSIFICATION IN DATA ANALYTICS

- **Life Science:** Predicting tumor cells as benign or malignant
- **Security:** Classifying credit card transactions as legitimate or fraudulent
- **Prediction:** Weather, voting, political dynamics, etc.
- **Entertainment:** Categorizing news stories as finance, weather, entertainment, sports, etc.
- **Social media:** Identifying the current trend and future growth

CLASSIFICATION : DEFINITION

- Classification is a form of data analysis to extract models describing important data classes.
- Essentially, it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as classes.
- The term “mutually exhaustive and exclusive” simply means that each object must be assigned to precisely one class
 - That is, never to more than one and never to no class at all.

CLASSIFICATION TECHNIQUES

- Classification consists of assigning a class label to a set of unclassified cases.

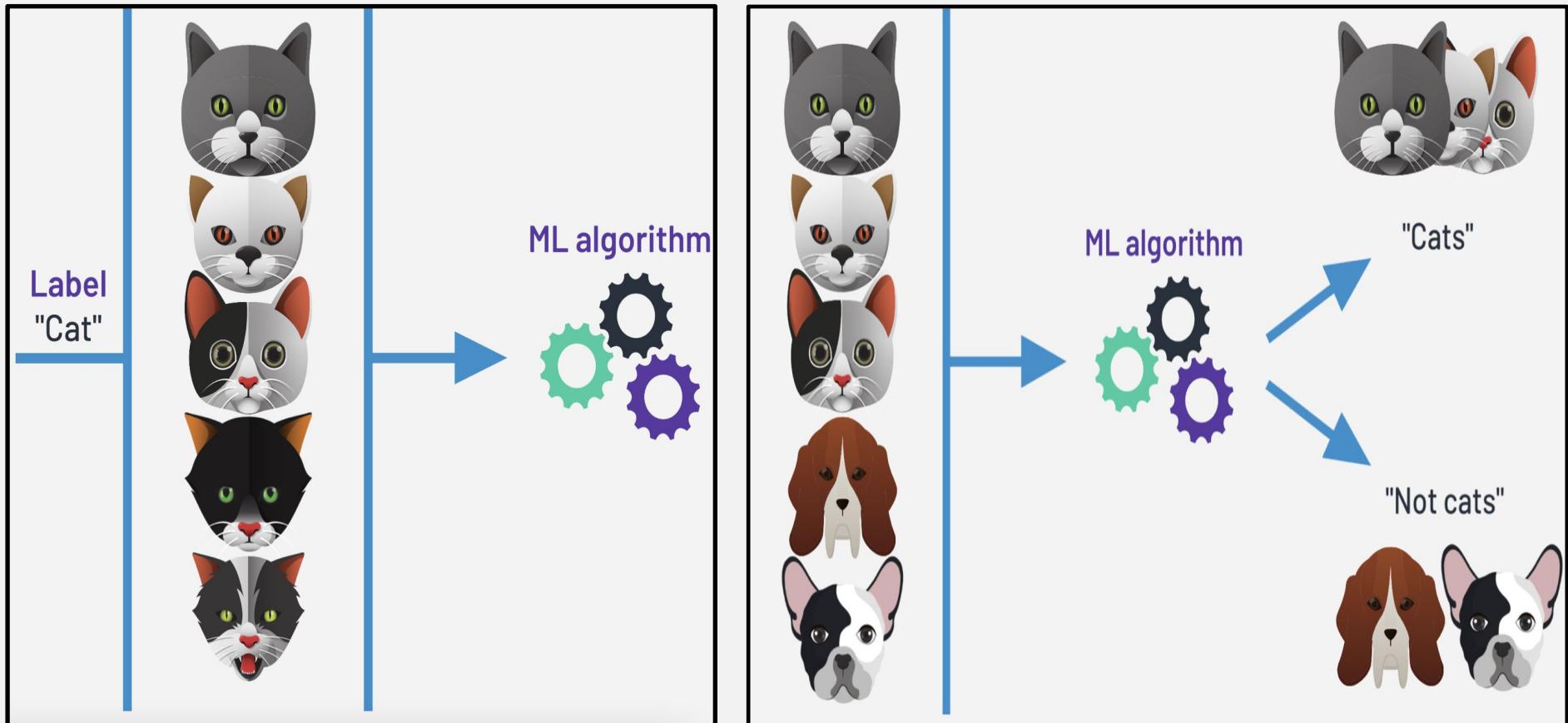
- **Supervised Classification**

- The set of possible classes is known in advance.

- **Unsupervised Classification**

- Set of possible classes is not known. After classification we can try to assign a name to that class.
 - Unsupervised classification is called **clustering**.

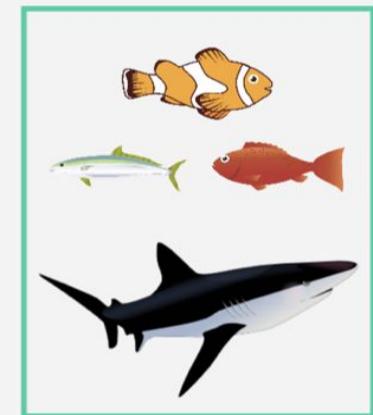
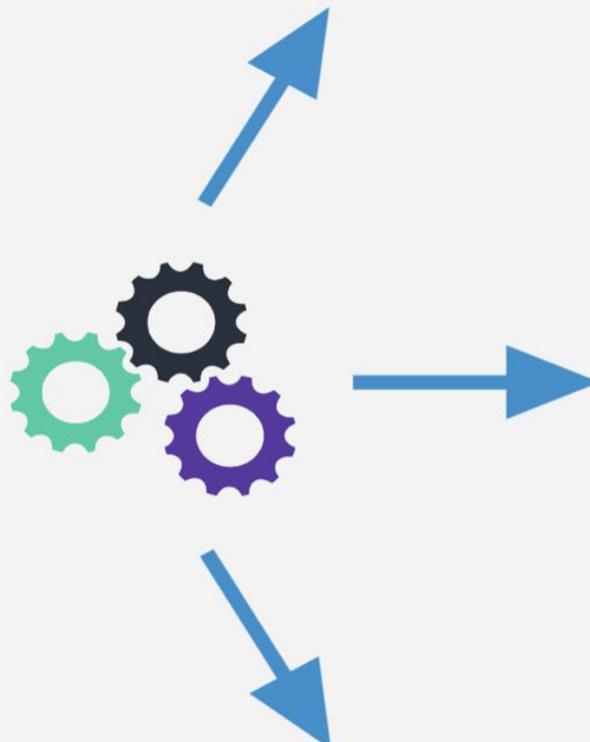
SUPERVISED CLASSIFICATION



Source: <https://www.g2.com/articles/supervised-vs-unsupervised-learning>

UNSUPERVISED CLASSIFICATION

No labels



SUPERVISED CLASSIFICATION TECHNIQUE

- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: Previously unseen records should be assigned a class as accurately as possible.
 - Satisfy the property of “mutually exclusive and exhaustive”

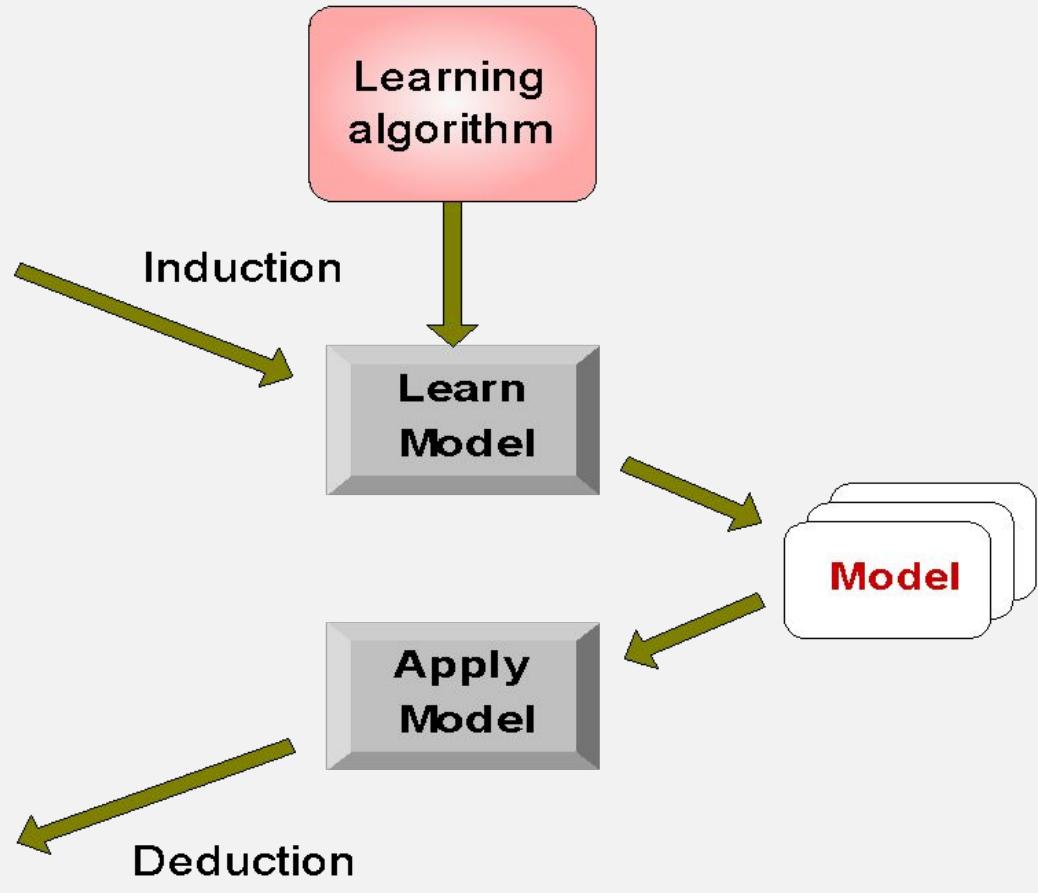
ILLUSTRATING CLASSIFICATION TASKS

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|-----|---------|---------|---------|-------|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

Training Set

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|-----|---------|---------|---------|-------|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

Test Set



CLASSIFICATION PROBLEM

- More precisely, a classification problem can be stated as below:

Definition: Classification Problem

Given a database $D = \{t_1, t_2, \dots, t_m\}$ of tuples and a set of classes $C = \{c_1, c_2, \dots, c_k\}$, the classification problem is to define a mapping $f : D \rightarrow C$,

Where each t_i is assigned to one class.

Note that tuple $t_i \in D$ is defined by a set of attributes $A = \{A_1, A_2, \dots, A_n\}$.

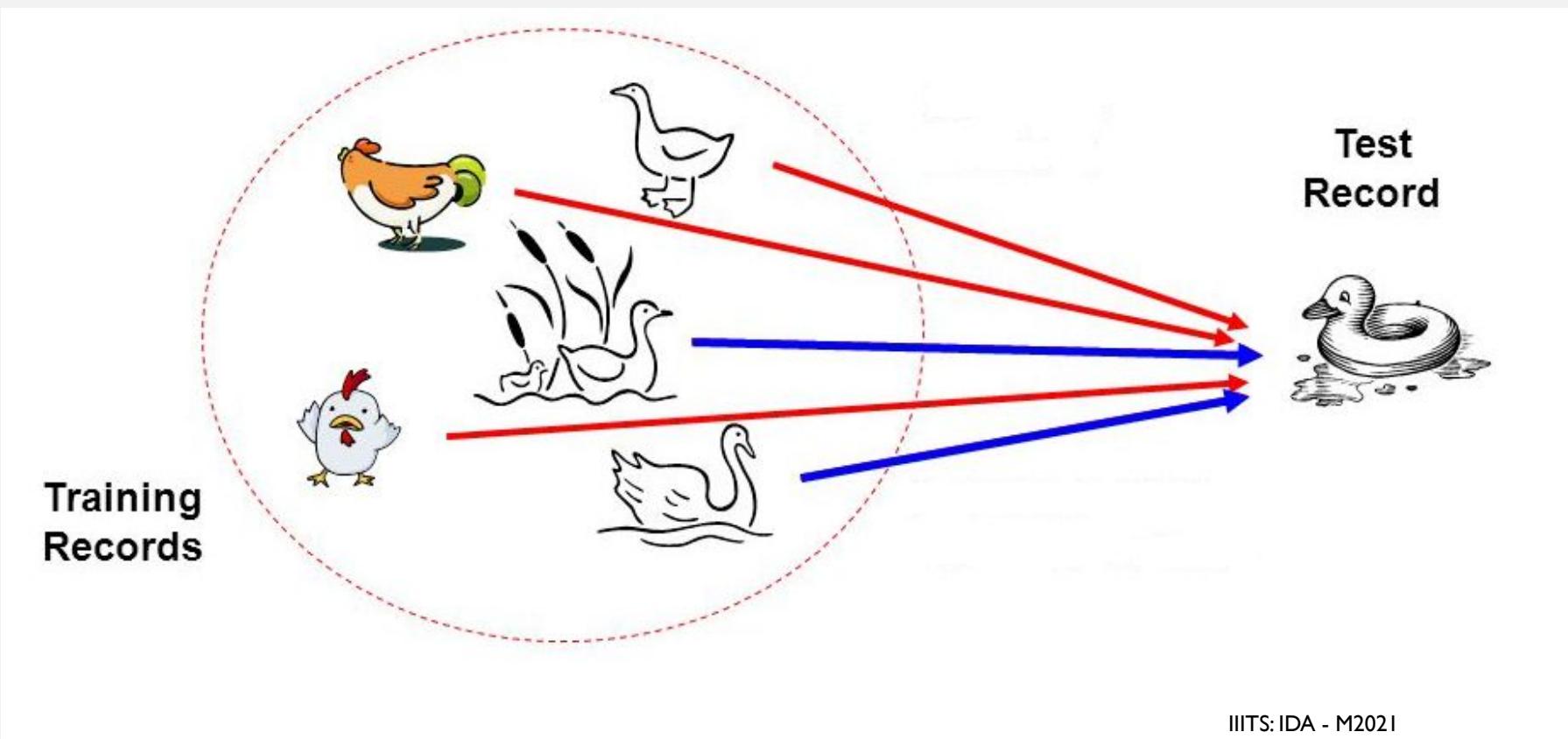
CLASSIFICATION TECHNIQUES

- A number of classification techniques are known, which can be broadly classified into the following categories:
 1. Statistical-Based Methods
 - Bayesian Classifier
 2. Distance-Based Classification
 - K-Nearest Neighbours
 3. Decision Tree-Based Classification
 - ID3, C 4.5, CART
 4. Support vector machine
 5. Classification using Neural Network (ANN)

Bayesian Classifier

BAYESIAN CLASSIFIER

- Principle
 - If it walks like a duck, quacks like a duck, then it is **probably** a duck



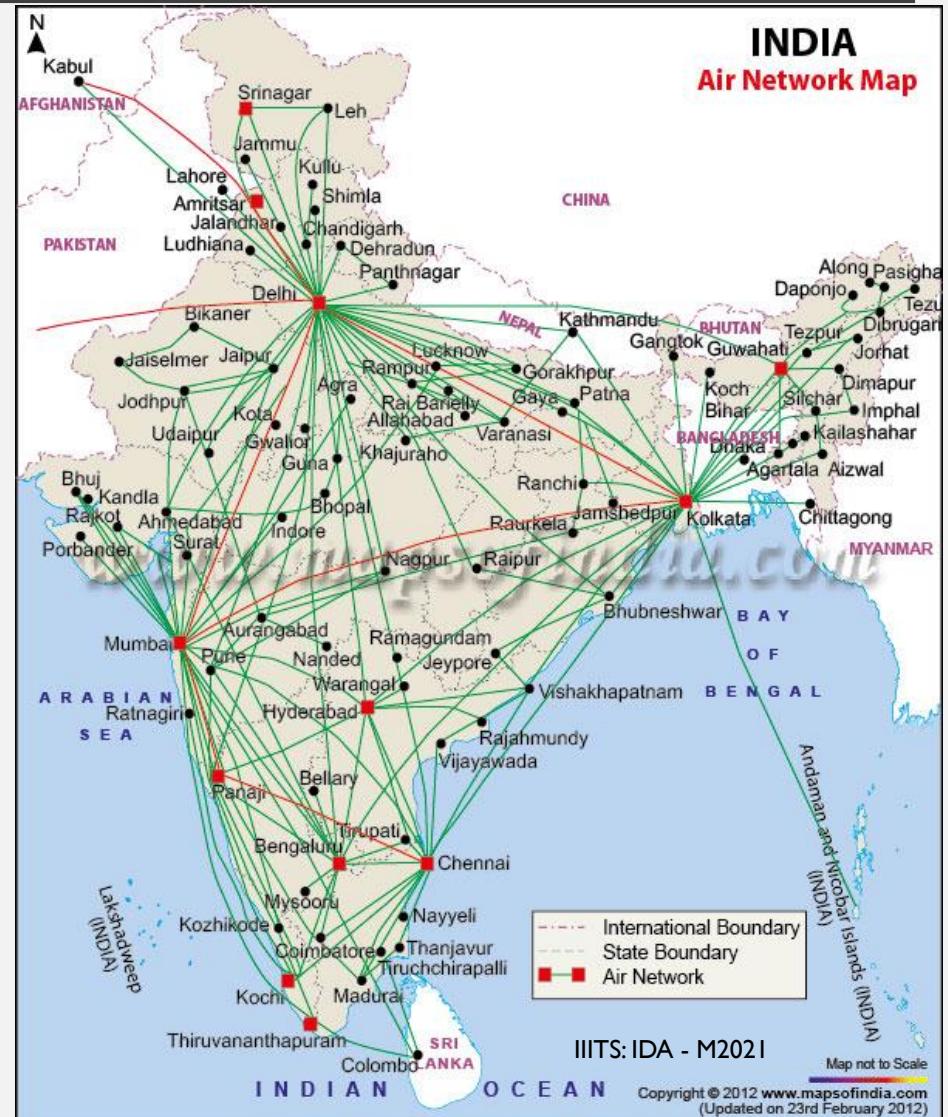
BAYESIAN CLASSIFIER

- A statistical classifier
 - Performs *probabilistic prediction*, i.e., predicts class membership probabilities
- Foundation
 - Based on Bayes' Theorem.
- Assumptions
 1. The classes are mutually exclusive and exhaustive.
 2. The attributes are independent given the class.
- Called “Naïve” classifier because of these assumptions.
 - Empirically proven to be useful.
 - Scales very well.

EXAMPLE: BAYESIAN CLASSIFICATION

- **Example 17.2:** Air Traffic Data

- Let us consider a set of observation recorded in a database
 - Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



AIR-TRAFFIC DATA

| Days | Season | Fog | Rain | Class |
|----------|--------|--------|--------|-----------|
| Weekday | Spring | None | None | On Time |
| Weekday | Winter | None | Slight | On Time |
| Weekday | Winter | None | None | On Time |
| Holiday | Winter | High | Slight | Late |
| Saturday | Summer | Normal | None | On Time |
| Weekday | Autumn | Normal | None | Very Late |
| Holiday | Summer | High | Slight | On Time |
| Sunday | Summer | Normal | None | On Time |
| Weekday | Winter | High | Heavy | Very Late |
| Weekday | Summer | None | Slight | On Time |

Cond. to next slide...

AIR-TRAFFIC DATA

Cond. from previous slide...

| Days | Season | Fog | Rain | Class |
|----------|--------|--------|--------|-----------|
| Saturday | Spring | High | Heavy | Cancelled |
| Weekday | Summer | High | Slight | On Time |
| Weekday | Winter | Normal | None | Late |
| Weekday | Summer | High | None | On Time |
| Weekday | Winter | Normal | Heavy | Very Late |
| Saturday | Autumn | High | Slight | On Time |
| Weekday | Autumn | None | Heavy | On Time |
| Holiday | Spring | Normal | Slight | On Time |
| Weekday | Spring | Normal | None | On Time |
| Weekday | Spring | Normal | Heavy | On Time |

AIR-TRAFFIC DATA

- In this database, there are four attributes

$$A = [\text{Day}, \text{Season}, \text{Fog}, \text{Rain}]$$

with 20 tuples.

- The categories of classes are:

$$C = [\text{On Time}, \text{Late}, \text{Very Late}, \text{Cancelled}]$$

- Given this is the knowledge of data and classes, we are to find most likely classification for any other **unseen instance**, for example:

| | | | | |
|----------|--------|------|------|-----|
| Week Day | Winter | High | None | ??? |
|----------|--------|------|------|-----|

- Classification technique eventually to map this tuple into an accurate class.

BAYESIAN CLASSIFIER

- In many applications, the relationship between the attributes set and the class variable is **non-deterministic**.
 - In other words, a test cannot be classified to a class label with certainty.
 - In such a situation, the classification can be achieved **probabilistically**.
- The Bayesian classifier is an approach for **modelling probabilistic relationships** between the attribute set and the class variable.
- More precisely, Bayesian classifier use **Bayes' Theorem of Probability** for classification.
- Before going to discuss the Bayesian classifier, we should have a quick look at the **Theory of Probability** and then **Bayes' Theorem**.

Bayes' Theorem of Probability

SIMPLE PROBABILITY

Definition: Simple Probability

If there are n elementary events associated with a random experiment and m of n of them are favorable to an event A , then the probability of happening or occurrence of A is

$$P(A) = \frac{m}{n}$$

MUTUALLY EXCLUSIVE EVENTS

- Suppose, A and B are any two events and $P(A)$, $P(B)$ denote the probabilities that the events A and B will occur, respectively.
- **Mutually Exclusive Events:**

- Two events are mutually exclusive, if the occurrence of one precludes the occurrence of the other.

Example: Tossing a coin (two events)

Tossing a ludo cube (Six events)

Kings and Aces are Mutually Exclusive

- Can you give an example, so that two events are not mutually exclusive?

Hint: Hearts and Kings, Weather (sunny, foggy, warm)

INDEPENDENT EVENTS

- **Independent events:** Two events are independent if occurrences of one does not alter the occurrence of other.

Example: Tossing both coin and ludo cube together.

(How many events are here?)

- Can you give an example, where an event is dependent on one or more other events(s)?

Hint: Receiving a message (A) through a communication channel (B) over a computer (C), rain and train.

JOINT PROBABILITY

Definition: Joint Probability

If $P(A)$ and $P(B)$ are the probability of two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive, then $P(A \cap B) = 0$

If A and B are independent events, then $P(A \cap B) = P(A).P(B)$

Thus, for mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

CONDITIONAL PROBABILITY

Definition: Conditional Probability

If events are dependent, then their probability is expressed by conditional probability. The probability that A occurs given that B is denoted by $P(A|B)$.

Suppose, A and B are two events associated with a random experiment. The probability of A under the condition that B has already occurred and $P(B) \neq 0$ is given by

$$\begin{aligned} P(A|B) &= \frac{\text{Number of events in } B \text{ which are favourable to } A}{\text{Number of events in } B} \\ &= \frac{\text{Number of events favourable to } A \cap B}{\text{Number of events favourable to } B} \\ &= \frac{P(A \cap B)}{P(B)} \end{aligned}$$

CONDITIONAL PROBABILITY

Corollary: Conditional Probability

$$P(A \cap B) = P(A) \cdot P(B|A), \quad \text{if } P(A) \neq 0$$

or $P(A \cap B) = P(B) \cdot P(A|B), \quad \text{if } P(B) \neq 0$

For three events A , B and C

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)$$

For n events A_1, A_2, \dots, A_n and if all events are mutually independent to each other

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \dots \cdot P(A_n)$$

Note:

$$P(A|B) = 0 \quad \text{if events are mutually exclusive}$$

$$P(A|B) = P(A) \quad \text{if } A \text{ and } B \text{ are independent}$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \text{ otherwise,}$$

$$P(A \cap B) = P(B \cap A)$$

CONDITIONAL PROBABILITY

- Generalization of Conditional Probability:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)} \\ &= \frac{P(B|A) \cdot P(A)}{P(B)} \quad \because P(A \cap B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B) \end{aligned}$$

By the law of total probability : $P(B) = P[(B \cap A) \cup (B \cap \bar{A})]$, where \bar{A} denotes the compliment of event A. Thus,

$$\begin{aligned} P(A|B) &= \frac{P(B|A) \cdot P(A)}{P[(B \cap A) \cup (B \cap \bar{A})]} \\ &= \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})} \end{aligned}$$

CONDITIONAL PROBABILITY

-

In general,

$$P(A|D) = \frac{P(A) \cdot P(D|A)}{P(A) \cdot P(D|A) + P(B) \cdot P(D|B) + P(C) \cdot P(D|C)}$$

Total Probability

Definition: Total Probability

Let E_1, E_2, \dots, E_n be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 or E_2 or ... E_n , then

$$P(A) = P(E_1).P(A|E_1) + P(E_2).P(A|E_2) + \dots + P(E_n).P(A|E_n)$$

Total Probability: An Example

Example 17.2

A bag contains 4 red and 3 black balls. A second bag contains 2 red and 4 black balls. One bag is selected at random. From the selected bag, one ball is drawn. What is the probability that the ball drawn is red?

This problem can be answered using the concept of Total Probability

E_1 =Selecting bag I

E_2 =Selecting bag II

A = Drawing the red ball

Thus, $P(A) = P(E_1).P(A|E_1) + P(E_2).P(A|E_2)$

where, $P(A|E_1)$ = Probability of drawing red ball when first bag has been chosen

and $P(A|E_2)$ = Probability of drawing red ball when second bag has been chosen

Reverse Probability

Example 17.3:

A bag (Bag I) contains 4 red and 3 black balls. A second bag (Bag II) contains 2 red and 4 black balls. You have chosen one ball at random. It is found as red ball. What is the probability that the ball is chosen from Bag I?

Here,

E_1 = Selecting bag I

E_2 = Selecting bag II

A = Drawing the red ball

We are to determine $P(E_1|A)$. Such a problem can be solved using Bayes' theorem of probability.

BAYES' THEOREM

Theorem: Bayes' Theorem

Let E_1, E_2, \dots, E_n be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 or E_2 or ... E_n , then

$$P(E_i|A) = \frac{P(E_i) \cdot P(A|E_i)}{\sum_{i=1}^n P(E_i) \cdot P(A|E_i)}$$

PRIOR AND POSTERIOR PROBABILITIES

- $P(A)$ and $P(B)$ are called prior probabilities
- $P(A|B)$, $P(B|A)$ are called posterior probabilities

Example: Prior versus Posterior Probabilities

- This table shows that the event Y has two outcomes namely A and B , which is dependent on another event X with various outcomes like x_1, x_2 and x_3 .
- **Case1:** Suppose, we don't have any information of the event A . Then, from the given sample space, we can calculate $P(Y = A) = \frac{5}{10} = 0.5$
- **Case2:** Now, suppose, we want to calculate $P(X = x_2 | Y = A) = \frac{2}{5} = 0.4$

The later is the conditional or posterior probability, where as the former is the prior probability.

| X | Y |
|-----|-----|
| | A |
| | A |
| | B |
| | A |
| | B |
| | A |
| | B |
| | B |
| | B |
| | A |

NAÏVE BAYESIAN CLASSIFIER

- Suppose, Y is a class variable and $X = \{X_1, X_2, \dots, X_n\}$ is a set of attributes, with instance of Y .

| INPUT (X) | CLASS(Y) |
|-----------|----------|
| ... | |
| ... | ... |
| | |
| ... | ... |

- The classification problem, then can be expressed as the class-conditional probability

$$P(Y = y_i | (X_1 = x_1) \text{ AND } (X_2 = x_2) \text{ AND } \dots \text{ AND } (X_n = x_n))$$

NAÏVE BAYESIAN CLASSIFIER

- Naïve Bayesian classifier calculate this posterior probability using Bayes' theorem, which is as follows.
- From Bayes' theorem on conditional probability, we have

$$\begin{aligned} P(Y|X) &= \frac{P(X|Y) \cdot P(Y)}{P(X)} \\ &= \frac{P(X|Y) \cdot P(Y)}{P(X|Y = y_1) \cdot P(Y = y_1) + \cdots + P(X|Y = y_k) \cdot P(Y = y_k)} \end{aligned}$$

where,

$$P(X) = \sum_{i=1}^k P(X|Y = y_i) \cdot P(Y = y_i)$$

Note:

- $P(X)$ is called the evidence (also the total probability) and it is a constant.
- The probability $P(Y|X)$ (also called class conditional probability) is therefore proportional to $P(X|Y) \cdot P(Y)$.
- Thus, $P(Y|X)$ can be taken as a measure of Y given that X .

$$P(Y|X) \approx P(X|Y) \cdot P(Y)$$

NAÏVE BAYESIAN CLASSIFIER

- Suppose, for a given instance of X (say $x = (X_1 = x_1)$ and $(X_n = x_n)$).
- There are any two class conditional probabilities namely $P(Y = y_i | X=x)$ and $P(Y = y_j | X=x)$.
- If $P(Y = y_i | X=x) > P(Y = y_j | X=x)$, then we say that y_i is more stronger than y_j for the instance $X = x$.
- The strongest y_i is the classification for the instance $X = x$.

NAÏVE BAYESIAN CLASSIFIER

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

| | | Class | | | |
|-----------|----------|---------------|---------------------|--------------|-----------|
| Attribute | | On Time | Late | Very Late | Cancelled |
| Day | Weekday | $9/14 = 0.64$ | $\frac{1}{2} = 0.5$ | $3/3 = 1$ | $0/1 = 0$ |
| | Saturday | $2/14 = 0.14$ | $\frac{1}{2} = 0.5$ | $0/3 = 0$ | $1/1 = 1$ |
| | Sunday | $1/14 = 0.07$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Holiday | $2/14 = 0.14$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| Season | Spring | $4/14 = 0.29$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Summer | $6/14 = 0.43$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Autumn | $2/14 = 0.14$ | $0/2 = 0$ | $1/3 = 0.33$ | $0/1 = 0$ |
| | Winter | $2/14 = 0.14$ | $2/2 = 1$ | $2/3 = 0.67$ | $0/1 = 0$ |

NAÏVE BAYESIAN CLASSIFIER

| | | Class | | | |
|-------------------|--------|----------------|---------------|---------------|---------------|
| Attribute | | On Time | Late | Very Late | Cancelled |
| Fog | None | $5/14 = 0.36$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | High | $4/14 = 0.29$ | $1/2 = 0.5$ | $1/3 = 0.33$ | $1/1 = 1$ |
| | Normal | $5/14 = 0.36$ | $1/2 = 0.5$ | $2/3 = 0.67$ | $0/1 = 0$ |
| Rain | None | $5/14 = 0.36$ | $1/2 = 0.5$ | $1/3 = 0.33$ | $0/1 = 0$ |
| | Slight | $8/14 = 0.57$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Heavy | $1/14 = 0.07$ | $1/2 = 0.5$ | $2/3 = 0.67$ | $1/1 = 1$ |
| Prior Probability | | $14/20 = 0.70$ | $2/20 = 0.10$ | $3/20 = 0.15$ | $1/20 = 0.05$ |

NAÏVE BAYESIAN CLASSIFIER

Instance:

| | | | | |
|----------|--------|------|-------|-----|
| Week Day | Winter | High | Heavy | ??? |
|----------|--------|------|-------|-----|

Case1: Class = On Time : $0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07 = 0.0013$

Case2: Class = Late : $0.10 \times 0.50 \times 1.0 \times 0.50 \times 0.50 = 0.0125$

Case3: Class = Very Late : $0.15 \times 1.0 \times 0.67 \times 0.33 \times 0.67 = 0.0222$

Case4: Class = Cancelled : $0.05 \times 0.0 \times 0.0 \times 1.0 \times 1.0 = 0.0000$

Case3 is the strongest; Hence correct classification is **Very Late**

NAÏVE BAYESIAN CLASSIFIER

- **Algorithm: Naïve Bayesian Classification**

Input: Given a set of k mutually exclusive and exhaustive classes $C = \{c_1, c_2, \dots, c_k\}$, which have prior probabilities $P(C_1), P(C_2), \dots, P(C_k)$.

There are n -attribute set $A = \{A_1, A_2, \dots, A_n\}$, which for a given instance have values $A_1 = a_1, A_2 = a_2, \dots, A_n = a_n$

Step: For each $c_i \in C$, calculate the class condition probabilities, $i = 1, 2, \dots, k$

$$p_i = P(C_i) \times \prod_{j=1}^n P(A_j = a_j | C_i)$$

$$p_x = \max\{p_1, p_2, \dots, p_k\}$$

Output: C_x is the classification

NAÏVE BAYESIAN CLASSIFIER

Pros and Cons

- The Naïve Bayes' approach is a very popular one, which often works well.
- However, it has a number of potential problems
 - It relies on all attributes being **categorical**.
 - If the data is **less**, then it **estimates poorly**.

NAÏVE BAYESIAN CLASSIFIER

Approach to overcome the limitations in Naïve Bayesian Classification

- Estimating the posterior probabilities for continuous attributes
 - In real life situation, all attributes are not necessarily be categorical, In fact, there is a mix of both categorical and continuous attributes.
 - In the following, we discuss the schemes to deal with continuous attributes in Bayesian classifier.
 1. We can discretize each continuous attributes and then replace the continuous values with its corresponding discrete intervals.
 2. We can assume a certain form of probability distribution for the continuous variable and estimate the parameters of the distribution using the training data. A Gaussian distribution is usually chosen to represent the posterior probabilities for continuous attributes. A general form of Gaussian distribution will look like

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

where, μ and σ^2 denote mean and variance, respectively.

NAÏVE BAYESIAN CLASSIFIER

- For each class C_i , the posterior probabilities for attribute A_j (it is the numeric attribute) can be calculated following Gaussian normal distribution as follows.

$$P(A_j = a_j | C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(a_j - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Here, the parameter μ_{ij} can be calculated based on the sample mean of attribute value of A_j for the training records that belong to the class C_i .

Similarly, σ_{ij}^2 can be estimated from the calculation of variance of such training records.

NAÏVE BAYESIAN CLASSIFIER

M-estimate of Conditional Probability

- The M-estimation is to deal with the potential problem of Naïve Bayesian Classifier when training data size is too poor.
 - If the posterior probability for one of the attribute is zero, then the overall class-conditional probability for the class vanishes.
 - In other words, if training data do not cover many of the attribute values, then we may not be able to classify some of the test records.
- This problem can be addressed by using the **M-estimate approach**.

M-ESTIMATE APPROACH

- M-estimate approach can be stated as follows

$$P(A_j = a_j | C_i) = \frac{n_{c_i} + mp}{n + m}$$

where, n = total number of instances from class C_i

n_{c_i} = number of training examples from class C_i that take the value $A_j = a_j$

m = it is a parameter known as the equivalent sample size, and

p = is a user specified parameter.

Note:

If $n = 0$, that is, if there is no training set available, then $P(a_i | C_i) = p$,
so, this is a different value, in absence of sample value.

A PRACTICE EXAMPLE

Example 8.4

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Data instance

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = fair)

| age | income | student | credit rating | com |
|---------|--------|---------|---------------|-----|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31...40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31...40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31...40 | medium | no | excellent | yes |
| 31...40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |

A PRACTICE EXAMPLE

- $P(C_i)$:
 $P(\text{buys_computer} = \text{"yes"}) = 9/14 = 0.643$
 $P(\text{buys_computer} = \text{"no"}) = 5/14 = 0.357$
- Compute $P(X|C_i)$ for each class
 - $P(\text{age} = \text{"<=30"} | \text{buys_computer} = \text{"yes"}) = 2/9 = 0.222$
 - $P(\text{age} = \text{"<= 30"} | \text{buys_computer} = \text{"no"}) = 3/5 = 0.6$
 - $P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"yes"}) = 4/9 = 0.444$
 - $P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$
 - $P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$
 - $P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"no"}) = 1/5 = 0.2$
 - $P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$
 - $P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$
- $\mathbf{X} = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit_rating} = \text{fair})$

$$\mathbf{P(X|C_i)} : P(X|\text{buys_computer} = \text{"yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$
$$P(X|\text{buys_computer} = \text{"no"}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$$

$$\mathbf{P(X|C_i)*P(C_i)} : P(X|\text{buys_computer} = \text{"yes"}) * P(\text{buys_computer} = \text{"yes"}) = 0.028$$
$$P(X|\text{buys_computer} = \text{"no"}) * P(\text{buys_computer} = \text{"no"}) = 0.007$$

Therefore, \mathbf{X} belongs to class ("buys_computer = yes")

REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 18

Naïve Bayes' Classifier: Example

Dr. Sreeja S R

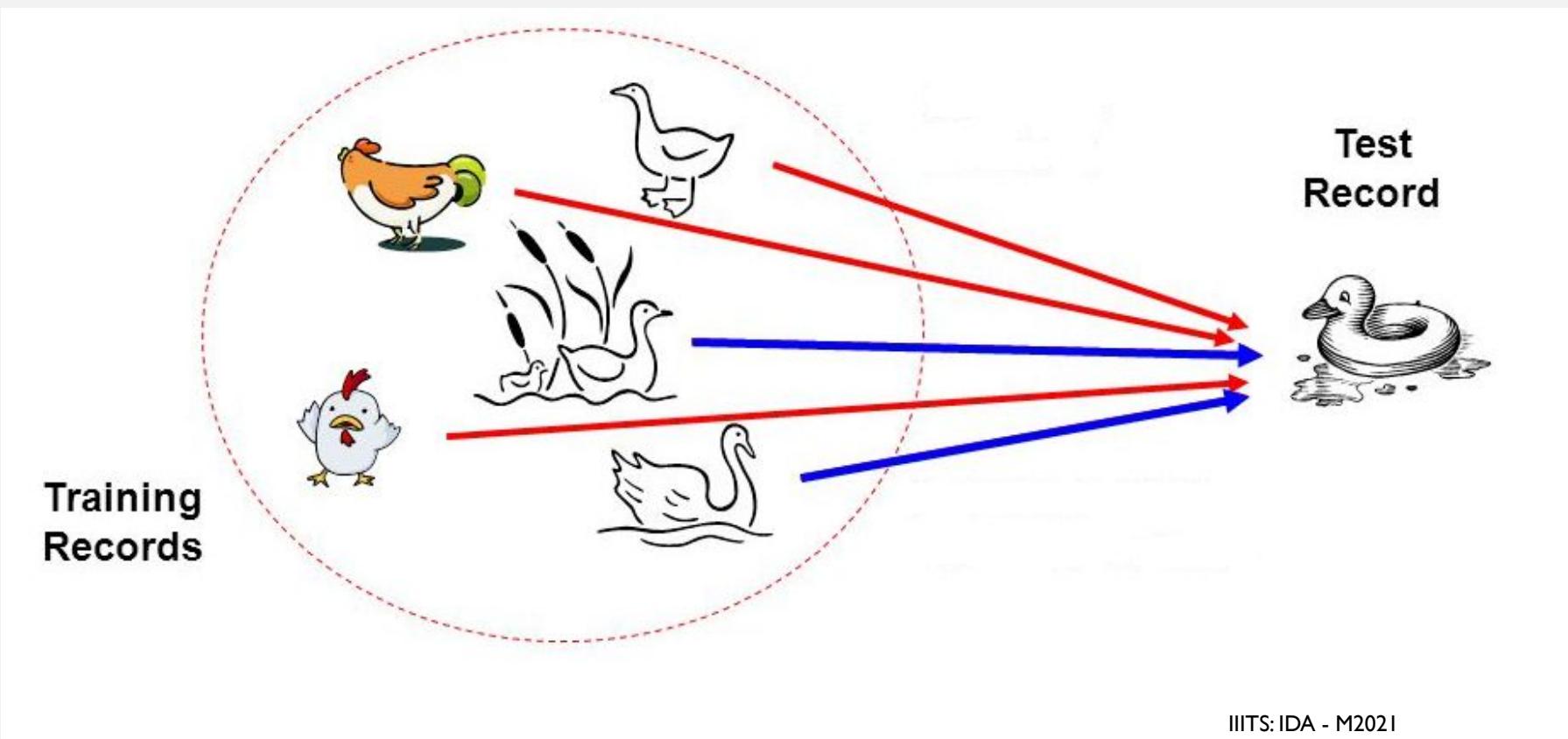
Assistant Professor

Indian Institute of Information Technology
IIIT Sri City

Bayesian Classifier

BAYESIAN CLASSIFIER

- Principle
 - If it walks like a duck, quacks like a duck, then it is **probably** a duck



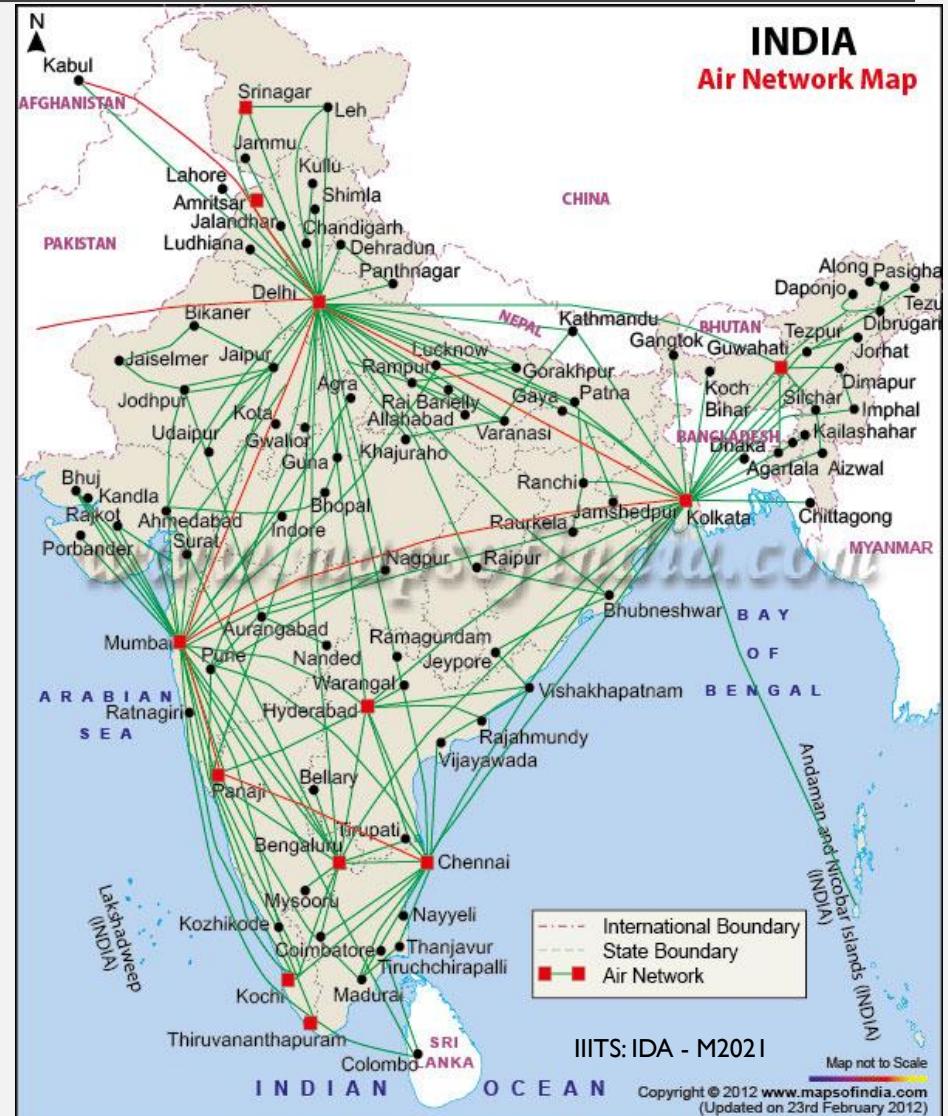
BAYESIAN CLASSIFIER

- A statistical classifier
 - Performs *probabilistic prediction*, i.e., predicts class membership probabilities
- Foundation
 - Based on Bayes' Theorem.
- Assumptions
 1. The classes are mutually exclusive and exhaustive.
 2. The attributes are independent given the class.
- Called “Naïve” classifier because of these assumptions.
 - Empirically proven to be useful.
 - Scales very well.

EXAMPLE: BAYESIAN CLASSIFICATION

- **Example 8.2: Air Traffic Data**

- Let us consider a set observation recorded in a database
 - Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



AIR-TRAFFIC DATA

| Days | Season | Fog | Rain | Class |
|----------|--------|--------|--------|-----------|
| Weekday | Spring | None | None | On Time |
| Weekday | Winter | None | Slight | On Time |
| Weekday | Winter | None | None | On Time |
| Holiday | Winter | High | Slight | Late |
| Saturday | Summer | Normal | None | On Time |
| Weekday | Autumn | Normal | None | Very Late |
| Holiday | Summer | High | Slight | On Time |
| Sunday | Summer | Normal | None | On Time |
| Weekday | Winter | High | Heavy | Very Late |
| Weekday | Summer | None | Slight | On Time |

Cond. to next slide...

AIR-TRAFFIC DATA

Cond. from previous slide...

| Days | Season | Fog | Rain | Class |
|----------|--------|--------|--------|-----------|
| Saturday | Spring | High | Heavy | Cancelled |
| Weekday | Summer | High | Slight | On Time |
| Weekday | Winter | Normal | None | Late |
| Weekday | Summer | High | None | On Time |
| Weekday | Winter | Normal | Heavy | Very Late |
| Saturday | Autumn | High | Slight | On Time |
| Weekday | Autumn | None | Heavy | On Time |
| Holiday | Spring | Normal | Slight | On Time |
| Weekday | Spring | Normal | None | On Time |
| Weekday | Spring | Normal | Heavy | On Time |

AIR-TRAFFIC DATA

- In this database, there are four attributes

$$A = [\text{Day}, \text{Season}, \text{Fog}, \text{Rain}]$$

with 20 tuples.

- The categories of classes are:

$$C = [\text{On Time}, \text{Late}, \text{Very Late}, \text{Cancelled}]$$

- Given this is the knowledge of data and classes, we are to find most likely classification for any other **unseen instance**, for example:

| | | | | |
|----------|--------|------|------|-----|
| Week Day | Winter | High | None | ??? |
|----------|--------|------|------|-----|

- Classification technique eventually to map this tuple into an accurate class.

Bayes' Theorem of Probability

BAYES' THEOREM

Theorem 8.4: Bayes' Theorem

Let E_1, E_2, \dots, E_n be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 or E_2 or ... E_n , then

$$P(E_i|A) = \frac{P(E_i) \cdot P(A|E_i)}{\sum_{i=1}^n P(E_i) \cdot P(A|E_i)}$$

PRIOR AND POSTERIOR PROBABILITIES

- $P(A)$ and $P(B)$ are called prior probabilities
- $P(A|B)$, $P(B|A)$ are called posterior probabilities

Example 8.6: Prior versus Posterior Probabilities

- This table shows that the event Y has two outcomes namely A and B , which is dependent on another event X with various outcomes like x_1, x_2 and x_3 .
- **Case1:** Suppose, we don't have any information of the event A . Then, from the given sample space, we can calculate $P(Y = A) = \frac{5}{10} = 0.5$
- **Case2:** Now, suppose, we want to calculate $P(X = x_2 | Y = A) = \frac{2}{5} = 0.4$.

The later is the conditional or posterior probability, where as the former is the prior probability.

| X | Y |
|-----|-----|
| | A |
| | A |
| | B |
| | A |
| | B |
| | A |
| | B |
| | B |
| | B |
| | A |

NAÏVE BAYESIAN CLASSIFIER

- Suppose, Y is a class variable and $X = \{X_1, X_2, \dots, X_n\}$ is a set of attributes, with instance of Y .

| INPUT (X) | CLASS(Y) |
|-----------|----------|
| ... | |
| ... | ... |
| | |
| ... | ... |

- The classification problem, then can be expressed as the class-conditional probability

$$P(Y = y_i | (X_1 = x_1) \text{ AND } (X_2 = x_2) \text{ AND } \dots \text{ AND } (X_n = x_n))$$

NAÏVE BAYESIAN CLASSIFIER

- Naïve Bayesian classifier calculate this posterior probability using Bayes' theorem, which is as follows.
- From Bayes' theorem on conditional probability, we have

$$\begin{aligned} P(Y|X) &= \frac{P(X|Y) \cdot P(Y)}{P(X)} \\ &= \frac{P(X|Y) \cdot P(Y)}{P(X|Y = y_1) \cdot P(Y = y_1) + \dots + P(X|Y = y_k) \cdot P(Y = y_k)} \end{aligned}$$

where,

$$P(X) = \sum_{i=1}^k P(X|Y = y_i) \cdot P(Y = y_i)$$

Note:

- $P(X)$ is called the evidence (also the total probability) and it is a constant.
- The probability $P(Y|X)$ (also called class conditional probability) is therefore proportional to $P(X|Y) \cdot P(Y)$.
- Thus, $P(Y|X)$ can be taken as a measure of Y given that X .

$$P(Y|X) \approx P(X|Y) \cdot P(Y)$$

NAÏVE BAYESIAN CLASSIFIER

- Suppose, for a given instance of X (say $x = (X_1 = x_1)$ and $(X_n = x_n)$).
- There are any two class conditional probabilities namely $P(Y = y_i | X=x)$ and $P(Y = y_j | X=x)$.
- If $P(Y = y_i | X=x) > P(Y = y_j | X=x)$, then we say that y_i is more stronger than y_j for the instance $X = x$.
- The strongest y_i is the classification for the instance $X = x$.

NAÏVE BAYESIAN CLASSIFIER

- **Algorithm: Naïve Bayesian Classification**

Input: Given a set of k mutually exclusive and exhaustive classes $C = \{c_1, c_2, \dots, c_k\}$, which have prior probabilities $P(C_1), P(C_2), \dots, P(C_k)$.

There are n -attribute set $A = \{A_1, A_2, \dots, A_n\}$, which for a given instance have values $A_1 = a_1, A_2 = a_2, \dots, A_n = a_n$

Step: For each $c_i \in C$, calculate the class condition probabilities, $i = 1, 2, \dots, k$

$$p_i = P(C_i) \times \prod_{j=1}^n P(A_j = a_j | C_i)$$

$$p_x = \max\{p_1, p_2, \dots, p_k\}$$

Output: C_x is the classification

Note: $\sum p_i \neq 1$, because they are not probabilities rather proportion values (to posterior probabilities)

NAÏVE BAYESIAN CLASSIFIER

- Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

$P(A_j = a_j | C_i)$ is the likelihood. Here, for the given example, $P(\text{Day} = \text{weekday} | C_1 = \text{On time})$

| | | Class | | | |
|-----------|----------|---------------|---------------------|--------------|-----------|
| Attribute | | On Time | Late | Very Late | Cancelled |
| Day | Weekday | $9/14 = 0.64$ | $\frac{1}{2} = 0.5$ | $3/3 = 1$ | $0/1 = 0$ |
| | Saturday | $2/14 = 0.14$ | $\frac{1}{2} = 0.5$ | $0/3 = 0$ | $1/1 = 1$ |
| | Sunday | $1/14 = 0.07$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Holiday | $2/14 = 0.14$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| Season | Spring | $4/14 = 0.29$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Summer | $6/14 = 0.43$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Autumn | $2/14 = 0.14$ | $0/2 = 0$ | $1/3 = 0.33$ | $0/1 = 0$ |
| | Winter | $2/14 = 0.14$ | $2/2 = 1$ | $2/3 = 0.67$ | $0/1 = 0$ |

NAÏVE BAYESIAN CLASSIFIER

| | | Class | | | |
|-------------------|--------|----------------|---------------|---------------|---------------|
| Attribute | | On Time | Late | Very Late | Cancelled |
| Fo g | None | $5/14 = 0.36$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | High | $4/14 = 0.29$ | $1/2 = 0.5$ | $1/3 = 0.33$ | $1/1 = 1$ |
| | Normal | $5/14 = 0.36$ | $1/2 = 0.5$ | $2/3 = 0.67$ | $0/1 = 0$ |
| Ra in | None | $5/14 = 0.36$ | $1/2 = 0.5$ | $1/3 = 0.33$ | $0/1 = 0$ |
| | Slight | $8/14 = 0.57$ | $0/2 = 0$ | $0/3 = 0$ | $0/1 = 0$ |
| | Heavy | $1/14 = 0.07$ | $1/2 = 0.5$ | $2/3 = 0.67$ | $1/1 = 1$ |
| Prior Probability | | $14/20 = 0.70$ | $2/20 = 0.10$ | $3/20 = 0.15$ | $1/20 = 0.05$ |

NAÏVE BAYESIAN CLASSIFIER

Instance:

| | | | | |
|----------|--------|------|-------|-----|
| Week Day | Winter | High | Heavy | ??? |
|----------|--------|------|-------|-----|

Case1: Class = On Time : $0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07 = 0.0013$

Case2: Class = Late : $0.10 \times 0.50 \times 1.0 \times 0.50 \times 0.50 = 0.0125$

Case3: Class = Very Late : $0.15 \times 1.0 \times 0.67 \times 0.33 \times 0.67 = 0.0222$

Case4: Class = Cancelled : $0.05 \times 0.0 \times 0.0 \times 1.0 \times 1.0 = 0.0000$

Case3 is the strongest; Hence correct classification is **Very Late**

NAÏVE BAYESIAN CLASSIFIER

Pros and Cons

- The Naïve Bayes' approach is a very popular one, which often works well.
- However, it has a number of potential problems
 - It relies on all attributes being **categorical**.
 - If the data is **less**, then it **estimates poorly**.

NAÏVE BAYESIAN CLASSIFIER

Approach to overcome the limitations in Naïve Bayesian Classification

- Estimating the posterior probabilities for continuous attributes
 - In real life situation, all attributes are not necessarily be categorical, In fact, there is a mix of both categorical and continuous attributes.
 - In the following, we discuss the schemes to deal with continuous attributes in Bayesian classifier.
 1. We can discretize each continuous attributes and then replace the continuous values with its corresponding discrete intervals.
 2. We can assume a certain form of probability distribution for the continuous variable and estimate the parameters of the distribution using the training data. A Gaussian distribution is usually chosen to represent the posterior probabilities for continuous attributes. A general form of Gaussian distribution will look like

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

where, μ and σ^2 denote mean and variance, respectively.

NAÏVE BAYESIAN CLASSIFIER

-

For each class C_i , the posterior probabilities for attribute A_j (it is the numeric attribute) can be calculated following Gaussian normal distribution as follows.

$$P(A_j = a_j | C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(a_j - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Here, the parameter μ_{ij} can be calculated based on the sample mean of attribute value of A_j for the training records that belong to the class C_i .

Similarly, σ_{ij}^2 can be estimated from the calculation of variance of such training records.

NAÏVE BAYESIAN CLASSIFIER

M-estimate of Conditional Probability

- The M-estimation is to deal with the potential problem of Naïve Bayesian Classifier when training data size is too poor.
 - If the posterior probability for one of the attribute is zero, then the overall class-conditional probability for the class vanishes.
 - In other words, if training data do not cover many of the attribute values, then we may not be able to classify some of the test records.
- This problem can be addressed by using the **M-estimate approach**.

M-ESTIMATE APPROACH

- M-estimate approach can be stated as follows

$$P(A_j = a_j | C_i) = \frac{n_{c_i} + mp}{n + m}$$

where, n = total number of instances from class C_i

n_{c_i} = number of training examples from class C_i that take the value $A_j = a_j$

m = it is a parameter known as the equivalent sample size, and

p = is a user specified parameter.

Note:

If $n = 0$, that is, if there is no training set available, then $P(a_i | C_i) = p$,
so, this is a different value, in absence of sample value.

A PRACTICE EXAMPLE

Example 17.4

Class:

C1:buys_computer = 'yes'
C2:buys_computer = 'no'

Data instance

X = (age <=30,
Income = medium,
Student = yes
Credit_rating = fair)

A PRACTICE EXAMPLE

Example 17.4

| age | income | student | credit rating | com |
|---------|--------|---------|---------------|-----|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31...40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31...40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31...40 | medium | no | excellent | yes |
| 31...40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |

A PRACTICE EXAMPLE

- $P(C_i)$:
 $P(\text{buys_computer} = \text{"yes"}) = 9/14 = 0.643$
 $P(\text{buys_computer} = \text{"no"}) = 5/14 = 0.357$
- Compute $P(X|C_i)$ for each class
 - $P(\text{age} = \text{"<=30"} | \text{buys_computer} = \text{"yes"}) = 2/9 = 0.222$
 - $P(\text{age} = \text{"<= 30"} | \text{buys_computer} = \text{"no"}) = 3/5 = 0.6$
 - $P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"yes"}) = 4/9 = 0.444$
 - $P(\text{income} = \text{"medium"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$
 - $P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$
 - $P(\text{student} = \text{"yes"} | \text{buys_computer} = \text{"no"}) = 1/5 = 0.2$
 - $P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"yes"}) = 6/9 = 0.667$
 - $P(\text{credit_rating} = \text{"fair"} | \text{buys_computer} = \text{"no"}) = 2/5 = 0.4$
- $\mathbf{X} = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit_rating} = \text{fair})$

$$\mathbf{P(X|C_i)} : P(X|\text{buys_computer} = \text{"yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$
$$P(X|\text{buys_computer} = \text{"no"}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$$

$$\mathbf{P(X|C_i)*P(C_i)} : P(X|\text{buys_computer} = \text{"yes"}) * P(\text{buys_computer} = \text{"yes"}) = 0.028$$
$$P(X|\text{buys_computer} = \text{"no"}) * P(\text{buys_computer} = \text{"no"}) = 0.007$$

Therefore, \mathbf{X} belongs to class ("buys_computer = yes")

REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?



INTRODUCTION TO DATA ANALYTICS

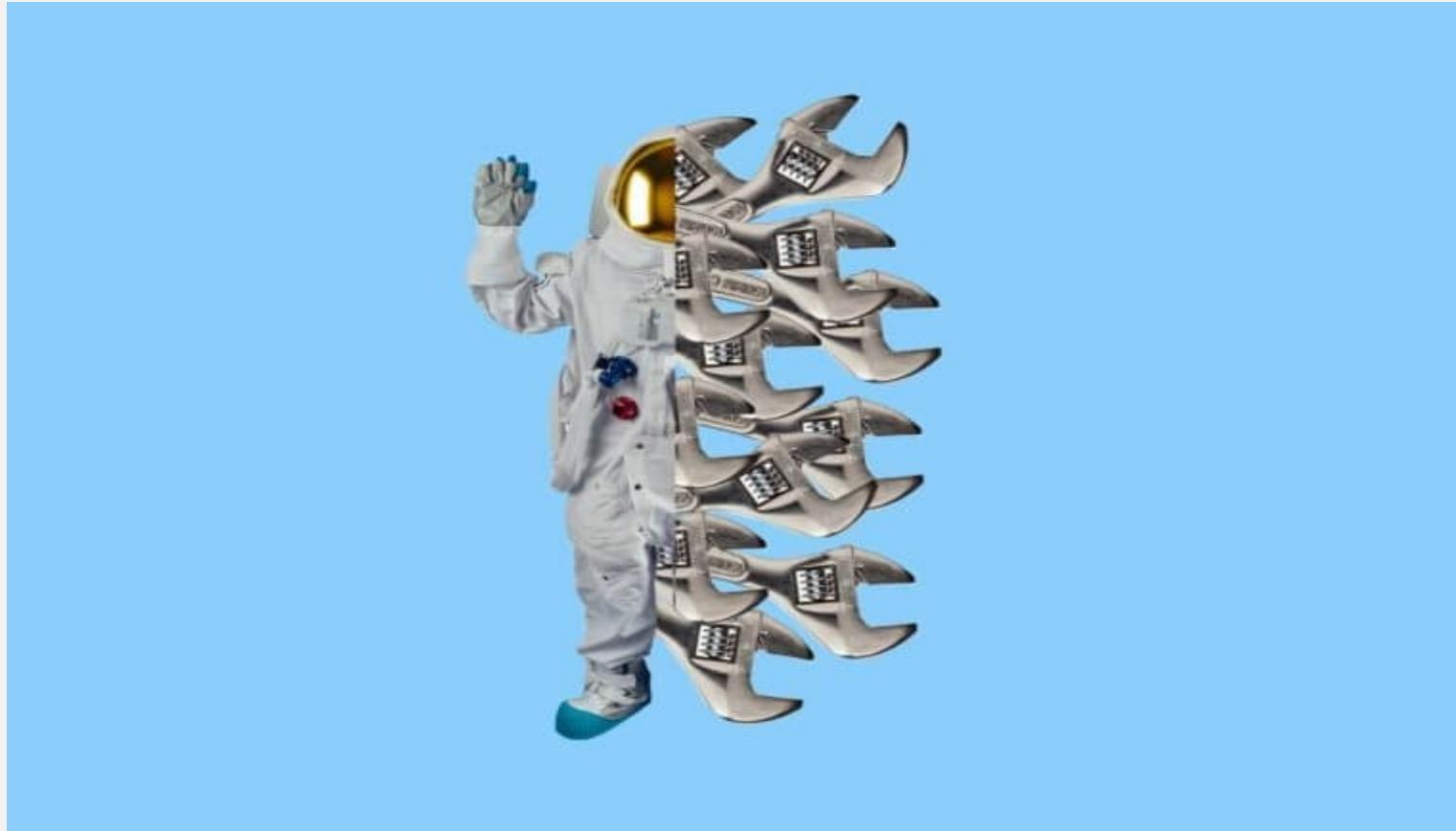
Class # 19

Classification: Decision Tree Induction

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**



Thanks to 3D printing, NASA can basically “email” tools to astronauts.

Getting new equipment to the Space Station used to take months or years, but the new technology means the tools are ready within hours.

THIS PRESENTATION SLIDES INCLUDES...

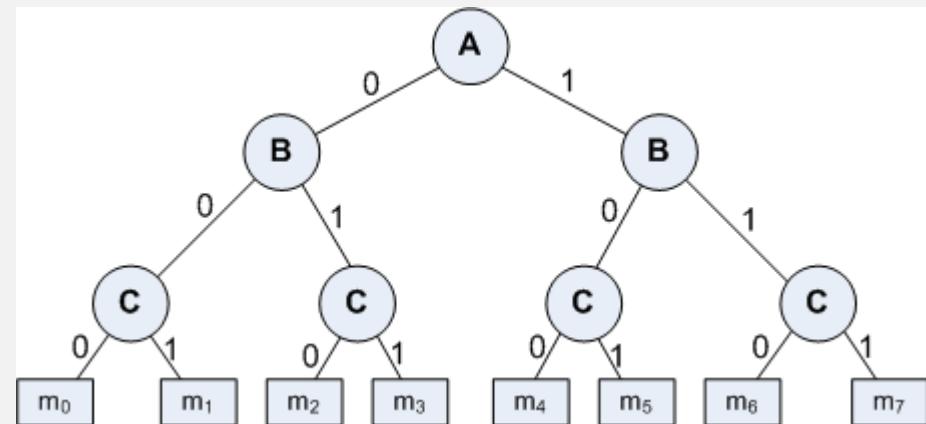
- Concept of Decision Tree
- Use of Decision Tree to classify data
- Basic algorithm to build Decision Tree
 - Some illustrations
- Concept of Entropy
 - Basic concept of entropy in information theory
 - Mathematical formulation of entropy
 - Calculation of entropy of a training set
- Decision Tree induction algorithms
 - ID3
 - CART
 - C4.5

BASIC CONCEPT

- A Decision Tree is an important data structure known to solve many computational problems

Example 19.1: Binary Decision Tree

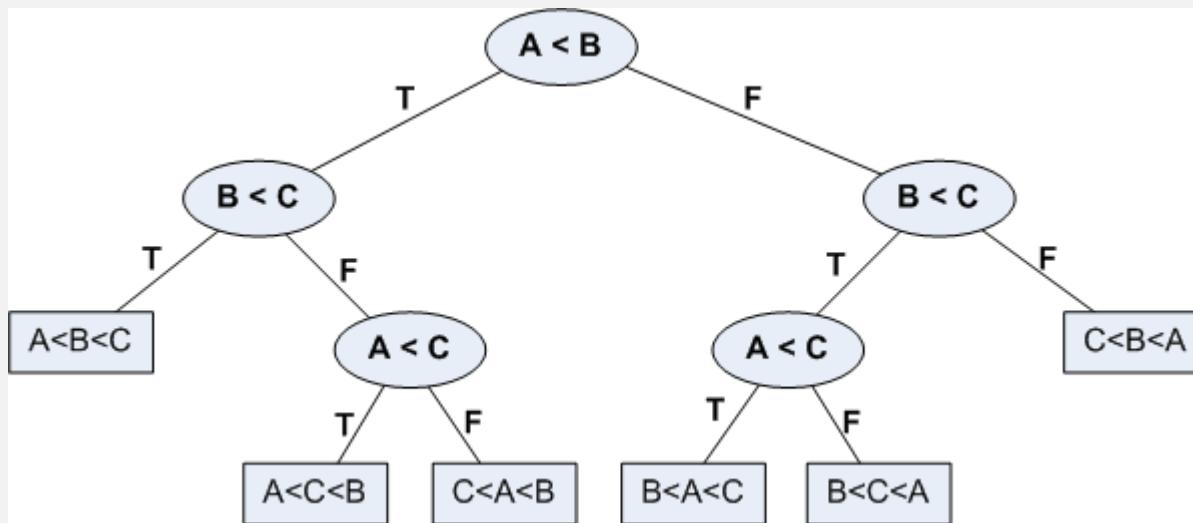
| A | B | C | f |
|---|---|---|-------|
| 0 | 0 | 0 | m_0 |
| 0 | 0 | 1 | m_1 |
| 0 | 1 | 0 | m_2 |
| 0 | 1 | 1 | m_3 |
| 1 | 0 | 0 | m_4 |
| 1 | 0 | 1 | m_5 |
| 1 | 1 | 0 | m_6 |
| 1 | 1 | 1 | m_7 |



BASIC CONCEPT

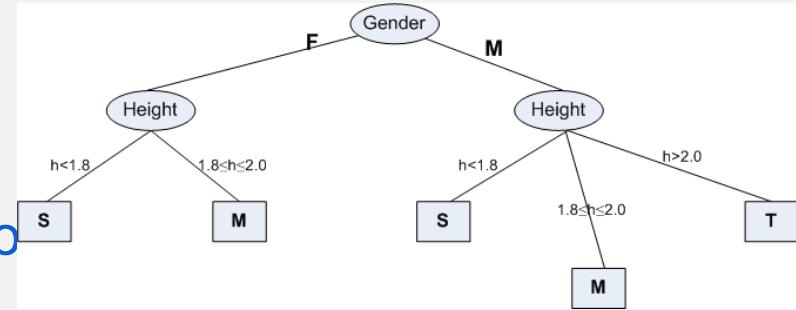
- In Example 19.1, we have considered a decision tree where values of any attribute if binary only. Decision tree is also possible where attributes are of continuous data type

Example 19.2: Decision Tree with numeric data



SOME CHARACTERISTICS

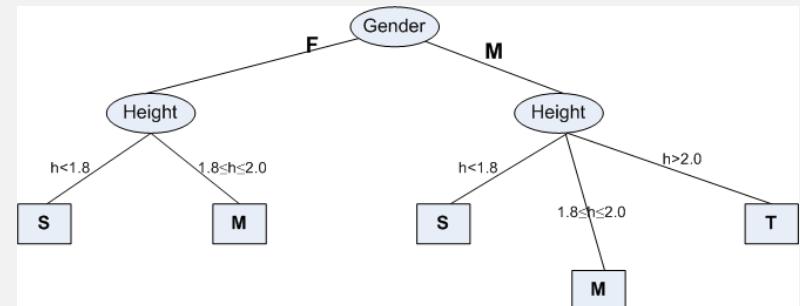
- Decision tree may be n -ary, $n \geq 2$.
- There is a special node called **root node**
- All nodes drawn with circle (ellipse) are called **internal nodes**.
- All nodes drawn with rectangle boxes are called **terminal nodes** or **leaf nodes**.
- Edges of a node represent the **outcome for a value** of the node.
- In a path, a node with same label **is never repeated**.
- Decision tree **is not unique**, as different ordering of internal nodes can give different decision trees.



DECISION TREE AND CLASSIFICATION TASK

- Decision tree helps us to classify data.

- Internal nodes are some attribute
- Edges are the values of attributes



- External nodes are the outcome of classification
- Such a classification is, in fact, made by posing questions starting from the root node to each terminal node.

DECISION TREE AND CLASSIFICATION TASK

Example 19.3 : Vertebrate Classification

| Name | Body Temperature | Skin Cover | Gives Birth | Aquatic Creature | Aerial Creature | Has Legs | Hibernates | Class |
|------------|------------------|------------|-------------|------------------|-----------------|----------|------------|-----------|
| Human | Warm | hair | yes | no | no | yes | no | Mammal |
| Python | Cold | scales | no | no | no | no | yes | Reptile |
| Salmon | Cold | scales | no | yes | no | no | no | Fish |
| Whale | Warm | hair | yes | yes | no | no | no | Mammal |
| Frog | Cold | none | no | semi | no | yes | yes | Amphibian |
| Komodo | Cold | scales | no | no | no | yes | no | Reptile |
| Bat | Warm | hair | yes | no | yes | yes | yes | Mammal |
| Pigeon | Warm | feathers | no | no | yes | yes | no | Bird |
| Cat | Warm | fur | yes | no | no | yes | no | Mammal |
| Leopard | Cold | scales | yes | yes | no | no | no | Fish |
| Turtle | Cold | scales | no | semi | no | yes | no | Reptile |
| Penguin | Warm | feathers | no | semi | no | yes | no | Bird |
| Porcupine | Warm | quills | yes | no | no | yes | yes | Mammal |
| Eel | Cold | scales | no | yes | no | no | no | Fish |
| Salamander | Cold | none | no | semi | no | yes | yes | Amphibian |

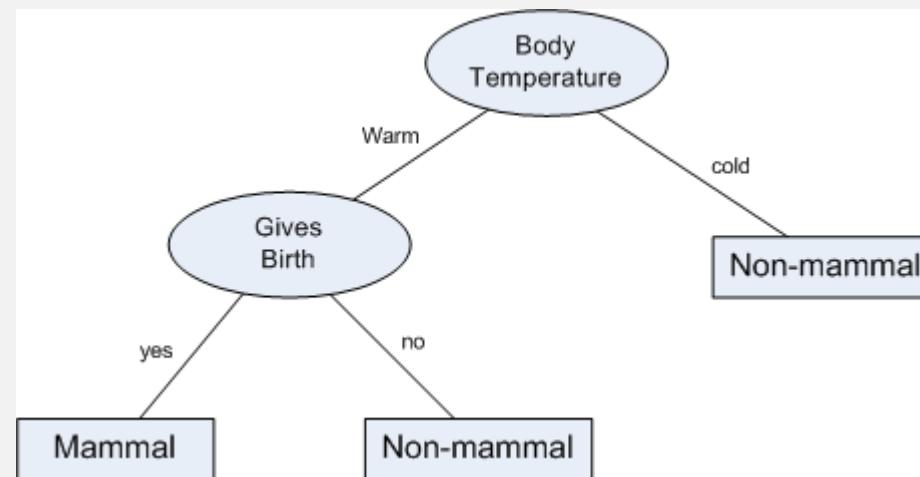
DECISION TREE AND CLASSIFICATION TASK

Example 19.3 : Vertebrate Classification

- Suppose, a new species is discovered as follows.

| Name | Body Temperature | Skin Cover | Gives Birth | Aquatic Creature | Aerial Creature | Has Legs | Hibernates | Class |
|--------------|------------------|------------|-------------|------------------|-----------------|----------|------------|-------|
| Gila Monster | cold | scale | no | no | no | yes | yes | ? |

Example 19.3) is as follows.



DECISION TREE AND CLASSIFICATION TASK

- Example 19.3 illustrates how we can solve a classification problem by asking a series of question about the attributes.
- Each time we receive an answer, a follow-up question is asked until we reach a conclusion about the class-label of the test.
- The series of questions and their answers can be organized in the form of a decision tree
 - As a hierarchical structure consisting of nodes and edges
- Once a decision tree is built, it is applied to any test to classify it.

DEFINITION OF DECISION TREE

Definition: **Decision Tree**

Given a database $D = \text{here}$ denotes a tuple, which is defined by a set of attribute set of classes $C = \text{.}$

A decision tree T is a tree associated with D that has the following properties:

- Each internal node is labeled with an attribute A_i ,
- Each edges is labeled with predicate that can be applied to the attribute associated with the parent node of it
- Each leaf node is labeled with class c_j

BUILDING DECISION TREE

- In principle, there are exponentially many decision tree that can be constructed from a given database (also called training data).
 - Some of the tree may not be optimum
 - Some of them may give inaccurate result
- Two approaches are known
 - **Greedy strategy**
 - A top-down recursive divide-and-conquer
 - **Modification of greedy strategy**
 - ID3
 - C4.5
 - CART, etc.

BUILT DECISION TREE ALGORITHM

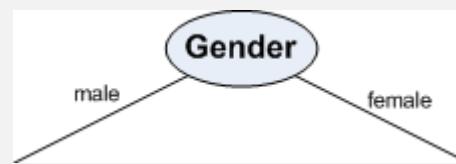
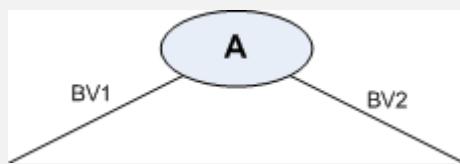
- **Algorithm BuiltDT**
- Input: D : Training data set
- Output: T : Decision tree

Steps

1. If all tuples in D belongs to the same class C_j
 Add a leaf node labeled as C_j
 Return *// Termination condition*
2. **Select** an attribute A_i (so that it is not selected twice in the same branch)
3. **Partition** $D = \{ D_1, D_2, \dots, D_p \}$ based on p different values of A_i in D
4. For each $D_k \in D$
 Create a node and add an edge between D and D_k with label as the A_i 's
 attribute value in D_k
5. For each $D_k \in D$
 BuildDT(D_k) *// Recursive call*
6. Stop

NODE SPLITTING IN BUILDDT ALGORITHM

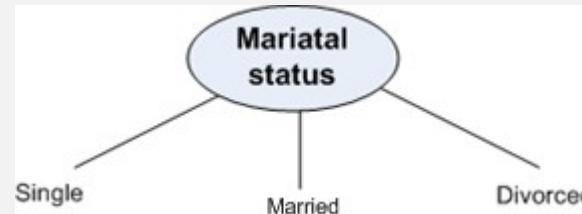
- BuildDT algorithm must provides a method for expressing **an attribute test condition** and **corresponding outcome** for different attribute type
- **Case: Binary attribute**
 - This is the simplest case of node splitting
 - The test condition for a binary attribute generates only two outcomes



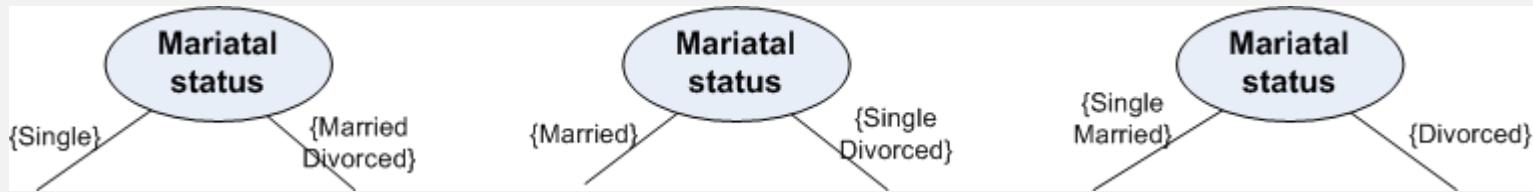
NODE SPLITTING IN BUILDDT ALGORITHM

- **Case: Nominal attribute**

- Since a nominal attribute can have many values, its test condition can be expressed in two ways:
 - A multi-way split
 - A binary split
- **Muti-way split:** Outcome depends on the number of distinct values for the corresponding attribute



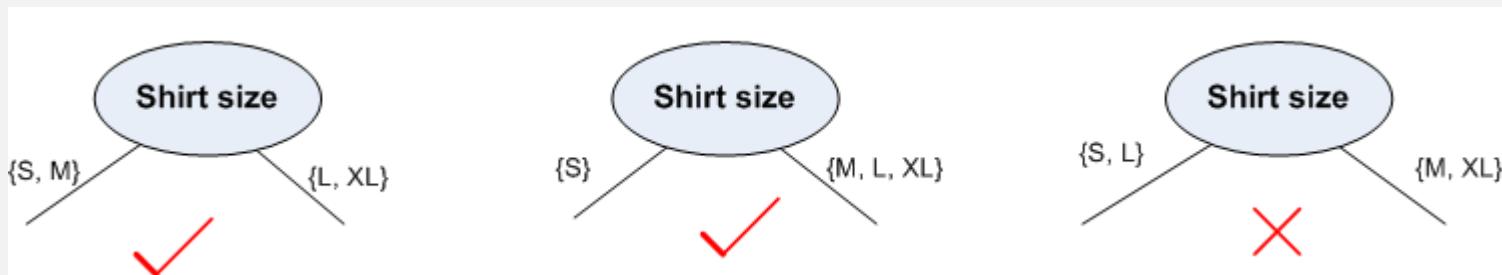
- **Binary splitting** by grouping attribute values



NODE SPLITTING IN BUILDDT ALGORITHM

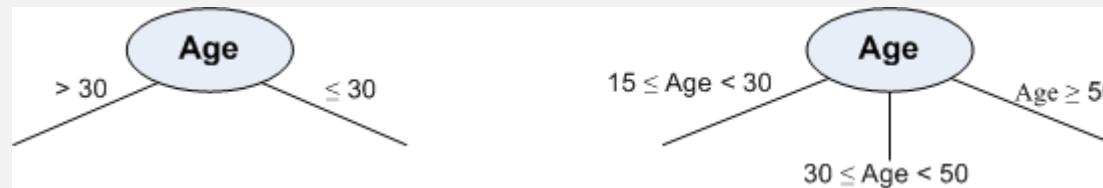
- **Case: Ordinal attribute**

- It also can be expressed in two ways:
 - A multi-way split
 - A binary split
- **Multi-way split:** It is same as in the case of nominal attribute
- **Binary splitting** attribute values should be grouped maintaining the **order** property of the attribute values



NODE SPLITTING IN BUILDDT ALGORITHM

- **Case: Numerical attribute**
 - For numeric attribute (with discrete or continuous values), a test condition can be expressed as a comparison set
 - **Binary outcome:** $A > v$ or $A \leq v$
 - In this case, decision tree induction must consider all possible split positions
 - **Range query :** $v_i \leq A < v_{i+1}$ for $i = 1, 2, \dots, q$ (if q number of ranges are chosen)
 - Here, q should be decided a priori



- For a numeric attribute, decision tree induction is a combinatorial optimization problem

ILLUSTRATION : BUILDDT ALGORITHM

Example 19.4: Illustration of BuildDT Algorithm

- Consider a training data set as shown.

| Person | Gender | Height | Class |
|--------|--------|--------|-------|
| 1 | F | 1.6 | S |
| 2 | M | 2.0 | M |
| 3 | F | 1.9 | M |
| 4 | F | 1.88 | M |
| 5 | F | 1.7 | S |
| 6 | M | 1.85 | M |
| 7 | F | 1.6 | S |
| 8 | M | 1.7 | S |
| 9 | M | 2.2 | T |
| 10 | M | 2.1 | T |
| 11 | F | 1.8 | M |
| 12 | M | 1.95 | M |
| 13 | F | 1.9 | M |
| 14 | F | 1.8 | M |
| 15 | F | 1.75 | S |

Attributes:

Gender = {Male(M), Female (F)} // Binary attribute

Height = {1.5, ..., 2.5} // Continuous attribute

Class = {Short (S), Medium (M), Tall (T)}

Given a person, we are to test in which class s/he belongs

ILLUSTRATION : BUILDDT ALGORITHM

- To built a decision tree, we can select an attribute in two different orderings: <Gender, Height> or <Height, Gender>
- Further, for each ordering, we can choose different ways of splitting
- Different instances are shown in the following.
- **Approach 1 : <Gender, Height>**

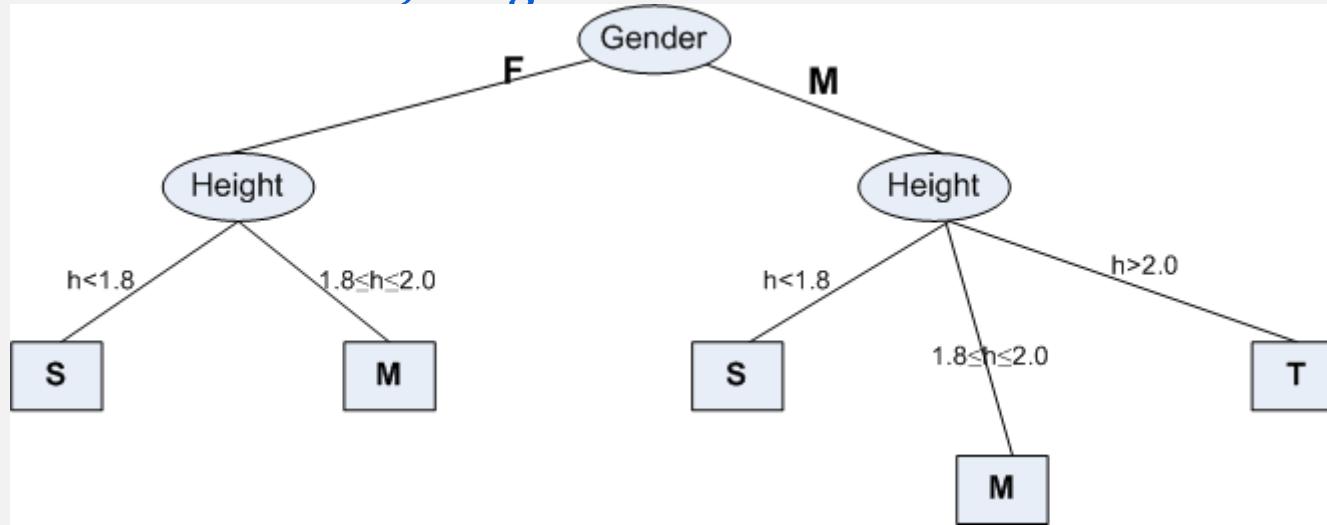


ILLUSTRATION : BUILDDT ALGORITHM

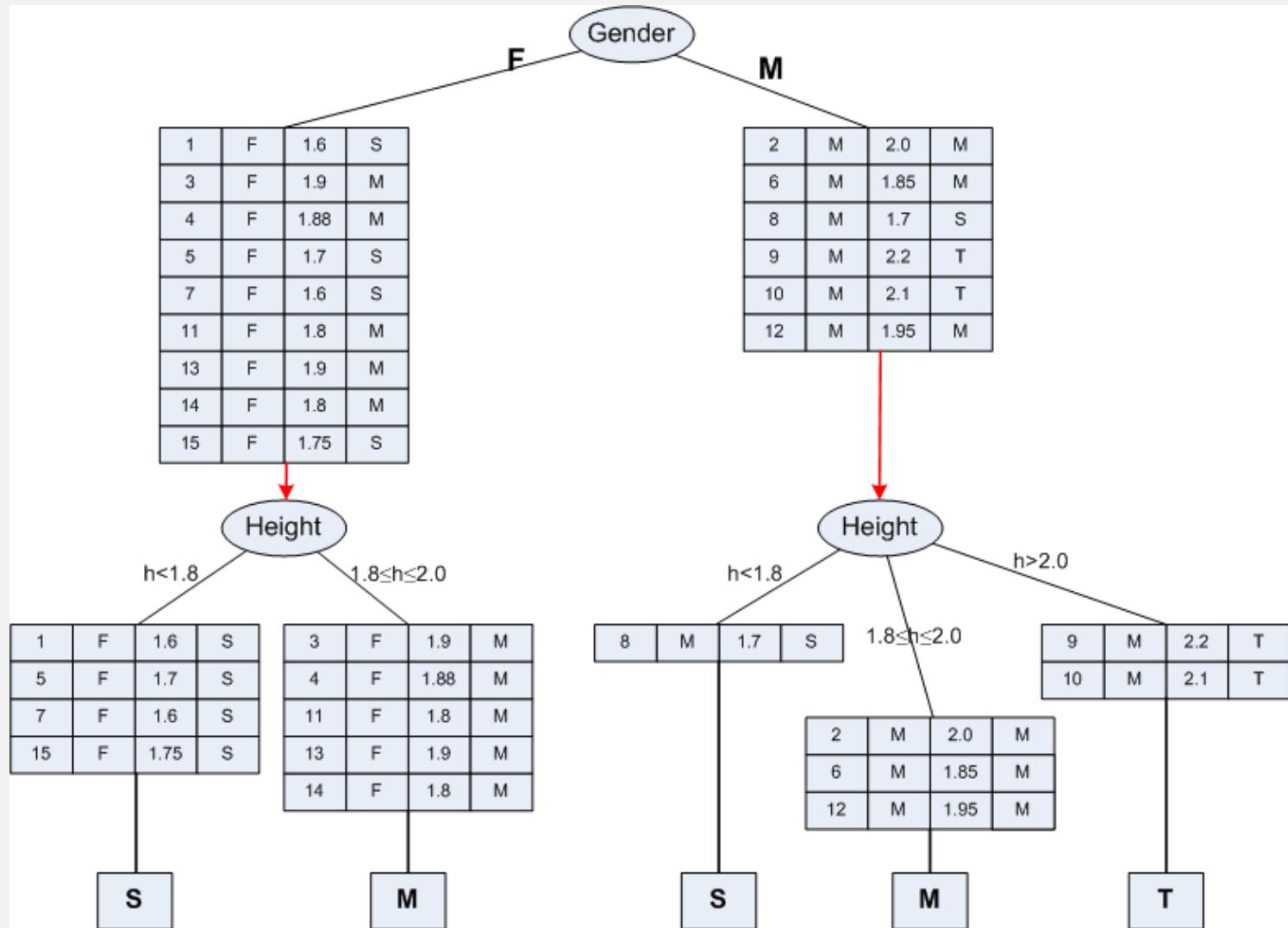


ILLUSTRATION : BUILDDT ALGORITHM

- Approach 2 : <Height, Gender>

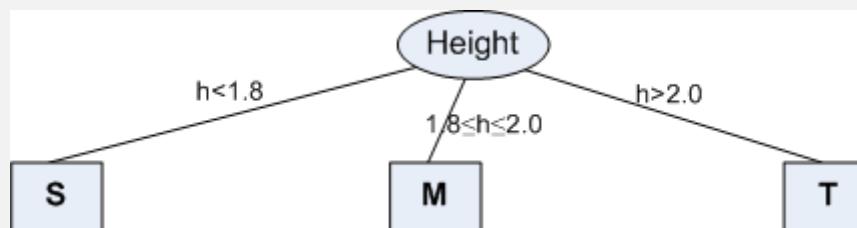
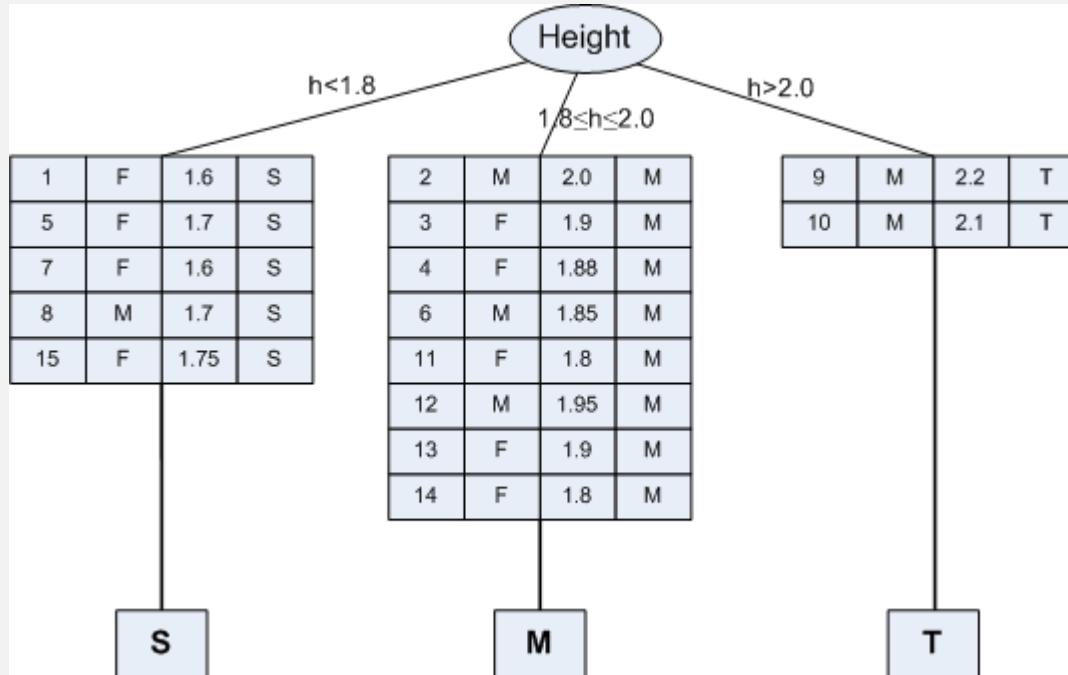


ILLUSTRATION : BUILDDT ALGORITHM

Example 19.5: Illustration of BuildDT Algorithm

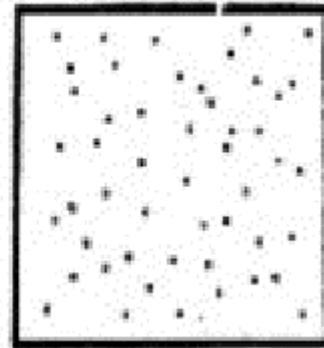
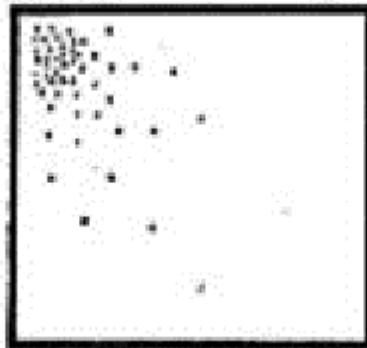
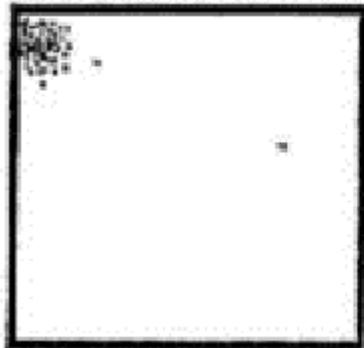
- Consider an anonymous database as shown.

| A1 | A2 | A3 | A4 | Class |
|-----|-----|-----|-----|-------|
| a11 | a21 | a31 | a41 | C1 |
| a12 | a21 | a31 | a42 | C1 |
| a11 | a21 | a31 | a41 | C1 |
| a11 | a22 | a32 | a41 | C2 |
| a11 | a22 | a32 | a41 | C2 |
| a12 | a22 | a31 | a41 | C1 |
| a11 | a22 | a32 | a41 | C2 |
| a11 | a22 | a31 | a42 | C1 |
| a11 | a21 | a32 | a42 | C2 |
| a11 | a22 | a32 | a41 | C2 |
| a12 | a22 | a31 | a41 | C1 |
| a12 | a22 | a31 | a42 | C1 |

- Is there any “clue” that enables to select the “best” attribute first?
- Suppose, following are two attempts:
 - A1□A2□A3□A4 [naïve]
 - A3□A2□A4□A1 [Random]
- Draw the decision trees in the above-mentioned two cases.
- Are the trees different to classify any test data?
- If any other sample data is added into the database, is that likely to alter the decision tree already obtained?

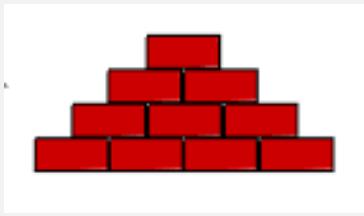
CONCEPT OF ENTROPY

CONCEPT OF ENTROPY

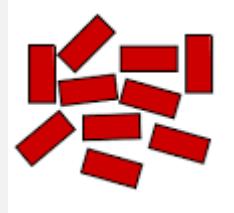


If a point represents a gas molecule,
then which system has the more
entropy?

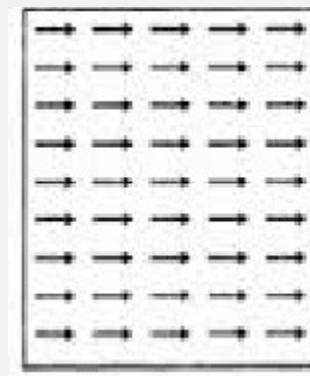
How to measure? ?



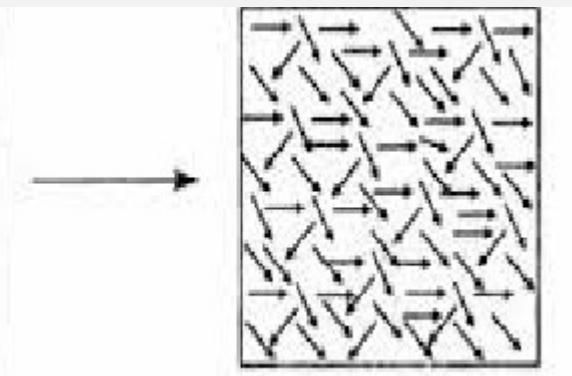
More ordered
less entropy



Less ordered
higher entropy



More organized or
ordered (less probable)



Less organized or
disordered (more probable)

CONCEPT OF ENTROPY



Universe!

What was its entropy value at its starting point?

ENTROPY AND ITS MEANING

- Entropy is an important concept used in Physics in the context of heat and thereby uncertainty of the states of a matter.
- At a later stage, with the growth of Information Technology, entropy becomes an important concept in **Information Theory**.
- To deal with the classification job, entropy is an important concept, which is considered as
 - an information-theoretic measure of the “uncertainty” contained in a training data
 - due to the presence of more than one classes.

ENTROPY IN INFORMATION THEORY

- The entropy concept in information theory first time coined by Claude Shannon (1850).
- The first time it was used to measure the “information content” in messages.
- According to his concept of entropy, presently entropy is widely being used as a way of representing messages for efficient transmission by Telecommunication Systems.

MEASURE OF INFORMATION CONTENT

- People, in general, are information hungry!
- Everybody wants to acquire information (from newspaper, library, nature, fellows, etc.)
 - Think how a crime detector do it to know about the crime from crime spot and criminal(s).
 - Kids annoyed their parents asking questions.
- In fact, fundamental thing is that we gather information asking questions (and decision tree induction is no exception).
 -
- We may note that information gathering may be with certainty or uncertainty.

MEASURE OF INFORMATION CONTENT

Example 19.6

- a) Guessing a birthday of your classmate

It is with uncertainty ~

Whereas guessing the day of his/her birthday is .

This uncertainty, we may say varies between 0 to 1, both inclusive.

- b) As another example, a question related to event with eventuality (or impossibility) will be answered with 0 or 1 uncertainty.

- Does sun rises in the East? (answer is with 0 uncertainty)
- Will mother give birth to male baby? (answer is with $\frac{1}{2}$ uncertainty)
- Is there a planet like earth in the galaxy? (answer is with an extreme uncertainty)

DEFINITION OF ENTROPY

Suppose there are m distinct objects, which we want to identify by asking a series of **Yes/No** questions. Further, we assume that m is an exact power of 2, say , where .

Definition: Entropy

The entropy of a set of m distinct values is the minimum number of yes/no questions needed to determine an unknown values from these m possibilities.

ENTROPY CALCULATION

- How can we calculate the minimum number of questions, that is, entropy?
 - There are two approaches:
 - Brute –force approach
 - Clever approach.

Example 19.7: City quiz

Suppose, Thee is a quiz relating to guess a city out of 8 cities, which are as follows:

Bangalore, Bhopal, Bhubaneshwar, Delhi, Hyderabad, Kolkata, Madras, Mumbai

The question is, “Which city is called **city of joy**”?

APPROACH 1: BRUTE-FORCE SEARCH

- Brute force approach
 - We can ask “Is it city X ”,
 - if yes stop, else ask next ...

In this approach, we can ask such questions randomly choosing one city at a time. As a matter of randomness, let us ask the questions, not necessarily in the order, as they are in the list.

Q.1: Is the city Bangalore? No

Q.2: Is the city Bhubaneswar? No

Q.3: Is the city Bhopal? No

Q.4: Is the city Delhi? No

Q.5: Is the city Hyderabad? No

Q.6: Is the city Madras? No

Q.7: Is the city Mumbai? No

No need to ask further question! Answer is already out by the Q.7. If asked randomly, each of these possibilities is equally likely with probability . Hence on the average, we need

questions.

APPROACH 2: CLEVER APPROACH

- Clever approach (binary search)
 - In this approach, we divide the list into two halves, pose a question for a half
 - Repeat the same recursively until we get yes answer for the unknown.

Q.1: Is it Bangalore, Bhopal, Bhubaneswar or Delhi? No

Q.2: Is it Madras or Mumbai? No

Q.3: Is it Hyderabad? No

So after fixing 3 questions, we are able to crack the answer.

Note:

Approach 2 is considered to be the best strategy because it will invariably find the answer and will do so with a minimum number of questions on the average than any other strategy.

Approach 1 occasionally do better (when you are lucky enough!)

- It is no coincidence that , and the minimum number of yes/no questions needed is 3.
- If $m = 16$, then , and we can argue that we need 4 questions to solve the problem. If $m = 32$, then 5 questions, $m = 256$, then 8 questions and so on.

REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 19

Decision Tree Induction – ID3

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

Entropy Calculation

Lemma 19.1: Entropy calculation

The minimum number of *yes/no* questions needed to identify an unknown object from $m = 2^n$ equally likely possible object is n .

If m is not a power of 2, then the entropy of a set of m distinct objects that are equally likely is $\log_2 m$

ENTROPY IN MESSAGES

- We know that the most conventional way to code information is using binary bits, that is, using 0s and 1s.
- The answer to a question that can only be answered *yes/no* (with equal probability) can be considered as containing one **unit of information**, that is, one bit.
- In other words, the unit of information can also be looked at as the amount of information that can be **coded** using only 0s and 1s.

ENTROPY IN MESSAGES

Example 19.7: Information coding

- If we have **two** possible objects say **male** and **female**, then we use the coding

$0 = \text{female}$

$1 = \text{male}$

$$m = 2 (= 2^n, n = 1)$$

- We can encode **four** possible objects say **East**, **West**, **North**, **South** using two bits, for example

$00 : \text{North}$

$01 : \text{East}$

$$m = 4 (= 2^n, n = 2)$$

$10 : \text{West}$

$11 : \text{South}$

- We can encode **eight** values say eight different colours, we need to use **three** bits, such as

$000 : \text{Violet}$

$100 : \text{Yellow}$

$001 : \text{Orange}$

$101 : \text{Indigo}$

$$m = 8 (= 2^n, n = 3)$$

$010 : \text{Red}$

$110 : \text{Blue}$

$011 : \text{White}$

$111 : \text{Green}$

Thus, in general, to code m values, each in a distinct manner, we need n bits such that $m = 2^n$.

ENTROPY IN MESSAGES

- In this point, we can note that to identify an object, if it is encoded with bits, then we have to ask questions in an alternative way. For example
 - Is the first bit 0?
 - Is the second bit 0?
 - Is the third bit 0? and so on
- Thus, we need n questions, if m objects are there such that $m = 2^n$.
- The above leads to (an alternative) and equivalent definition of entropy

Definition 19.1: Entropy

The entropy of a set of m distinct values is the number of bits needed to encode all the values in the most efficient way.

MESSAGES WHEN ($m \neq 2^n$)

- In the previous discussion, we have assumed that m , the number of distinct objects is exactly a power of 2, that is $m = 2^n$ for some $n \geq 1$ and all m objects are equally likely.
- This is mere an assumption to make the discussion simplistic.
- In the following we try to redefine the entropy calculation in more general case, that is, when $m \neq 2^n$ and not necessarily m objects are equally probable. Let us consider a different instance of yes/no question game, which is as follows.

Example 19.8: Name game

- There are seven days: Sun, Mon, Tue, Wed, Thu, Fri, Sat.
- We are to identify a sequence of $k \geq 1$ such values (each one chosen independently of the others, that is, repetitions are allowed).
- We denote the minimum number of yes/no questions needed to identify a sequence of k unknown values drawn independently from m possibilities as E_k^m , the entropy in this case.
- In other words, E_k^m is the number of questions required to discriminate amongst m^k distinct possibilities.

MESSAGES WHEN ($m \neq 2^n$)

- Here, $m = 7$ (as stated in the game of sequence of days) and $k = 6$ (say).
- An arbitrary sequence may be {Tue, Thu, Tue, Mon, Sun, Tue}, etc. There are $7^6 = 117649$ possible sequences of six days.
- From our previous understanding, we can say that the minimum number of yes/no questions that is required to identify such a sequence is $\log_2 117649 = 16.8443$.
- Since, this is a non integer number, and the number of question should be an integer, we can say 17 questions are required. Thus,

$$E_6^7 = \lceil \log_2 7^6 \rceil$$

- In general,

$$E_k^m = \lceil \log_2 m^k \rceil$$

- Alternatively, the above can be written as,

$$\lceil \log_2 m^k \rceil \leq E_k^m \leq \lceil \log_2 m^k \rceil + 1$$

- Or

$$\lceil \log_2 m \rceil \leq \frac{E_k^m}{k} \leq \lceil \log_2 m \rceil + \frac{1}{k}$$

ENTROPY OF MESSAGES WHEN ($m \neq 2^n$)

Note that here $\frac{E_k^m}{k}$ is the average number of questions needed to determine each of the values in a sequence of k values. By choosing a large enough value of k , that is, a long enough sequence, the value of $\frac{1}{k}$ can be made as small as we wish. Thus, the average number of questions required to determine each value can be made arbitrarily close to $\log_2 m$. This is evident from our earlier workout, for example, tabulated below, for $m = 7$.

$$E_k^m = \lceil \log_2 m^k \rceil$$

| k | m^k | $\lceil \log_2 m^k \rceil$ | No. Q | $\frac{\text{No. } Q}{k}$ |
|-----------------------|-------------------------|--|--------------|---|
| 6 | 117649 | 16.84413 | 17 | 2.8333 |
| 21 | | 58.95445 | 59 | 2.8095 |
| 1000 | | 2807.3549 | 2808 | 2.8080 |
| | | | | |

No. Q = Number of questions

Note that $\log_2 7 \approx 2.8074$ and $\frac{\text{No. } Q}{k} \approx \log_2 7$. Further, $\frac{\text{No. } Q}{k} = \frac{E_k^7}{k}$ i.e. $\frac{E_k^7}{k} = \log_2 7$ (is independent of k and is a constant!)

ENTROPY OF MESSAGES WHEN ($m \neq 2^n$)

Lemma 19.3: Entropy Calculation

The entropy of a set of m distinct objects is $\log_2 m$ even when m is not exactly a power of 2.

- We have arrived at a conclusion that $E = \log_2 m$ for any value of m , irrespective of whether it is a power of 2 or not.

Note: E is not necessarily be an integer always.

- Next, we are to have our observation, if all m objects are not equally probable.
- Suppose, p_i denotes the frequency with which the i^{th} of the m objects occurs, where $0 \leq p_i \leq 1$ for all p_i such that

$$\sum_{i=1}^m p_i = 1$$

DISCRIMINATING AMONGST M VALUES ($m \neq 2^n$)

Example 19.8: Discriminating among objects

- Suppose four objects A, B, C and D which occur with frequencies $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ and $\frac{1}{8}$, respectively. (A B C D A A A B)
- Thus, in this example, $m = 4$ and $p_1 = \frac{1}{2}, p_2 = \frac{1}{4}, p_3 = \frac{1}{8}$ and $p_4 = \frac{1}{8}$.
- Using standard 2-bit encoding, we can represent them as

$$A = 00,$$

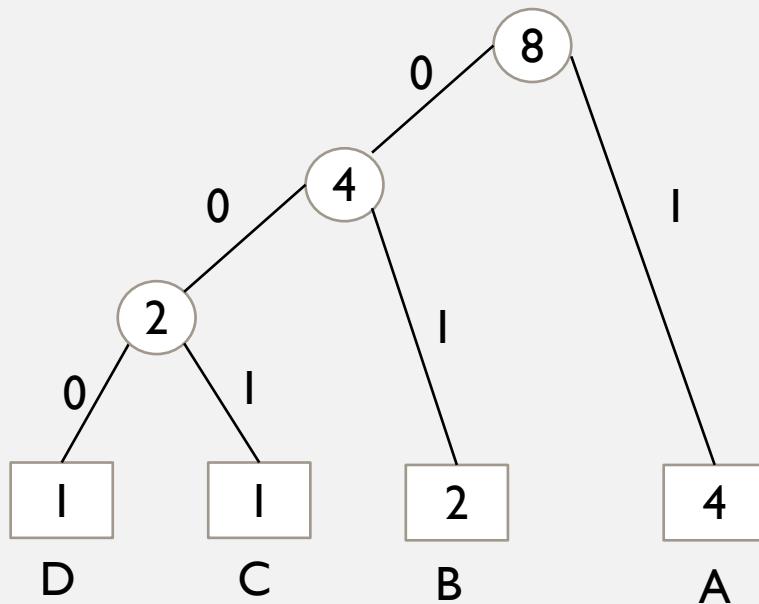
$$B = 01,$$

$$C = 10,$$

$$D = 11.$$

- Also, we can follow variable length coding (also called Huffman coding) as an improved way of representing them.

HUFFMAN CODING



- The Huffman coding of *A, B, C and D* with their frequencies $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{8}$ are shown below.

A = 1

B = 01

C = 001

D = 000

DISCRIMINATING AMONGST M VALUES ($m \neq 2^n$)

- With the above representation say, if A is to be identified, then we need to examine only one question, for B it is 2 and for C and D both, it is 3.
- Thus, on the average, we need

$$\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1.75 \text{ bits}$$

- This is the number of yes/no questions to identify any one of the four objects, whose frequency of occurrences are not uniform.
- This is simply in contrast to 2-bit encoding, where we need 2-bits (questions) on the average.

DISCRIMINATING AMONGST M VALUES ($m \neq 2^n$)

- It may be interesting to note that even with variable length encoding, there are several ways of encoding. Few of them are given below.

$$1) A = 0$$

$$B = 11$$

$$C = 100$$

$$D = 101$$

$$2) A = 01$$

$$B = 1$$

$$C = 001$$

$$D = 000$$

$$3) A = 101$$

$$B = 001$$

$$C = 10011$$

$$D = 100001$$

- The calculation of entropy in the observed cases can be obtained as:
 - 1.75
 - 2
 - 3) 3.875
- Anyway, key to finding the most efficient way of encoding is to **assign a smallest number of bits to the object with highest frequency** and so on.
- The above observation is also significant in the sense that it provides a **systematic way of finding a sequence of well-chosen question in order to identify an object at a faster rate**.

INFORMATION CONTENT

Based on the previous discussion we can easily prove the following lemma.

Lemma 19.4: Information content

If an object occurs with frequency p , then the most efficient way to represent it with $\log_2(1/p)$ bits.

Example 19.9: Information content

- A which occurs with frequency $\frac{1}{2}$ is represented by 1-bit, B which occurs with frequency $\frac{1}{4}$ represented by 2-bits and both C and D which occurs with frequency $\frac{1}{8}$ are represented by 3 bits each.

ENTROPY CALCULATION

We can generalize the above understanding as follows.

- If there are m objects with frequencies p_1, p_2, \dots, p_m , then the average number of bits (i.e. questions) that need to be examined a value, that is, entropy is the frequency of occurrence of the i^{th} value multiplied by the number of bits that need to be determined, summed up values of i from 1 to m .

Theorem 19.4: Entropy calculation

If p_i denotes the frequencies of occurrences of m distinct objects, then the entropy E is

$$E = \sum_{i=1}^m p_i \log(1/p_i) \text{ and } \sum_{i=1}^m p_i = 1$$

Note:

- If all are equally likely, then $p_i = \frac{1}{m}$ and $E = \log_2 m$; it is the special case.

ENTROPY OF A TRAINING SET

- If there are k classes c_1, c_2, \dots, c_k and p_i for $i = 1$ to k denotes the number of occurrences of classes c_i divided by the total number of instances (i.e., the frequency of occurrence of c_i) in the training set, then entropy of the training set is denoted by

$$E = - \sum_{i=1}^m p_i \log_2 p_i$$

Here, E is measured in “bits” of information.

Note:

- The above formula should be summed over the non-empty classes only, that is, classes for which $p_i \neq 0$
- E is always a positive quantity
- E takes its minimum value (zero) if and only if all the instances have the same class (i.e., the training set with only one non-empty class, for which the probability 1).
- Entropy takes its maximum value when the instances are equally distributed among k possible classes. In this case, the maximum value of E is $\log_2 k$.

ENTROPY OF A TRAINING SET

Example 19.10: OPTH dataset

Consider the OTPH data shown in the following table with total 24 instances in it.

| Age | Eye sight | Astigmatic | Use Type | Class |
|-----|-----------|------------|----------|-------|
| 1 | 1 | 1 | 1 | 3 |
| 1 | 1 | 1 | 2 | 2 |
| 1 | 1 | 2 | 1 | 3 |
| 1 | 1 | 2 | 2 | 1 |
| 1 | 2 | 1 | 1 | 3 |
| 1 | 2 | 1 | 2 | 2 |
| 1 | 2 | 2 | 1 | 3 |
| 1 | 2 | 2 | 2 | 1 |
| 2 | 1 | 1 | 1 | 3 |
| 2 | 1 | 1 | 2 | 2 |
| 2 | 1 | 2 | 1 | 3 |
| 2 | 1 | 2 | 2 | 1 |
| 2 | 2 | 1 | 1 | 3 |
| 2 | 2 | 1 | 2 | 2 |
| 2 | 2 | 2 | 1 | 3 |
| 2 | 2 | 2 | 2 | 3 |
| 3 | 1 | 1 | 1 | 3 |
| 3 | 1 | 1 | 2 | 3 |
| 3 | 2 | 2 | 1 | 3 |
| 3 | 2 | 1 | 2 | 2 |
| 3 | 2 | 2 | 1 | 3 |
| 3 | 2 | 2 | 2 | 3 |

A coded forms for all values of attributes are used to avoid the cluttering in the table.

ENTROPY OF A TRAINING SET

Specification of the attributes are as follows.

| Age | Eye Sight | Astigmatic | Use Type |
|----------------|------------------|------------|-------------|
| 1: Young | 1: Myopia | 1: No | 1: Frequent |
| 2: Middle-aged | 2: Hypermetropia | 2: Yes | 2: Less |
| 3: Old | | | |

Class: 1: Contact Lens 2:Normal glass 3: Nothing

In the OPTH database, there are 3 classes and 4 instances with class 1, 5 instances with class 2 and 15 instances with class 3. Hence, entropy E of the database is:

$$E = -\frac{4}{24} \log_2 \frac{4}{24} - \frac{5}{24} \log_2 \frac{5}{24} - \frac{15}{24} \log_2 \frac{15}{24} = 1.3261$$

Note:

- The entropy of a training set implies the number of yes/no questions, on the average, needed to determine an unknown test to be classified.
- It is very crucial to decide the series of questions about the value of a set of attribute, which collectively determine the classification. Sometimes it may take one question, sometimes many more.
- Decision tree induction helps us to ask such a series of questions. In other words, we can utilize entropy concept to build a better decision tree.

How entropy can be used to build a decision tree ?

DECISION TREE INDUCTION TECHNIQUES

- Decision tree induction is a top-down, recursive and divide-and-conquer approach.
- The procedure is to choose an attribute and split it into from a larger training set into smaller training sets.
- Different algorithms have been proposed to take a good control over
 1. Choosing the best attribute to be splitted, and
 2. Splitting criteria
- Several algorithms have been proposed for the above tasks. In this lecture, we shall limit our discussions into three important of them
 - **ID3**
 - **C 4.5**
 - **CART**

ALGORITHM ID3

ID3: DECISION TREE INDUCTION ALGORITHMS

- Quinlan [1986] introduced the ID3, a popular short form of Iterative Dichotomizer 3 for decision trees from a set of training data.
- In ID3, each node corresponds to a splitting attribute and each arc is a possible value of that attribute.
- At each node, the splitting attribute is selected to be the most informative among the attributes not yet considered in the path starting from the root.

ALGORITHM ID3

- In ID3, **entropy is used** to measure how informative a node is.
 - It is observed that splitting on any attribute has **the property that average entropy of the resulting training subsets will be less than or equal to** that of the previous training set.
- ID3 algorithm defines a measurement of a splitting called **Information Gain** to determine the goodness of a split.
 - The attribute with the **largest value of information gain** is chosen as the splitting attribute and
 - it partitions into a number of smaller training sets based on the **distinct values of attribute** under split.

DEFINING INFORMATION GAIN

- We consider the following symbols and terminologies to define information gain, which is denoted as α .
- $D \equiv$ denotes the training set at any instant
- $|D| \equiv$ denotes the size of the training set D
- $E(D) \equiv$ denotes the entropy of the training set D
- The entropy of the training set D

$$E(D) = -\sum_{i=1}^k p_i \log_2(p_i)$$

- where the training set D has c_1, c_2, \dots, c_k , the k number of distinct classes and
- $p_i, 0 < p_i \leq 1$ is the probability that an arbitrary tuple in D belongs to class c_i ($i = 1, 2, \dots, k$).

DEFINING INFORMATION GAIN

- p_i can be calculated as

$$p_i = \frac{|C_{i,D}|}{|D|}$$

- where $C_{i,D}$ is the set of tuples of class c_i in D .
- Suppose, we want to partition D on some attribute A having m distinct values $\{a_1, a_2, \dots, a_m\}$.
- Attribute A can be considered to split D into m partitions $\{D_1, D_2, \dots, D_m\}$, where D_j ($j = 1, 2, \dots, m$) contains those tuples in D that have outcome a_j of A .

DEFINING INFORMATION GAIN

Definition 19.4: Weighted Entropy

The weighted entropy denoted as $E_A(D)$ for all partitions of D with respect to A is given by:

$$E_A(D) = \sum_{j=1}^m \frac{|D_j|}{|D|} E(D_j)$$

Here, the term $\frac{|D_j|}{|D|}$ denotes the weight of the j -th training set.

More meaningfully, $E_A(D)$ is the expected information required to classify a tuple from D based on the splitting of A .

DEFINING INFORMATION GAIN

- Our objective is to take A on splitting to produce an exact classification (also called pure), that is, all tuples belong to one class.
- However, it is quite likely that the partitions is impure, that is, they contain tuples from two or more classes.
- In that sense, $E_A(D)$ is a measure of impurities (or purity). A lesser value of $E_A(D)$ implying more power the partitions are.

Definition 19.5: Information Gain

Information gain, $\alpha(A, D)$ of the training set D splitting on the attribute A is given by

$$\alpha(A, D) = E(D) - E_A(D)$$

In other words, $\alpha(A, D)$ gives us an estimation how much would be gained by splitting on A . The attribute A with the highest value of α should be chosen as the splitting attribute for D .

INFORMATION GAIN CALCULATION

Example 19.11 : Information gain on splitting OPTH

- Let us refer to the OPTH database discussed earlier.
- Splitting on **Age** at the root level, it would give three subsets D_1, D_2 and D_3 as shown in the tables in the following three slides.
- The entropy $E(D_1), E(D_2)$ and $E(D_3)$ of training sets D_1, D_2 and D_3 and corresponding weighted entropy $E_{Age}(D_1), E_{Age}(D_2)$ and $E_{Age}(D_3)$ are also shown alongside.
- The Information gain $\alpha (Age, OPTH)$ is then can be calculated as **0.0394**.
- Recall that entropy of OPTH data set, we have calculated as $E(OPTH) = \textcolor{red}{1.3261}$
(see Slide #17)

INFORMATION GAIN CALCULATION

Example 19.11 : Information gain on splitting OPTH

Training set: $D_1(\text{Age} = 1)$

| Age | Eye-sight | Astigmatism | Use type | Class |
|-----|-----------|-------------|----------|-------|
| 1 | 1 | 1 | 1 | 3 |
| 1 | 1 | 1 | 2 | 2 |
| 1 | 1 | 2 | 1 | 3 |
| 1 | 1 | 2 | 2 | 1 |
| 1 | 2 | 1 | 1 | 3 |
| 1 | 2 | 1 | 2 | 2 |
| 1 | 2 | 2 | 1 | 3 |
| 1 | 2 | 2 | 2 | 1 |

$$E(D_1) = -\frac{2}{8} \log_2\left(\frac{2}{8}\right) - \frac{2}{8} \log_2\left(\frac{2}{8}\right) \\ - \frac{4}{8} \log_2\left(\frac{4}{8}\right) = 1.5$$

$$E_{Age}(D_1) = \frac{8}{24} \times 1.5 = 0.5000$$

CALCULATING INFORMATION GAIN

Training set: $D_2(\text{Age} = 2)$

| Age | Eye-sight | Astigmatism | Use type | Class |
|-----|-----------|-------------|----------|-------|
| 2 | 1 | 1 | 1 | 3 |
| 2 | 1 | 1 | 2 | 2 |
| 2 | 1 | 2 | 1 | 3 |
| 2 | 1 | 2 | 2 | 1 |
| 2 | 2 | 1 | 1 | 3 |
| 2 | 2 | 1 | 2 | 2 |
| 2 | 2 | 2 | 1 | 3 |
| 2 | 2 | 2 | 2 | 3 |

$$E(D_2) = -\frac{1}{8} \log_2\left(\frac{1}{8}\right) - \frac{2}{8} \log_2\left(\frac{2}{8}\right) - \frac{5}{8} \log_2\left(\frac{5}{8}\right) \\ = 1.2988$$

$$E_{Age}(D_2) = \frac{8}{24} \times 1.2988 = 0.4329$$

CALCULATING INFORMATION GAIN

Training set: $D_3(\text{Age} = 3)$

| Age | Eye-sight | Astigmatism | Use type | Class |
|-----|-----------|-------------|----------|-------|
| 3 | 1 | 1 | 1 | 3 |
| 3 | 1 | 1 | 2 | 3 |
| 3 | 1 | 2 | 1 | 3 |
| 3 | 1 | 2 | 2 | 1 |
| 3 | 2 | 1 | 1 | 3 |
| 3 | 2 | 1 | 2 | 2 |
| 3 | 2 | 2 | 1 | 3 |
| 3 | 2 | 2 | 2 | 3 |

$$E(D_3) = -\frac{1}{8} \log_2\left(\frac{1}{8}\right) - \frac{1}{8} \log_2\left(\frac{1}{8}\right) - \frac{6}{8} \log_2\left(\frac{6}{8}\right) = 1.0613$$

$$E_{Age}(D_3) = \frac{8}{24} \times 1.0613 = 0.3504$$

$$\alpha(Age, D) = 1.3261 - (0.5000 + 0.4329 + 0.3504) = 0.0394$$

INFORMATION GAINS FOR DIFFERENT ATTRIBUTES

- In the same way, we can calculate the information gains, when splitting the OPTH database on **Eye-sight**, **Astigmatic** and **Use Type**. The results are summarized below.
- Splitting attribute: **Age**

$$\alpha(Age, OPTH) = 0.0394$$

- Splitting attribute: **Eye-sight**

$$\alpha(Eye_sight, OPTH) = 0.0395$$

- Splitting attribute: **Astigmatic**

$$\alpha(Astigmatic, OPTH) = 0.3770$$

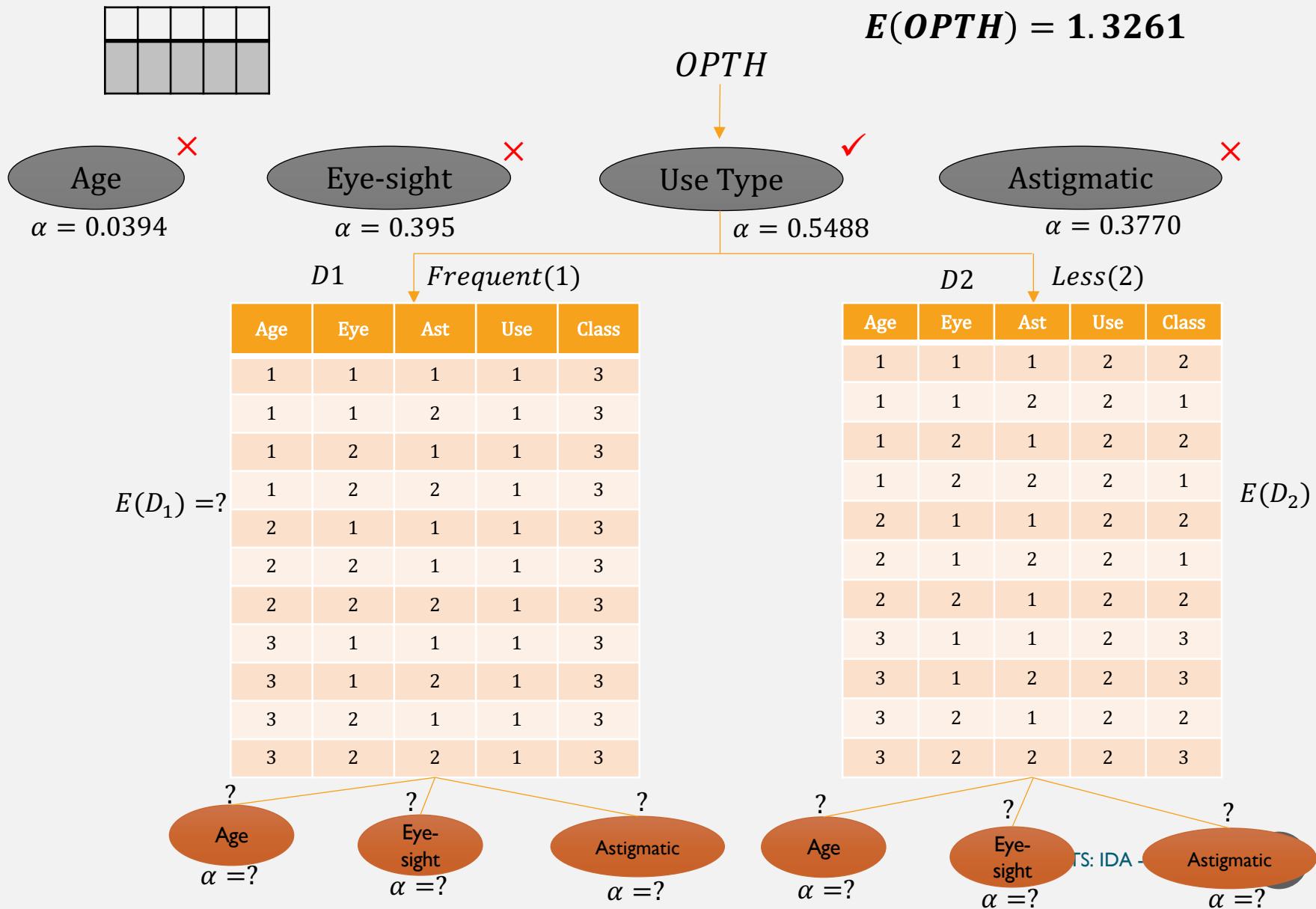
- Splitting attribute: **Use Type**

$$\alpha(Use\ Type, OPTH) = 0.5488$$

DECISION TREE INDUCTION : ID3 WAY

- The ID3 strategy of attribute selection is to choose to split on the attribute that gives the greatest reduction in the weighted average entropy
 - The one that maximizes the value of information gain
- In the example with OPTH database, the larger values of information gain is $\alpha(\text{Use Type}, OPTH) = 0.5488$
 - Hence, the attribute should be chosen for splitting is “**Use Type**”.
- The process of splitting on nodes is repeated for each branch of the evolving decision tree, and the final tree, which would look like is shown in the following slide and calculation is left for practice.

DECISION TREE INDUCTION : ID3 WAY



FREQUENCY TABLE : CALCULATING A

- Calculation of entropy for each table and hence information gain for a particular split appears tedious (at least manually)!
- As an alternative, we discuss **a short-cut method** of doing the same using a special data structure called **Frequency Table**.
- **Frequency Table:** Suppose, $X = \{x_1, x_2, \dots, x_n\}$ denotes an attribute with $n - \text{different}$ attribute values in it. For a given database D , there are a set of k classes say $C = \{c_1, c_2, \dots, c_k\}$. Given this, a frequency table will look like as follows.

FREQUENCY TABLE : CALCULATING A

| | X | | | | | |
|-------|-------|-------|-------|----------|-------|-------|
| | x_1 | x_2 | | x_i | | x_n |
| c_1 | | | | | | |
| c_2 | | | | | | |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ |
| c_j | | | | f_{ij} | | |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ |
| c_k | | | | | | |

- Number of rows = Number of classes
- Number of columns = Number of attribute values
- f_{ij} = Frequency of x_i for class c_j

Assume that $|D| = N$, the number of total instances of D .

CALCULATION OF A USING FREQUENCY TABLE

Example 19.12 : OTPH Dataset

With reference to OPTH dataset, and for the attribute Age, the frequency table would look like

| | Age=1 | Age=2 | Age=3 | Row Sum |
|------------|-------|-------|-------|---------|
| Class 1 | 2 | 1 | 1 | 4 |
| Class 2 | 2 | 2 | 1 | 5 |
| Class 3 | 4 | 5 | 6 | 15 |
| Column Sum | 8 | 8 | 8 | 24 |

Column Sums

N=24

CALCULATION OF A USING FREQUENCY TABLE

- The weighted average entropy $E_X(D)$ then can be calculated from the frequency table following the
 - Calculate $V = f_{ij} \log_2 f_{ij}$ for all $i = 1, 2, \dots, k$
(Entry Sum) $j = 1, 2, \dots, n$ and $v_{ij} \neq 0$
 - Calculate $S = s_i \log_2 s_i$ for all $i = 1, 2, \dots, n$
(Column Sum) in the row of column sum
 - Calculate $E_X(D) = (-V + S)/N$

Example 19.13: OTPH Dataset

For the frequency table in Example 18.12, we have

$$V$$

$$= 2 \log 2 + 1 \log 1 + 1 \log 1 + 2 \log 2 + 2 \log 2 + 1 \log 1 + 4 \log 4 + 5 \log 5 + 6 \log 6$$
$$S = 8 \log 8 + 8 \log 8 + 8 \log 8$$

$$E_{Age}(OPTH) = 1.2867$$

PROOF OF EQUIVALENCE

- In the following, we prove the equivalence of the short-cut of entropy calculation using [Frequency Table](#).
- Splitting on an attribute A with n values produces n subsets of the training dataset D (of size $|D| = N$). The $j - th$ subset ($j = 1, 2, \dots, n$) contains all the instances for which the attribute takes its $j - th$ value. Let N_j denotes the number of instances in the $j - th$ subset. Then

$$\sum_{j=1}^n N_j = N$$

- Let f_{ij} denotes the number of instances for which the classification is c_i and attribute A takes its $j - th$ value. Then

$$\sum_{i=1}^k f_{ij} = N_j$$

PROOF OF EQUIVALENCE

Denoting E_j as the entropy of the $j - th$ subset, we have

$$E_j = - \sum_{i=1}^k \frac{f_{ij}}{N_j} \log_2 \frac{f_{ij}}{N_j}$$

Therefore, the weighted average entropy of the splitting attribute A is given by

$$\begin{aligned} E_A(D) &= \sum_{j=1}^n \frac{N_j}{N} \cdot E_j \\ &= - \sum_{j=1}^n \sum_{i=1}^k \frac{N_j}{N} \cdot \frac{f_{ij}}{N_j} \cdot \log_2 \frac{f_{ij}}{N_j} \\ &= - \sum_{j=1}^n \sum_{i=1}^k \frac{f_{ij}}{N_j} \cdot \log_2 \frac{f_{ij}}{N_j} \end{aligned}$$

PROOF OF EQUIVALENCE

$$= - \sum_{j=1}^n \sum_{i=1}^k \frac{f_{ij}}{N_j} \cdot \log_2 f_{ij} + \sum_{j=1}^n \sum_{i=1}^k \frac{f_{ij}}{N_j} \cdot \log_2 N_j$$

$$= - \sum_{j=1}^n \sum_{i=1}^k \frac{f_{ij}}{N_j} \cdot \log_2 f_{ij} + \sum_{j=1}^1 \frac{N_j}{N} \cdot \log_2 N_j$$

$$\because \sum_{i=1}^k f_{ij} = N_j$$

$$= \left(- \sum_{j=1}^n \sum_{i=1}^k f_{ij} \cdot \log_2 f_{ij} + \sum_{j=1}^n N_j \log_2 N_j \right) / N$$
$$= (-V + S) / N$$

where $V = \sum_{j=1}^n \sum_{i=1}^k f_{ij} \cdot \log_2 f_{ij}$ (Entries sum)

and $S = \sum_{j=1}^n N_j \log_2 N_j$ (Column Sum)

Hence, the equivalence is proved.

LIMITING VALUES OF INFORMATION GAIN

- The Information gain metric used in ID3 **always** should be positive or zero.
- It is always positive value because information is always gained (i.e., purity is improved) by splitting on an attribute.
- On the other hand, when a training set is such that if there are k classes, and the entropy of training set takes the largest value i.e., $\log_2 k$ (this occurs when the classes are balanced), then the information gain will be zero.

LIMITING VALUES OF INFORMATION GAIN

Example 19.14: Limiting values of Information gain

Consider a training set shown below.

| Data set <i>Table A</i> | | |
|-------------------------|---|-------|
| X | Y | Class |
| 1 | 1 | A |
| 1 | 2 | B |
| 2 | 1 | A |
| 2 | 2 | B |
| 3 | 2 | A |
| 3 | 1 | B |
| 4 | 2 | A |
| 4 | 1 | B |

| | | X | Table X | |
|-------|---|---|---------|---|
| | 1 | 2 | 3 | 4 |
| A | 1 | 1 | 1 | 1 |
| B | 1 | 1 | 1 | 1 |
| C.Sum | 2 | 2 | 2 | 2 |

Frequency table of X

| | Y | Table Y |
|-------|---|---------|
| | 1 | 2 |
| A | 2 | 2 |
| B | 2 | 2 |
| C.Sum | 4 | 4 |

Frequency table of Y

LIMITING VALUES OF INFORMATION GAIN

- Entropy of Table A is

$$E = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2 = 1 \text{ (The maximum entropy).}$$

- In this example, whichever attribute is chosen for splitting, each of the branches will also be balanced thus each with maximum entropy.
- In other words, information gain in both cases (i.e., splitting on X as well as Y) will be zero.

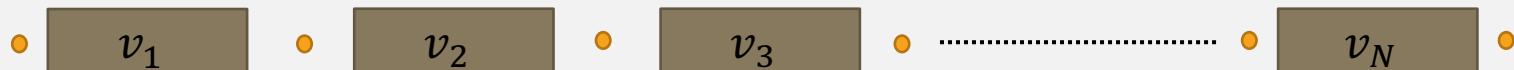
Note:

- The absence of information gain does not imply that there is no profit for splitting on the attribute.
- Even if it is chosen for splitting, ultimately it will lead to a final decision tree with the branches terminated by a leaf node and thus having an entropy of zero.
- Information gain can never be a negative value.

SPLITTING OF CONTINUOUS ATTRIBUTE VALUES

- In the foregoing discussion, we assumed that an attribute to be splitted is with a finite number of discrete values. Now, there is a great deal if the attribute is not so, rather it is a continuous-valued attribute.
- There are two approaches mainly to deal with such a case.

1. **Data Discretization:** All values of the attribute can be discretized into a finite number of group values and then split point can be decided at each boundary point of the groups.



So, if there are $n - groups$ of discrete values, then we have $(n + 1)$ split points.

SPLITTING OF CONTINUOUS ATTRIBUTE VALUES

2. Mid-point splitting: Another approach to avoid the data discretization.

- It sorts the values of the attribute and take the distinct values only in it.
- Then, the mid-point between each pair of adjacent values is considered as a split-point.



- Here, if n -distinct values are there for the attribute A , then we choose $n - 1$ split points as shown above.
- For example, there is a split point $s = \frac{v_i + v_{(i+1)}}{2}$ in between v_i and $v_{(i+1)}$
- For each split-point, we have two partitions: $A \leq s$ and $A > s$, and finally the point with maximum information gain is the desired split point for that attribute.

REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?



DATA ANALYTICS

Class # 20

Decision Tree Induction – CART & C4.5

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

ALGORITHM CART

CART ALGORITHM

- It is observed that information gain measure used in ID3 **is biased towards test with many outcomes**, that is, it prefers to select attributes having a large number of values.
- L. Breiman, J. Friedman, R. Olshen and C. Stone in 1984 proposed an algorithm to build a binary decision tree also called CART decision tree.
 - CART stands for **Classification and Regression Tree**
 - In fact, invented independently at the same time as ID3 (1984).
 - ID3 and CART are two cornerstone algorithms spawned a flurry of work on decision tree induction.
- CART is a technique that generates a **binary decision tree**; That is, unlike ID3, in CART, for each node only two children is created.
- ID3 uses Information gain as a measure to select the best attribute to be splitted, whereas CART do the same but using another measurement called **Gini index**. It is also known as **Gini Index of Diversity** and is denote as γ .

GINI INDEX OF DIVERSITY

Definition 20.1: Gini Index

Suppose, D is a training set with size $|D|$ and $C = \{c_1, c_2, \dots, c_k\}$ be the set of k classifications and $A = \{a_1, a_2, \dots, a_m\}$ be any attribute with m different values of it. Like entropy measure in ID3, CART proposes Gini Index (denoted by G) as the measure of impurity of D . It can be defined as follows.

$$G(D) = 1 - \sum_{i=1}^k p_i^2$$

where p_i is the probability that a tuple in D belongs to class c_i and p_i can be estimated as

$$p_i = \frac{|C_{i,D}|}{D}$$

where $|C_{i,D}|$ denotes the number of tuples in D with class c_i .

GINI INDEX OF DIVERSITY

-

Note

- $G(D)$ measures the “impurity” of data set D .
- The smallest value of $G(D)$ is zero
 - which it takes when all the classifications are same.
- It takes its largest value = $1 - \frac{1}{k}$
 - when the classes are evenly distributed between the tuples, that is the frequency of each class is $\frac{1}{k}$.

GINI INDEX OF DIVERSITY

Definition 20.2: Gini Index of Diversity

Suppose, a binary partition on A splits D into D_1 and D_2 , then the **weighted average Gini Index of splitting** denoted by $G_A(D)$ is given by

$$G_A(D) = \frac{|D_1|}{D} \cdot G(D_1) + \frac{|D_2|}{D} \cdot G(D_2)$$

This binary partition of D reduces the impurity and the reduction in impurity is measured by

$$\gamma(A, D) = G(D) - G_A(D)$$

GINI INDEX OF DIVERSITY AND CART

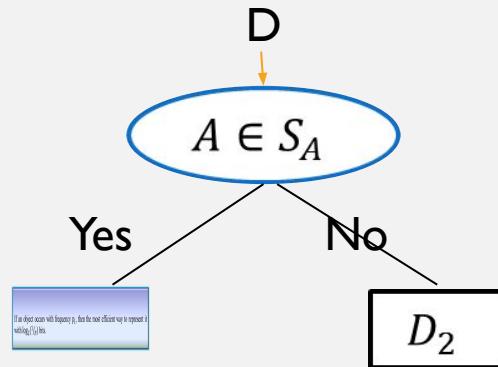
- This $\gamma(A, D)$ is called the Gini Index of diversity.
- It is also called as “impurity reduction”.
- The attribute that **maximizes** the reduction in impurity (or equivalently, has the **minimum value of $G_A(D)$**) is selected for the attribute to be splitted.

N-ARY ATTRIBUTE VALUES TO BINARY SPLITTING

- The CART algorithm considers a binary split for each attribute.
- We shall discuss how the same is possible for attribute with more than two values.
- **Case 1: Discrete valued attributes**
- Let us consider the case where A is a discrete-valued attribute having m discrete values a_1, a_2, \dots, a_m .
- To determine the best binary split on A , we examine all of the possible subsets say 2^A of A that can be formed using the values of A .
- Each subset $S_A \in 2^A$ can be considered as a binary test for attribute A of the form " $A \in S_A?$ ".

N-ARY ATTRIBUTE VALUES TO BINARY SPLITTING

- Thus, given a data set D , we have to perform a test for an attribute value A like



- This test is satisfied if the value of A for the tuples is among the values listed in S_A .
- If A has m distinct values in D , then there are 2^m possible subsets, out of which the empty subset $\{ \}$ and the power set $\{a_1, a_2, \dots, a_n\}$ should be excluded (as they really do not represent a split).
- Thus, there are $2^m - 2$ possible ways to form two partitions of the dataset D , based on the binary split of A .

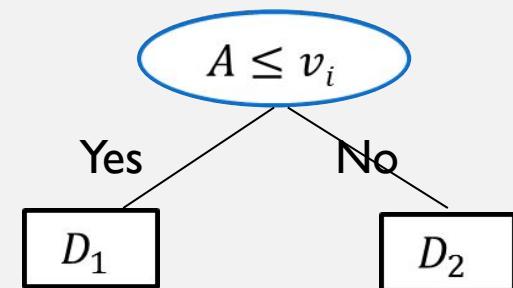
N-ARY ATTRIBUTE VALUES TO BINARY SPLITTING

Case2: Continuous valued attributes

- For a continuous-valued attribute, each possible split point must be taken into account.
- The strategy is similar to that followed in ID3 to calculate information gain for the continuous –valued attributes.
- According to that strategy, the mid-point between a_i and a_{i+1} , let it be v_i , then

Messages when ($m \neq 2^n$) ● Entropy of Messages when ($m \neq 2^n$) a_i ● Entropy of Messages when ($m \neq 2^n$)
independent of this record Entropy of Messages when ($m \neq 2^n$)

$$v_i = \frac{a_i + a_{i+1}}{2}$$



N-ARY ATTRIBUTE VALUES TO BINARY SPLITTING

- Each pair of (sorted) adjacent values is taken as a possible split-point say v_i .
- D_1 is the set of tuples in D satisfying $A \leq v_i$ and D_2 in the set of tuples in D satisfying $A > v_i$.
- The point giving the minimum Gini Index $G_A(D)$ is taken as the split-point of the attribute A .

Note

- The attribute A and either its splitting subset S_A (for discrete-valued splitting attribute) or split-point v_i (for continuous valued splitting attribute) together form the splitting criteria.

CART ALGORITHM : ILLUSTRATION

Example 20.1 : CART Algorithm

Suppose we want to build decision tree for the data set EMP as given in the table below.

Age
 Y : young
 M : middle-aged
 O : old

Salary
 L : low
 M : medium
 H : high

Job
 G : government
 P : private

Performance
 A : Average
 E : Excellent

Class : Select
 Y : yes
 N : no

| | Tuple# | Age | Salary | Job | Performance | Select |
|--|--------|-----|--------|-----|-------------|--------|
| | 1 | Y | H | P | A | N |
| | 2 | Y | H | P | E | N |
| | 3 | M | H | P | A | Y |
| | 4 | O | M | P | A | Y |
| | 5 | O | L | G | A | Y |
| | 6 | O | L | G | E | N |
| | 7 | M | L | G | E | Y |
| | 8 | Y | M | P | A | N |
| | 9 | Y | L | G | A | Y |
| | 10 | O | M | G | A | Y |
| | 11 | Y | M | G | E | Y |
| | 12 | M | M | P | E | Y |
| | 13 | M | H | G | A | Y |
| | 14 | O | M | P | E | N |

CART ALGORITHM : ILLUSTRATION

For the EMP data set,

$$\begin{aligned} G(EMP) &= 1 - \sum_{i=1}^2 p_i^2 \\ &= 1 - \left[\left(\frac{9}{14} \right)^2 + \left(\frac{5}{14} \right)^2 \right] \\ &= \mathbf{0.4592} \end{aligned}$$

Now let us consider the calculation of $G_A(EMP)$ for **Age**, **Salary**, **Job** and **Performance**.

CART ALGORITHM : ILLUSTRATION

Attribute of splitting: Age

The attribute age has three values, namely Y, M and O. So there are 6 subsets, that should be considered for splitting as:

$$G_{age_1}(D) = \frac{5}{14} * \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) + \frac{9}{14} \left(1 - \left(\frac{6}{14}\right)^2 - \left(\frac{8}{14}\right)^2\right) = \mathbf{0.4862}$$

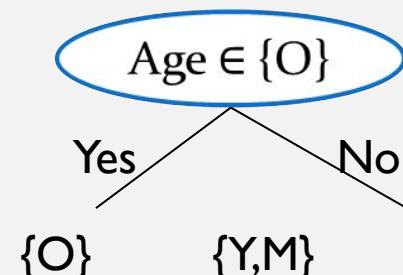
$$G_{age_2}(D) = ?$$

$$G_{age_3}(D) = ?$$

$$G_{age_4}(D) = G_{age_3}(D)$$

$$G_{age_5}(D) = G_{age_2}(D)$$

$$G_{age_6}(D) = G_{age_1}(D)$$



The best value of Gini Index while splitting attribute Age is $\gamma(Age_3, D) = \mathbf{0.3750}$

CART ALGORITHM : ILLUSTRATION

Attribute of Splitting: Salary

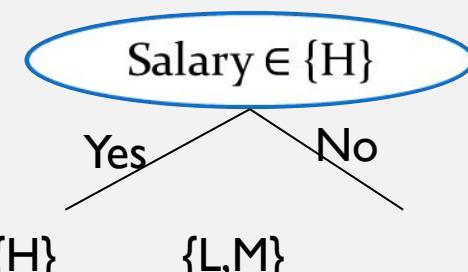
The attribute salary has three values namely L , M and H . So, there are 6 subsets, that should be considered for splitting as:

$$\begin{array}{llllll} \{L\} & \{M, H\} & \{M\} & \{L, H\} & \{H\} & \{L, M\} \\ sal_1' & sal_2' & sal_3' & sal_4' & sal_5' & sal_6 \end{array}$$

$$G_{sal_1}(D) = G_{sal_2}(D) = 0.3000$$

$$G_{sal_3}(D) = G_{sal_4}(D) = 0.3150$$

$$G_{sal_5}(D) = G_{sal_6}(D) = 0.4508$$



$$\gamma(salary_{(5,6)}, D) = 0.4592 - 0.4508 = 0.0084$$

CART ALGORITHM : ILLUSTRATION

Attribute of Splitting: job

Job being the binary attribute , we have

$$\begin{aligned}G_{job}(D) &= \frac{7}{14} G(D_1) + \frac{7}{14} G(D_2) \\&= \frac{7}{14} \left[1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 \right] + \frac{7}{14} \left[1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 \right] = ?\end{aligned}$$

$$\gamma(job, D) = ?$$

CART ALGORITHM : ILLUSTRATION

Attribute of Splitting: Performance

Job being the binary attribute , we have

$$G_{Performance}(D) = ?$$

$$\gamma(Performance, D) = ?$$

Out of these $\gamma(salary, D)$ gives the minimum value and hence, the attribute **Salary** would be chosen for splitting subset $\{M, H\}$ or $\{L\}$.

Note:

It can be noted that the procedure following “information gain” calculation (i.e. $\propto (A, D)$) and that of “impurity reduction” calculation (i.e. $\gamma(A, D)$) are near about.

CALCULATING Γ USING FREQUENCY TABLE

- We have learnt that splitting on an attribute gives a reduction in the average Gini Index of the resulting subsets (as it does for entropy).
- Thus, in the same way the average weighted Gini Index can be calculated using the same frequency table used to calculate information gain $\alpha(A, D)$, which is as follows.

The $G(D_j)$ for the j^{th} subset D_j

$$G(D_j) = 1 - \sum_{i=1}^k \left(\frac{f_{ij}}{|D_j|} \right)^2$$

CALCULATING Γ USING FREQUENCY TABLE

The average weighted Gini Index, $G_A(D_j)$ (assuming that attribute has m distinct values is)

$$\begin{aligned} G_A(D_j) &= \sum_{j=1}^k \frac{|D_j|}{|D_1|} \cdot G(D_j) \\ &= \sum_{j=1}^m \frac{|D_j|}{|D|} - \sum_{j=1}^m \sum_{i=1}^k \frac{|D_j|}{|D|} \cdot \left(\frac{f_{ij}}{|D_j|} \right)^2 \\ &= 1 - \frac{1}{|D|} \sum_{j=1}^m \frac{1}{D_j} \cdot \sum_{i=1}^k f_{ij}^2 \end{aligned}$$

The above gives a formula for m -attribute values; however, it can be fine tuned to subset of attributes also.

ILLUSTRATION: CALCULATING γ USING FREQUENCY TABLE

Example 20.2 : Calculating γ using frequency table of OPTH

Let us consider the frequency table for OPTH database considered earlier. Also consider the attribute A_1 with three values 1, 2 and 3. The frequency table is shown below.

| | 1 | 2 | 3 |
|------------|---|---|---|
| Class 1 | 2 | 1 | 1 |
| Class 2 | 2 | 2 | 1 |
| Class 3 | 4 | 5 | 6 |
| Column sum | 8 | 8 | 8 |

ILLUSTRATION: CALCULATING Γ USING FREQUENCY TABLE

Now we can calculate the value of Gini Index with the following steps:

1. For each non-empty column, form the sum of the squares of the values in the body of the table and divide by the column sum.
2. Add the values obtained for all columns and divided by $|D|$, the size of the database.
3. Subtract the total from 1.

As an example, with reference to the frequency table as mentioned just.

$$A_1 = 1 = \frac{(2^2 + 2^2 + 4^2)}{24} = 3.0$$

$$A_1 = 2 = \frac{(1^2 + 2^2 + 5^2)}{24} = 3.75$$

$$A_1 = 3 = \frac{(1^2 + 1^2 + 6^2)}{24} = 4.75$$

$$\text{So, } G_{A1}(D) = 1 - \frac{1+3.75+4.75}{24} = 0.5208$$

Illustration: Calculating γ using Frequency Table

Thus, the reduction in the value of Gini Index on splitting attribute A_1 is

$$\gamma(A_1, D) = 0.5382 - 0.5208 = 0.0174$$

where $G(D) = 0.5382$

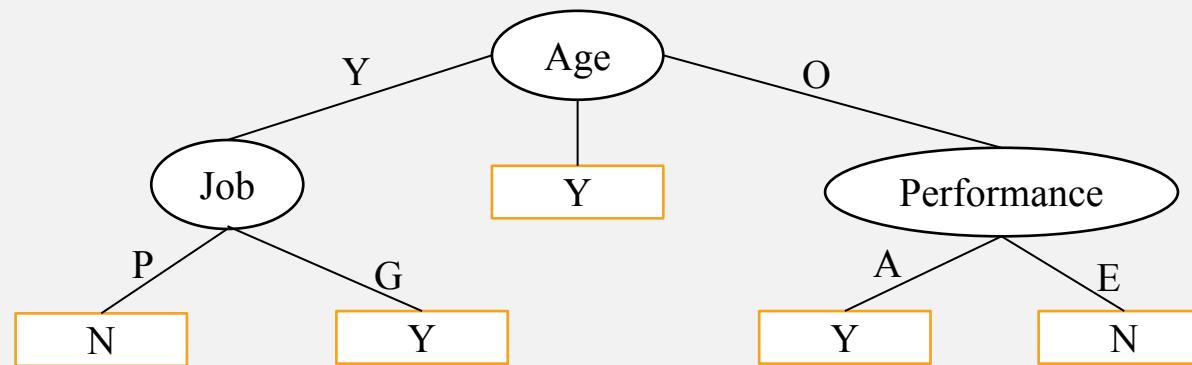
The calculation can be extended to other attributes in the OTPH database and is left as an exercise.



DECISION TREES WITH ID3 AND CART ALGORITHMS

Example 20.3 : Comparing Decision Trees of EMP Data set

Compare two decision trees obtained using ID3 and CART for the EMP dataset. The decision tree according to ID3 is given for your ready reference (subject to the verification)



Decision Tree using ID3

?

Decision Tree using CART

ALGORITHM C4.5

ALGORITHM C4.5 : INTRODUCTION

- J. Ross Quinlan, a researcher in machine learning developed a decision tree induction algorithm in 1984 known as ID3 (Iterative Dichotometer 3).
- Quinlan later presented C4.5, a successor of ID3, addressing some limitations in ID3.
- ID3 uses information gain measure, which is, in fact **biased towards splitting attribute having a large number of outcomes**.
- For example, if an attribute has distinct values for all tuples, then it would result in a large number of partitions, each one containing just one tuple.
- In such a case, note that each partition is pure, and hence the purity measure of the partition, that is $E_A(D) = 0$

ALGORITHM C4.5 : INTRODUCTION

Example 20.4 : Limitation of ID3

In the following, each tuple belongs to a unique class. The splitting on A is shown.

| A | ----- | class |
|-------|-------|-------|
| a_1 | | |
| a_2 | | |
| ⋮ | | |
| a_j | | |
| ⋮ | | |
| a_n | | |

a_1 ----- |
 a_2 ----- |
⋮
 a_j ----- |
 a_n ----- |

$E(D_j) = l \log_2 l = 0$

$$E_A(D) = \sum_{j=1}^n \frac{|D_j|}{|D|} \cdot E(D_j) = \sum_{j=1}^n \frac{1}{|D|} \cdot 0 = 0$$

Thus, $\alpha(A, D) = E(D) - E_A(D)$ is maximum in such a situation.

ALGORITHM: C 4.5 : INTRODUCTION

- Although, the previous situation is an extreme case, intuitively, we can infer that **ID3 favours splitting attributes having a large number of values**
 - compared to other attributes, which have a less variations in their values.
- Such a partition appears to be useless for classification.
- This type of problem is called **overfitting problem**.

Note:

Decision Tree Induction Algorithm ID3 may suffer from overfitting problem.

ALGORITHM: C 4.5 : INTRODUCTION

- The overfitting problem in ID3 is due to the measurement of information gain.
- In order to reduce the effect of the use of the bias due to the use of information gain, C4.5 uses a different measure called **Gain Ratio**, denoted as β .
- Gain Ratio is a kind of normalization to information gain using a **split information**.

ALGORITHM: C4.5 : GAIN RATIO

Definition 20.3: Gain Ratio

The gain ratio can be defined as follows. We first define **split information** $E_A^*(D)$ as

$$E_A^*(D) = - \sum_{j=1}^m \frac{|D_j|}{|D|} \cdot \log \frac{|D_j|}{|D|}$$

Here, m is the number of distinct values in A .

The gain ratio is then defined as $\beta(A, D) = \frac{\alpha(A, D)}{E_A^*(D)}$, where $\alpha(A, D)$ denotes the information gain on splitting the attribute A in the dataset D .

PHYSICAL INTERPRETATION OF $E_A^*(D)$

Split information $E_A^*(D)$

- The value of split information depends on
 - the number of (distinct) values an attribute has and
 - how uniformly those values are distributed.
- In other words, it represents the **potential information** generated by splitting a data set D into m partitions, corresponding to the m outcomes of on attribute A .
- Note that for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D .

PHYSICAL INTERPRETATION OF $E_A^*(D)$

Example 20.5 : Split information $E_A^*(D)$

- To illustrate $E_A^*(D)$, let us examine the case where there are 32 instances and splitting an attribute A which has a_1, a_2, a_3 and a_4 sets of distinct values.
- Distribution 1 : Highly non-uniform distribution of attribute values

| Frequency | 32 | 0 | 0 | 0 |
|-----------|----|---|---|---|

$$E_A^*(D) = - \frac{32}{32} \log_2 \left(\frac{32}{32} \right) = -\log_2 1 = 0$$

- Distribution 2

| Frequency | 16 | 16 | 0 | 0 |
|-----------|----|----|---|---|

$$E_A^*(D) = - \frac{16}{32} \log_2 \left(\frac{16}{32} \right) - \frac{16}{32} \log_2 \left(\frac{16}{32} \right) = \log_2 2 = 1$$

PHYSICAL INTERPRETATION OF $E_A^*(D)$

Distribution 3

| | | | | |
|-----------|----|---|---|---|
| | | | | |
| Frequency | 16 | 8 | 8 | 0 |

$$E_A^*(D) = -\frac{16}{32} \log_2\left(\frac{16}{32}\right) - \frac{8}{32} \log_2\left(\frac{8}{32}\right) - \frac{8}{32} \log_2\left(\frac{8}{32}\right) = 1.5$$

- Distribution 4

| | | | | |
|-----------|----|---|---|---|
| | | | | |
| Frequency | 16 | 8 | 4 | 4 |

$$E_A^*(D) = 1.75$$

- Distribution 5: Uniform distribution of attribute values

| | | | | |
|-----------|---|---|---|---|
| | | | | |
| Frequency | 8 | 8 | 8 | 8 |

$$E_A^*(D) = \left(-\frac{8}{32} \log_2\left(\frac{8}{32}\right)\right) * 4 = -\log_2\left(\frac{1}{4}\right) = 2.0$$

PHYSICAL INTERPRETATION OF $E_A^*(D)$

- In general, if there are m attribute values, each occurring equally frequently, then the split information is $\log_2 m$.
- Based on the Example 20.5, we can summarize our observation on split information as under:
 - Split information is 0 when there is a single attribute value. It is a trivial case and implies *the minimum possible value of split information*.
 - For a given data set, when instances are uniformly distributed with respect to the attribute values, split information increases as the number of different attribute values increases.
 - The maximum value of split information occur when there are many possible attribute values, all are equally frequent.

Note:

- Split information varies between 0 and $\log_2 m$ (both inclusive)

PHYSICAL INTERPRETATION OF $\beta(A, B)$

- Information gain signifies how much information will be gained on partitioning the values of attribute A
 - Higher information gain means splitting of A is more desirable.
 - On the other hand, split information forms the denominator in the gain ratio formula.
 - This implies that higher the value of split information is, lower the gain ratio.
 - In turns, it decreases the information gain.
- Further, information gain is large when there are many distinct attribute values.
 - When many distinct values, split information is also a large value.
 - This way split information reduces the value of gain ratio, thus resulting a balanced value for information gain.
- Like information gain (in ID3), the attribute with the maximum gain ratio is selected as the splitting attribute in C4.5.

CALCULATION OF β USING FREQUENCY TABLE

- The frequency table can be used to calculate the gain ratio for a given data set and an attribute.
- We have already learned the calculation of information gain using Frequency Table.
- To calculate gain ratio, in addition to information gain, we are also to calculate split information.
- This split information can be calculated from frequency table as follows.
- For each non-zero column sum say s_j contribute $|D_j|$ for the j -th column (i.e., the j -th value of the attribute). Thus the split information is

$$E_A^*(D) = - \sum_{j=1}^m \frac{s_j}{|D|} \log_2 \frac{s_j}{|D|}$$

If there are m -columns in the frequency table.

Practice:

Using Gain ratio as the measurement of splitting attributes, draw the decision trees for OPTH and EMP data sets. Give calculation of gain ratio at each node.

SUMMARY OF DECISION TREE INDUCTION ALGORITHMS

- We have learned the building of a decision tree given a training data.
 - The decision tree is then used to classify a test data.
- For a given training data D , the important task is to build the decision tree so that:
 - All test data can be classified accurately
 - The tree is balanced and with as minimum depth as possible, thus the classification can be done at a faster rate.
- In order to build a decision tree, several algorithms have been proposed. These algorithms differ from the chosen splitting criteria, so that they satisfy the above mentioned objectives as well as the decision tree can be induced with minimum time complexity. We have studied three decision tree induction algorithms namely ID3, CART and C4.5. A summary of these three algorithms is presented in the following table.

TABLE 20.6

| Algorithm | Splitting Criteria | Remark |
|-----------|--------------------|--------|
| ID3 | | |

| Algorithm | Splitting Criteria | Remark |
|-----------|--------------------|--------|
| CART | | |

| Algorithm | Splitting Criteria | Remark |
|-----------|--------------------|--------|
| C4.5 | | |

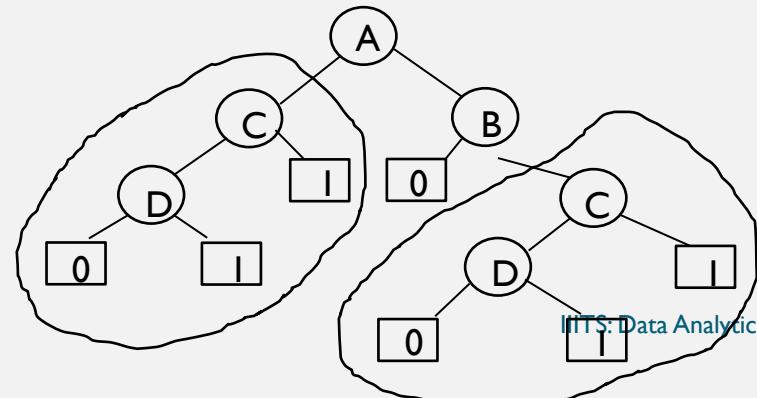
In addition to this, we also highlight few important characteristics of decision tree induction algorithms in the following.

NOTES ON DECISION TREE INDUCTION ALGORITHMS

1. **Optimal Decision Tree:** Finding an optimal decision tree is an NP-complete problem. Hence, decision tree induction algorithms **employ a heuristic based approach** to search for the best in a large search space. Majority of the algorithms follow a greedy, top-down recursive divide-and-conquer strategy to build decision trees.
2. **Missing data and noise:** Decision tree induction algorithms are quite robust to the data set with missing values and presence of noise. However, proper data pre-processing can be followed to nullify these discrepancies.
3. **Redundant Attributes:** The presence of redundant attributes does not adversely affect the accuracy of decision trees. It is observed that if an attribute is chosen for splitting, then another attribute which is redundant is unlikely to be chosen for splitting.
4. **Computational complexity:** Decision tree induction algorithms are computationally inexpensive, in particular, when the sizes of training sets are large. Moreover, once a decision tree is known, classifying a test record is extremely fast, with a worst-case time complexity of $O(d)$, where d is the maximum depth of the tree.

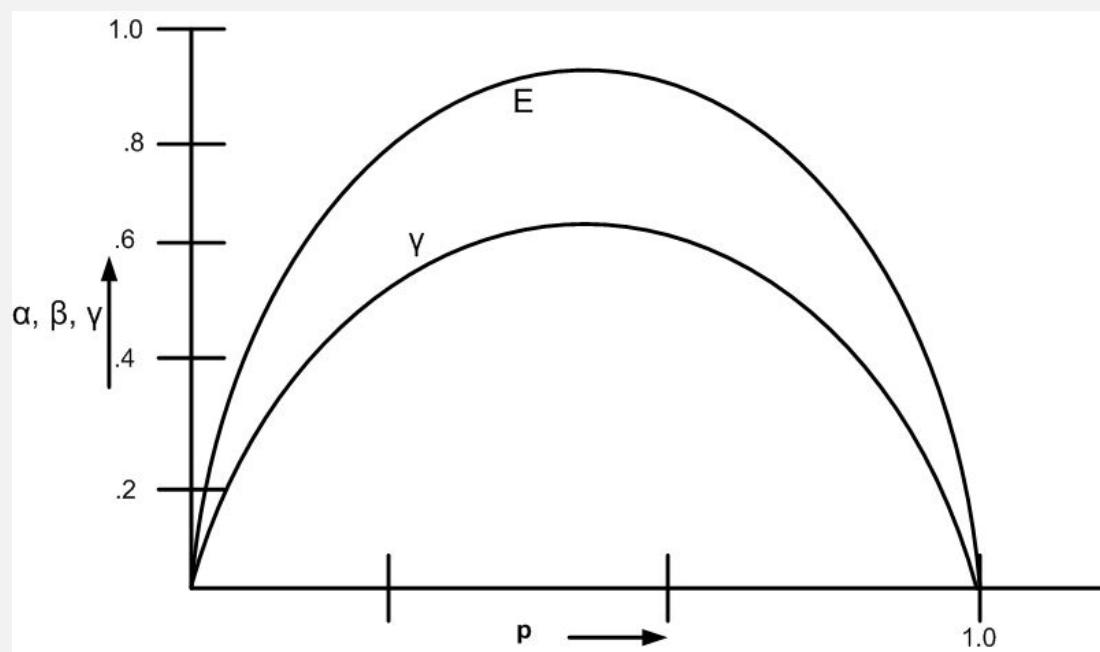
Notes on Decision Tree Induction algorithms

5. **Data Fragmentation Problem:** Since the decision tree induction algorithms employ a top-down, recursive partitioning approach, the number of tuples becomes smaller as we traverse down the tree. At a time, the number of tuples may be too small to make a decision about the class representation, such a problem is known as the data fragmentation. To deal with this problem, further splitting can be stopped when the number of records falls below a certain threshold.
6. **Tree Pruning:** A sub-tree can replicate two or more times in a decision tree (see figure below). This makes a decision tree unambiguous to classify a test record. To avoid such a sub-tree replication problem, all sub-trees except one can be pruned from the tree.



Notes on Decision Tree Induction algorithms

7. **Decision tree equivalence:** The different splitting criteria followed in different decision tree induction algorithms have little effect on the performance of the algorithms. This is because the different heuristic measures (such as information gain (α), Gini index (γ) and Gain ratio (β) are quite consistent with each other); also see the figure below.



REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 22

Sensitivity Analysis

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

TOPICS COVERED IN THIS PRESENTATION

- Introduction
- Estimation Strategies
- Accuracy Estimation
- Error Estimation
- Statistical Estimation
- Performance Estimation
- ROC Curve

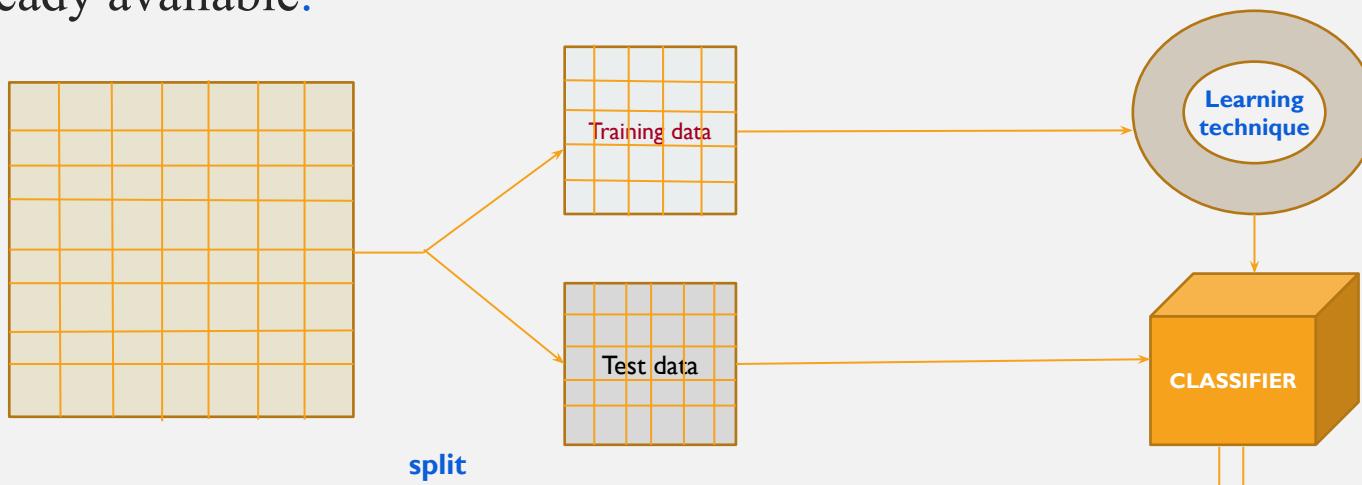
INTRODUCTION

- A classifier is used to predict an outcome of a test data
 - Such a prediction is useful in many applications
 - Business forecasting, cause-and-effect analysis, etc.
 - A number of classifiers have been evolved to support the activities.
 - Each has their own merits and demerits
- There is a need to estimate the accuracy and performance of the classifier with respect to few controlling parameters in data sensitivity
- As a task of sensitivity analysis, we have to focus on
 - Estimation strategy
 - Metrics for measuring accuracy
 - Metrics for measuring performance

Estimation Strategy

PLANNING FOR ESTIMATION

- Using some “**training data**”, building a classifier based on certain principle is called “**learning a classifier**”.
- After building a classifier and before using it for classification of unseen instance, we have to validate it using some “**test data**”.
- Usually training data and test data are outsourced from a large pool of data already available.



ESTIMATION STRATEGIES

- Accuracy and performance measurement should follow a strategy. As the topic is important, many strategies have been advocated so far. Most widely used strategies are
 - Holdout method
 - Random subsampling
 - Cross-validation
 - Bootstrap approach

HOLDOUT METHOD

- This is a basic concept of estimating a prediction.
 - Given a dataset, it is partitioned into **two disjoint sets** called **training set** and **testing set**.
 - Classifier is **learned** based on the training set and get **evaluated** with testing set.
 - Proportion of training and testing sets is at the discretion of analyst; typically **1:1 or 2:1**, and there is **a trade-off between these sizes** of these two sets.
 - If the training set is **too large**, then **model may be good enough**, but **estimation may be less reliable** due to small testing set and vice-versa.

RANDOM SUBSAMPLING

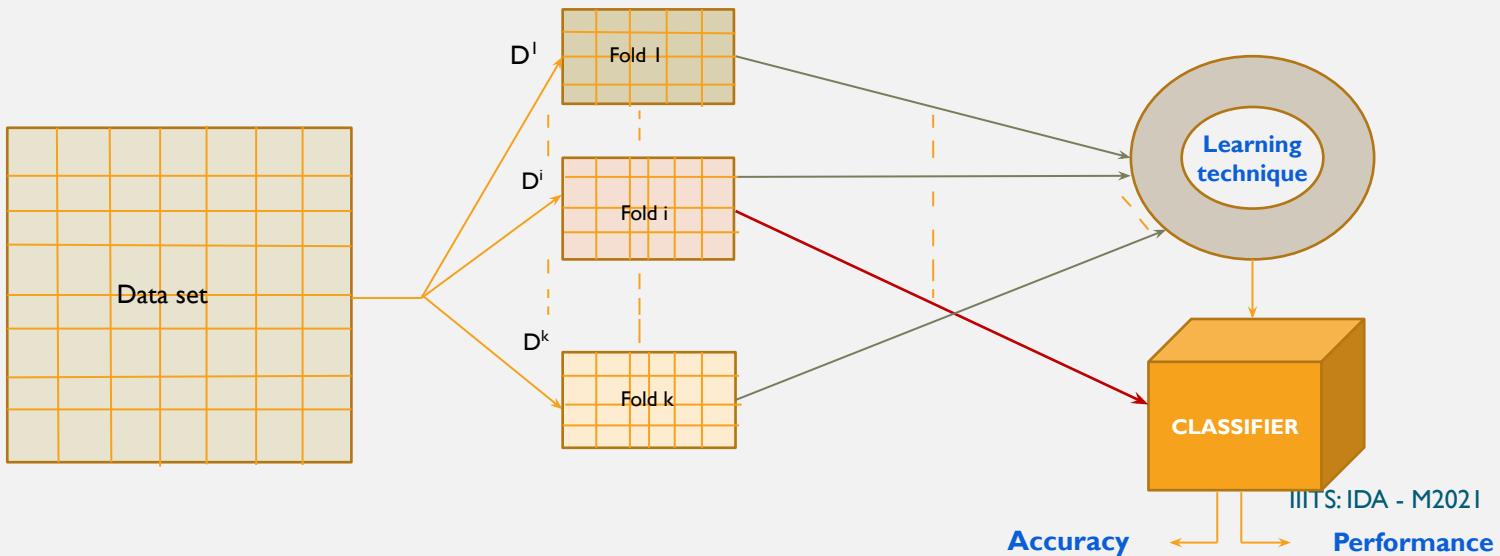
- It is a variation of Holdout method to overcome the drawback of over-presenting a class in one set thus under-presenting it in the other set and vice-versa.
- In this method, Holdout method is repeated k times, and in each time, two disjoint sets are chosen at random with a predefined sizes.
- Overall estimation is taken as the average of estimations obtained from each iteration.

CROSS-VALIDATION

- The main drawback of Random subsampling is, it does not have control over the number of times each tuple is used for training and testing.
- Cross-validation is proposed to overcome this problem.
- There are two variations in the cross-validation method.
 - k-fold cross-validation
 - N -fold cross-validation

K-FOLD CROSS-VALIDATION

- Dataset consisting of N tuples is divided into k (usually, 5 or 10) equal, mutually exclusive parts or folds (D_1, D_2, \dots, D_k), and if N is not divisible by k , then the last part will have fewer tuples than other ($k-1$) parts.
- A series of k runs is carried out with this decomposition, and in i^{th} iteration D_i is used as test data and other folds as training data
 - Thus, each tuple is used same number of times for training and once for testing.
- Overall estimate is taken as the average of estimates obtained from each iteration.



N-FOLD CROSS-VALIDATION

- In k -fold cross-validation method, $\frac{k-1}{N}$ part of the given data is used in training with k -tests.
- N -fold cross-validation is an **extreme case** of k -fold cross validation, often known as “**Leave-one-out**” cross-validation.
- Here, dataset is divided into as many folds as there are instances; thus, all most each tuple forming a training set, building N classifiers.
- In this method, therefore, N classifiers are built from $N-1$ instances, and each tuple is used to classify a single test instances.
- Test sets are mutually exclusive and effectively cover the entire set (in sequence). This is as if **trained by entire data as well as tested by entire data** set.
- Overall estimation is then averaged out of the results of N classifiers.

N-FOLD CROSS-VALIDATION : ISSUE

- So far the estimation of accuracy and performance of a classifier model is concerned, the *N-fold cross-validation is comparable to the others* we have just discussed.
- The drawback of *N-fold cross validation* strategy is that it is **computationally expensive**, as here we have to repeat the run N times; this is particularly true when data set is large.
- In practice, the **method is extremely beneficial with very small data set** only, where as much data as possible to need to be used to train a classifier.

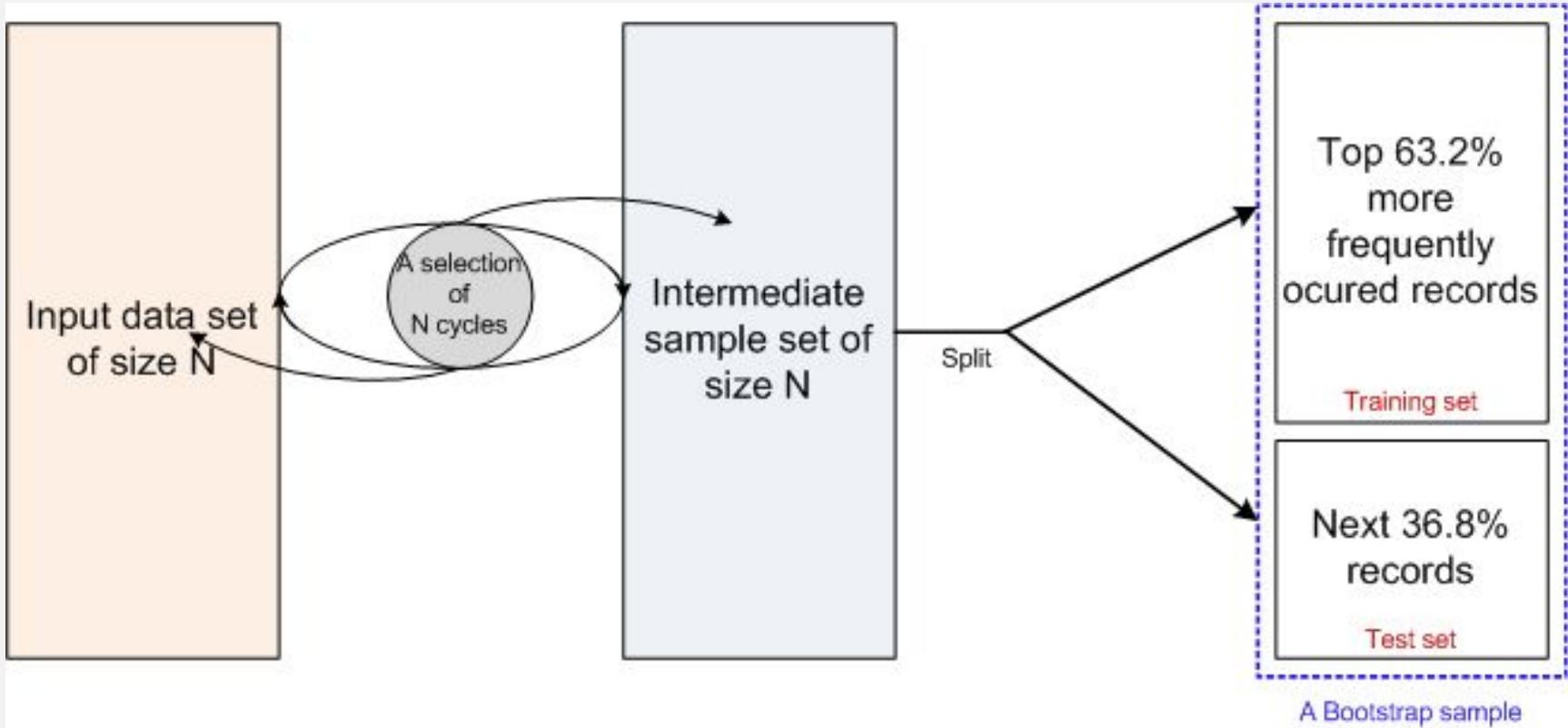
BOOTSTRAP METHOD

- The Bootstrap method is a variation of **repeated version of Random sampling** method.
- The method suggests the **sampling of training records with replacement**.
 - Each time a record is selected for training set, is put back into the original pool of records, so that it is equally likely to be redrawn in the next run.
 - In other words, the Bootstrap method samples the given data set **uniformly with replacement**.
- The rational of having this strategy is that let some records be occur **more than once** in the samples of both training as well as testing.
- **What is the probability that a record will be selected more than once?**

BOOTSTRAP METHOD

- Suppose, we have given a data set of N records. The data set is sampled N times with replacement, resulting in a bootstrap sample (i.e., training set) of I samples.
 - Note that the entire runs are called a bootstrap sample in this method.
- There are certain chance (i.e., probability) that a particular tuple occurs **one or more** times in the training set
 - If they do not appear in the training set, then they will end up in the test set.
 - Each tuple has a probability of being selected $\frac{1}{N}$ (and the probability of not being selected is $\left(1 - \frac{1}{N}\right)$).
 - We have to select N times, so the probability that a record will not be chosen during the whole run is $\left(1 - \frac{1}{N}\right)^N$
 - Thus, the probability that a record is chosen by a bootstrap sample is $1 - \left(1 - \frac{1}{N}\right)^N$
 - For a large value of N , it can be proved that $\left(1 - \frac{1}{N}\right)^N \approx e^{-1}$
 - **Thus, the probability that a record chosen in a bootstrap sample is $1 - e^{-1} = 0.632$**

BOOTSTRAP METHOD : IMPLICATION



- This is why, the Bootstrap method is also known as 0.632 bootstrap method

Accuracy Estimation

ACCURACY ESTIMATION

- We have learned how a classifier system can be tested. Next, we are to learn the metrics with which a classifier should be estimated.
- There are mainly two things to be measured for a given classifier
 - Accuracy
 - Performance
- **Accuracy estimation**
 - If N is the number of instances with which a classifier is tested and p is the number of correctly classified instances, the accuracy can be denoted as

$$\epsilon = \frac{p}{N}$$

- Also, we can say the **error rate** (i.e., is misclassification rate) denoted by $\bar{\epsilon}$ is denoted by
$$\bar{\epsilon} = 1 - \epsilon$$

ACCURACY : TRUE AND PREDICTIVE

- Now, this accuracy may be **true** (or absolute) accuracy or **predicted** (or optimistic) accuracy.
- True accuracy** of a classifier is the accuracy when the classifier is tested with **all possible unseen instances** in the given classification space.
 - However, the number of possible unseen instances is potentially very large (if it is not infinite)
 - For example, classifying a hand-written character
 - Hence, measuring the true accuracy beyond the dispute is impractical.
- Predictive accuracy** of a classifier is an **accuracy estimation** for a given test data (which are mutually exclusive with training data).
 - If the predictive accuracy for test set is ϵ and if we test the classifier with a different test set it is very likely that a different accuracy would be obtained.
 - The predictive accuracy when estimated with a given test set it should be acceptable without any objection

PREDICTIVE ACCURACY

Example 21.1 : Universality of predictive accuracy

- Consider a classifier model M^D developed with a training set D using an algorithm M.
- Two predictive accuracies when M^D is estimated with two different training sets T_1 and T_2 are

$$(M^D)_{T_1} = 95\%$$

$$(M^D)_{T_2} = 70\%$$

- Further, assume the size of T_1 and T_2 are

$$|T_1| = 100 \text{ records}$$

$$|T_2| = 5000 \text{ records.}$$

- Based on the above mentioned estimations, neither estimation is acceptable beyond doubt.

PREDICTIVE ACCURACY

- With the above-mentioned issue in mind, researchers have proposed two heuristic measures
 - Error estimation using **Loss Functions**
 - Statistical Estimation using **Confidence Level**
- In the next few slides, we will discuss about the two estimations

Error Estimation using Loss Functions

- Let T be a matrix comprising with N test tuples

$$\begin{bmatrix} X_1 & y_1 \\ X_2 & y_2 \\ \vdots & \vdots \\ X_N & y_N \end{bmatrix} N \times (n+1)$$

where X_i ($i = 1, 2, \dots, N$) is the n -dimensional test tuples with associated outcome y_i .

- Suppose, corresponding to (X_i, y_i) , classifier produces the result (X_i, y'_i)
- Also, assume that $(y_i - y'_i)$ denotes a difference between y_i and y'_i (following certain difference (or similarity), (e.g., $(y_i - y'_i) = 0$, if there is a match else 1)
- The two loss functions measure the error between y_i (the actual value) and y'_i (the predicted value) are

$$\text{Absolute error: } |y_i - y'_i|$$

$$\text{Squared error: } |y_i - y'_i|^2$$

Error Estimation using Loss Functions

- Based on the two loss functions, the test error (rate) also called **generalization error**, is defined as the average loss over the test set T. The following two measures for test errors are

Mean Absolute Error (MAE):

$$\frac{\sum_{i=1}^N |y_i - y_i'|}{N}$$

Mean Squared Error(MSE):

$$\frac{\sum_{i=1}^N (y_i - y_i')^2}{N}$$

- Note that, MSE aggregates the presence of outlier.
- In addition to the above, a relative error measurement is also known. In this measure, the error is measured relative to the mean value \tilde{y} calculated as the mean of y_i ($i = 1, 2, \dots, N$) of the training data say D. Two measures are

Relative Absolute Error (RAE):

$$\frac{\sum_{i=1}^N |y_i - y_i'|}{\sum_{i=1}^N |y_i - \tilde{y}|}$$

Relative Squared Error (RSE):

$$\frac{\sum_{i=1}^N (y_i - y_i')^2}{\sum_{i=1}^N (y_i - \tilde{y})^2}$$

Statistical Estimation using Confidence Level

- In fact, if we know the value of predictive accuracy, say ϵ , then we can guess the true accuracy within a certain range given a **confidence level**.
- **Confidence level:** The concept of “confidence level ” can be better understood with the following two experiments, related to tossing a coin.
- **Experiment 1:** When a coin is tossed, there is a probability that the head will occur. We have to experiment the value for this probability value. A simple experiment is that the coin is tossed many times and both numbers of heads and tails are recorded.

| N=10 | | N=50 | | N=100 | | N=250 | | N=500 | | N=1000 | |
|------|------|------|------|-------|------|-------|------|-------|------|--------|------|
| H | T | H | T | H | T | H | T | H | T | H | T |
| 3 | 7 | 29 | 21 | 54 | 46 | 135 | 115 | 241 | 259 | 490 | 510 |
| 0.30 | 0.70 | 0.58 | 0.42 | 0.54 | 0.46 | 0.54 | 0.46 | 0.48 | 0.42 | 0.49 | 0.51 |

- Thus, we can say that $p \rightarrow 0.5$ after a large number of trials in each experiment.

Statistical Estimation using Confidence Level

- **Experiment 2:** A similar experiment but with different counting is conducted to learn the probability that a coin is flipped its head 20 times out of 50 trials. This experiment is popularly known as Bernoulli's trials. It can be stated as follows.

$$P(X = v) = \binom{N}{v} p^v (1 - p)^{N-v}$$

- where N = Number of trials
- v = Number of outcomes that an event occurs.
- p = Probability that the event occur
- Thus, if $p = 0.5$, then $P(X = 20) = \binom{50}{20} 0.5^{20} \times 0.5^{30} = 0.0419$
- **Note:**
 - Also, we may note the following
 - Mean = $N \times p = 50 \times 0.5 = 25$ and Variance = $p \times (1-p) \times N = 50 \times 0.5 \times 0.5 = 12.5$

Statistical Estimation using Confidence Level

- The task of predicting the class labels of test records can also be considered as a binomial experiment, which can be understood as follows. Let us consider the following.
 - N = Number of records in the test set.
 - n = Number of records predicted correctly by the classifier.
 - $\epsilon = n/N$, the observed accuracy (it is also called the empirical accuracy).
 - $\tilde{\epsilon}$ = the true accuracy.
- Let τ_{α}^L and τ_{α}^U denotes the lower and upper bound of a confidence level α . Then the confidence interval for α is given by
$$P \left(\tau_{\alpha}^L \leq \frac{\epsilon - \tilde{\epsilon}}{\sqrt{\epsilon(1-\epsilon)/N}} \leq \tau_{\alpha}^U \right) = \alpha$$
- If τ_{α} is the mean of τ_{α}^L and τ_{α}^U , then we can write

$$\tilde{\epsilon} = \epsilon \pm \tau_{\alpha} \times \sqrt{\epsilon(1-\epsilon)/N}$$

Statistical Estimation using Confidence Level

$$\tilde{\epsilon} = \epsilon \pm \tau_{\alpha} \times \sqrt{\epsilon(1-\epsilon)/N}$$

- A table of τ_{α} with different values of α can be obtained from any book on statistics. A small part of the same is given below.

| | 0.5 | 0.7 | 0.8 | 0.9 | 0.95 | 0.98 | 0.99 |
|--|------|------|------|------|------|------|------|
| | 0.67 | 1.04 | 1.28 | 1.65 | 1.96 | 2.33 | 2.58 |

- Thus, given a confidence level α , we shall be able to know the value of τ_{α} and hence the true accuracy ($\tilde{\epsilon}$), if we have the value of the observed accuracy (ϵ).
- Thus, knowing a test data set of size N , it is possible to estimate the true accuracy!

Statistical Estimation using Confidence Level

Example 21.2: True accuracy from observed accuracy

A classifier is tested with a test set of size 100. Classifier predicts 80 test tuples correctly. We are to calculate the following.

- a) Observed accuracy
- b) Mean error rate
- c) Standard error
- d) True accuracy with confidence level 0.95.

Solution:

- a) The observed accuracy(ϵ) = $80/100 = 0.80$ So error (p) = 0.2
- b) Mean error rate = $p \times N = 0.2 \times 100 = 20$
- c) Standard error rate (σ) = $\sqrt{\epsilon(1-\epsilon)/N} = \sqrt{\frac{0.8 \times 0.2}{100}} = 0.04$
- d) $\tilde{\epsilon} = \epsilon \pm \tau_\alpha \times \sqrt{\epsilon(1-\epsilon)/N} = 0.8 \pm 0.04 \times 1.96 = 0.7216$ with $\tau_\alpha=1.96$ and $\alpha = 0.95$.

Statistical Estimation using Confidence Level

Note:

- Suppose, a classifier is tested k times with k different test sets. If ϵ_i denotes the predicted accuracy when tested with test set N_i in the i -th run ($1 \leq i \leq k$), then the overall predicted accuracy is

$$\epsilon = \sum_{i=1}^k \frac{\epsilon_i \times N_i}{\sum N_i}$$

Thus, ϵ is the weighted average of ϵ_i values. The standard error and true accuracy at a confidence α are

$$\text{Standard error} = \sqrt{\epsilon (1 - \epsilon) / \sum_{i=1}^k N_i}$$

$$\text{True accuracy} = \epsilon \pm \sqrt{\frac{\epsilon(1-\epsilon)}{\sum_{i=1}^k N_i}} \times \tau_\alpha$$

REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 23

Sensitivity Analysis – Performance Estimation

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

Performance Estimation

PERFORMANCE ESTIMATION OF A CLASSIFIER

- Predictive accuracy works fine, when the **classes are balanced**
 - That is, every class in the data set are equally important
- In fact, data sets with imbalanced class distributions are quite common in many real life applications
- When the classifier classified a test data set with imbalanced class distributions then, predictive accuracy on its own is not a reliable indicator of a classifier's effectiveness.

Example 22.1: Effectiveness of Predictive Accuracy

- Given a data set of stock markets, we are to classify them as “good” and “worst”. Suppose, in the data set, out of 100 entries, 98 belong to “good” class and only 2 are in “worst” class.
 - With this data set, if classifier’s predictive accuracy is 0.98, a very high value!
 - Here, there is a high chance that 2 “worst” stock markets may incorrectly be classified as “good”
 - On the other hand, if the predictive accuracy is 0.02, then none of the stock markets may be classified as “good”

PERFORMANCE ESTIMATION OF A CLASSIFIER

- Thus, when the classifier classified a test data set with imbalanced class distributions, then predictive accuracy on its own is not a reliable indicator of a classifier's effectiveness.
- This necessitates an alternative metrics to judge the classifier.
- Before exploring them, we introduce the concept of **Confusion matrix**.

CONFUSION MATRIX

- A confusion matrix for a two classes (+, -) is shown below.

| | C ₁ | C ₂ |
|----------------|----------------|----------------|
| C ₁ | True positive | False negative |
| C ₂ | False positive | True negative |

| | + | - |
|---|----|----|
| + | ++ | +- |
| - | -+ | -- |

- There are four quadrants in the confusion matrix, which are symbolized as below.
 - True Positive** (TP: f_{++}) : The number of instances that were positive (+) and correctly classified as positive (+v).
 - False Negative** (FN: f_{+}): The number of instances that were positive (+) and incorrectly classified as negative (-). It is also known as **Type 2 Error**.
 - False Positive (FP: f_{+}): The number of instances that were negative (-) and incorrectly classified as (+). This also known as **Type 1 Error**.
 - True Negative** (TN: f_{-}): The number of instances that were negative (-) and correctly classified as (-).

CONFUSION MATRIX

Note:

- $N_p = \text{TP}(f_{++}) + \text{FN}(f_{+-})$
= is the total number of positive instances.
- $N_n = \text{FP}(f_{-+}) + \text{TN}(f_{--})$
= is the total number of negative instances.
- $N = N_p + N_n$
= is the total number of instances.
- $(\text{TP} + \text{TN})$ denotes the number of correct classification
- $(\text{FP} + \text{FN})$ denotes the number of errors in classification.
- For a perfect classifier $\text{FP} = \text{FN} = 0$, that is, there would be no Type 1 or Type 2 errors.

CONFUSION MATRIX

Example 22.2: Confusion matrix

A classifier is built on a dataset regarding Good and Worst classes of stock markets. The model is then tested with a test set of 10000 unseen instances. The result is shown in the form of a confusion matrix. The result is self explanatory.

| Class | Good | Worst | Total |
|--------------|-------------|-------------|--------------|
| Good | 6954 | 46 | 7000 |
| Worst | 412 | 2588 | 3000 |
| Total | 7366 | 2634 | 10000 |

Predictive accuracy?

CONFUSION MATRIX FOR MULTICLASS CLASSIFIER

- Having m classes, confusion matrix is a table of size $m \times m$, where, element at (i, j) indicates the number of instances of class i but classified as class j .
- To have good accuracy for a classifier, ideally most diagonal entries should have large values with the rest of entries being close to zero.
- Confusion matrix may have additional rows or columns to provide total or recognition rates per class.

CONFUSION MATRIX FOR MULTICLASS CLASSIFIER

Example 22.3: Confusion matrix with multiple class

Following table shows the confusion matrix of a classification problem with six classes labeled as C_1, C_2, C_3, C_4, C_5 and C_6 .

| Class | C_1 | C_2 | C_3 | C_4 | C_5 | C_6 |
|-------|-------|-------|-------|-------|-------|-------|
| C_1 | 52 | 10 | 7 | 0 | 0 | 1 |
| C_2 | 15 | 50 | 6 | 2 | 1 | 2 |
| C_3 | 5 | 6 | 6 | 0 | 0 | 0 |
| C_4 | 0 | 2 | 0 | 10 | 0 | 1 |
| C_5 | 0 | 1 | 0 | 0 | 7 | 1 |
| C_6 | 1 | 3 | 0 | 1 | 0 | 24 |

Predictive accuracy?

CONFUSION MATRIX FOR MULTICLASS CLASSIFIER

- In case of multiclass classification, sometimes one class is important enough to be regarded as positive with all other classes combined together as negative.
- Thus a large confusion matrix of $m \times m$ can be concised into 2×2 matrix.

Example 22.4: $m \times m$ CM to 2×2 CM

- For example, the CM shown in the above Example is transformed into a CM of size 2×2 considering the class C_1 as the positive class and classes C_2, C_3, C_4, C_5 and C_6 combined together as negative.

| Class | + | - |
|-------|----|-----|
| + | 52 | 18 |
| - | 21 | 123 |

How we can calculate the predictive accuracy of the classifier model in this case?

Are the predictive accuracy same in both Example 22.3 and Example 22.4?

PERFORMANCE EVALUATION METRICS

- We now define a number of metrics for the measurement of a classifier.
 - In our discussion, we shall make the assumptions that there are only two classes: + (positive) and – (negative)
 - Nevertheless, the metrics can easily be extended to multi-class classifiers (with some modifications)
- **True Positive Rate (TPR):** It is defined as the fraction of the positive examples predicted correctly by the classifier.

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{f_{++}}{f_{++}+f_{+-}}$$

- This metric is also known as **Recall**, **Sensitivity** or **Hit rate**.
- **False Positive Rate (FPR):** It is defined as the fraction of negative examples classified as positive class by the classifier.

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = \frac{f_{-+}}{f_{-+}+f_{--}}$$

- This metric is also known as **False Alarm Rate**.

PERFORMANCE EVALUATION METRICS

- **False Negative Rate (FNR):** It is defined as the fraction of positive examples classified as a negative class by the classifier.

$$FNR = \frac{FN}{P} = \frac{FN}{TP + FN} = \frac{f_{+-}}{f_{++} + f_{+-}}$$

- **True Negative Rate (TNR):** It is defined as the fraction of negative examples classified correctly by the classifier

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = \frac{f_{--}}{f_{--} + f_{-+}}$$

- This metric is also known as ***Specificity***.

PERFORMANCE EVALUATION METRICS

- **Positive Predictive Value (PPV):** It is defined as the fraction of the positive examples classified as positive that are really positive

$$PPV = \frac{TP}{TP + FP} = \frac{f_{++}}{f_{++} + f_{-+}}$$

- It is also known as *Precision*.
- **F₁ Score (F₁):** Recall (r) and Precision (p) are two widely used metrics employed in analysis, where detection of one of the classes is considered more significant than the others.
 - It is defined in terms of (r or TPR) and (p or PPV) as follows.

$$\begin{aligned} F_1 &= \frac{2r \cdot p}{r + p} = \frac{2TP}{2TP + FP + FN} \\ &= \frac{2f_{++}}{2f_{++} + f_{\mp} + f_{+-}} = \frac{2}{\frac{1}{r} + \frac{1}{p}} \end{aligned}$$

Note

- F₁ represents the harmonic mean between recall and precision
- High value of F₁ score ensures that both Precision and Recall are reasonably high.

PERFORMANCE EVALUATION METRICS

- More generally, F_β score can be used to determine the trade-off between **Recall** and **Precision** as

$$F_\beta = \frac{(\beta + 1)rp}{r + \beta p} = \frac{(\beta + 1)TP}{(\beta + 1)TP + \beta FN + FP}$$

- Both, **Precision** and **Recall** are special cases of F_β when $\beta = 0$ and $\beta = 1$, respectively.

$$F_\beta = \frac{TP}{TP + FP} = Precision$$

$$F_\alpha = \frac{TP}{TP + FN} = Recall$$

PERFORMANCE EVALUATION METRICS

- A more general metric that captures Recall, Precision is defined in the following.

$$F_{\omega} = \frac{\omega_1 TP + \omega_4 TN}{\omega_1 TP + \omega_2 FP + \omega_3 FN + \omega_4 TN}$$

| Metric | | | | |
|-----------|---|---|---|---|
| Recall | 1 | 1 | 0 | 1 |
| Precision | 1 | 0 | 1 | 0 |
| | | | 1 | 0 |

Note

- In fact, given TPR , FPR , p and r , we can derive all others measures.
- That is, these are the universal metrics.

PREDICTIVE ACCURACY (E)

- It is defined as the fraction of the number of examples that are correctly classified by the classifier to the total number of instances.

$$\varepsilon = \frac{TP + TN}{P + N}$$

$$= \frac{TP + TN}{TP + FP + FN + TN}$$

$$= \frac{f_{++} + f_{--}}{f_{++} + f_{+-} + f_{-+} + f_{--}}$$

- This accuracy is equivalent to F_w with $w_1 = w_2 = w_3 = w_4 = 1$.

ERROR RATE ($\bar{\epsilon}$)

- The error rate $\bar{\epsilon}$ is defined as the fraction of the examples that are incorrectly classified.

$$\begin{aligned}\bar{\epsilon} &= \frac{FP + FN}{P + N} \\ &= \frac{FP + FN}{TP + TN + FP + FN} \\ &= \frac{f_{+-} + f_{-_+}}{f_{++} + f_{+-} + f_{-_+} + f_{--}}\end{aligned}$$

Note

$$\bar{\epsilon} = 1 - \epsilon.$$

ACCURACY, SENSITIVITY AND SPECIFICITY

- Predictive accuracy (ε) can be expressed in terms of sensitivity and specificity.
- We can write

$$\varepsilon = \frac{TP + TN}{TP + FP + FN + TN}$$

$$= \frac{TP + TN}{P + N}$$

$$\varepsilon = \frac{TP}{P} \times \frac{P}{P + N} + \frac{TN}{N} \times \frac{N}{P + N}$$

Thus,

$$\varepsilon = \text{Sensitivity} \times \frac{P}{P+N} + \text{Specificity} \times \frac{N}{P+N}$$

ANALYSIS WITH PERFORMANCE MEASUREMENT METRICS

- Based on the various performance metrics, we can characterize a classifier.
- We do it in terms of TPR, FPR, Precision and Recall and Accuracy
- **Case 1: Perfect Classifier**

When every instance is **correctly** classified, it is called the **perfect classifier**. In this case, $TP = P$, $TN = N$ and CM is

$$TPR = \frac{P}{P} = 1$$

$$FPR = \frac{0}{N} = 0$$

$$Precision = \frac{P}{P} = 1$$

$$F_1 Score = \frac{2 \times 1}{1+1} = 1$$

$$Accuracy = \frac{P+N}{P+N} = 1$$

| | | Predicted Class | |
|--------------|---|-----------------|---|
| | | + | - |
| Actual class | + | P | 0 |
| | - | 0 | N |

ANALYSIS WITH PERFORMANCE MEASUREMENT METRICS

- **Case 2: Worst Classifier**

When every instance is **wrongly** classified, it is called the **worst classifier**. In this case, $TP = 0$, $TN = 0$ and the CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{N}{N} = 1$$

$$Precision = \frac{0}{N} = 0$$

F_1 Score = Not applicable

as $Recall + Precision = 0$

$$\text{Accuracy} = \frac{0}{P+N} = 0$$

| | | Predicted Class | |
|--------------|---|-----------------|---|
| | | + | - |
| Actual class | + | 0 | P |
| | - | N | 0 |

ANALYSIS WITH PERFORMANCE MEASUREMENT METRICS

- Case 3: Ultra-Liberal Classifier

The classifier always predicts the + class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$$TPR = \frac{P}{P} = 1$$

$$FPR = \frac{N}{N} = 1$$

$$Precision = \frac{P}{P+N}$$

$$F_1 Score = \frac{2P}{2P+N}$$

$$Accuracy = \frac{P}{P+N}$$

| | | Predicted Class | |
|--------------|---|-----------------|---|
| | | + | - |
| Actual class | + | P | 0 |
| | - | N | 0 |

ANALYSIS WITH PERFORMANCE MEASUREMENT METRICS

- **Case 4: Ultra-Conservative Classifier**

This classifier always predicts the - class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{0}{N} = 0$$

Precision = Not applicable
(as $TP + FP = 0$)

F_1 Score = Not applicable

$$\text{Accuracy} = \frac{N}{P+N}$$

| | | Predicted Class | |
|--------------|---|-----------------|---|
| | | + | - |
| Actual class | + | 0 | p |
| | - | 0 | N |

PREDICTIVE ACCURACY VERSUS TPR AND FPR

- One strength of characterizing a classifier by its *TPR* and *FPR* is that they do not depend on the relative size of P and N .
 - The same is also applicable for *FNR* and *TNR* and others measures from CM.
- In contrast, the *Predictive Accuracy*, *Precision*, *Error Rate*, F_1 *Score*, etc. are affected by the relative size of P and N .
- *FPR*, *TPR*, *FNR* and *TNR* are calculated from the different rows of the CM.
 - On the other hand Predictive Accuracy, etc. are derived from the values in both rows.
- This suggests that *FPR*, *TPR*, *FNR* and *TNR* are more effective than *Predictive Accuracy*, etc.

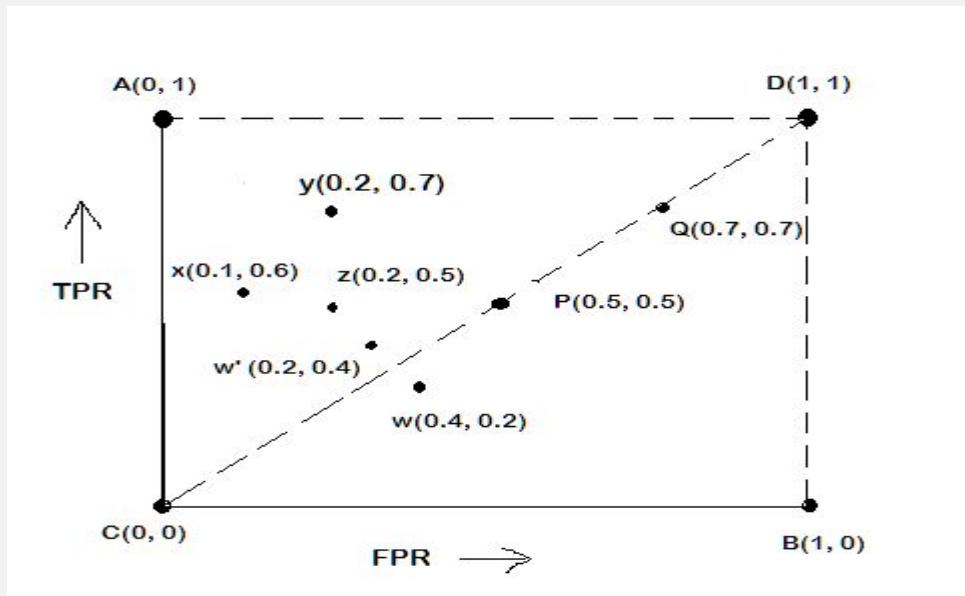
ROC Curves

ROC CURVES

- ROC is an abbreviation of **Receiver Operating Characteristic** come from the signal detection theory, developed during World War 2 for analysis of radar images.
- In the context of classifier, ROC plot is a useful tool to study the behaviour of a classifier or **comparing two or more classifiers**.
- A ROC plot is **a two-dimensional graph**, where, X-axis represents FP rate (FPR) and Y-axis represents TP rate (TPR).
- Since, the values of FPR and TPR varies from 0 to 1 both inclusive, the two axes thus from 0 to 1 only.
- Each point (x, y) on the plot indicating that the FPR has value x and the TPR value y .

ROC PLOT

- A typical look of ROC plot with few points in it is shown in the following figure.

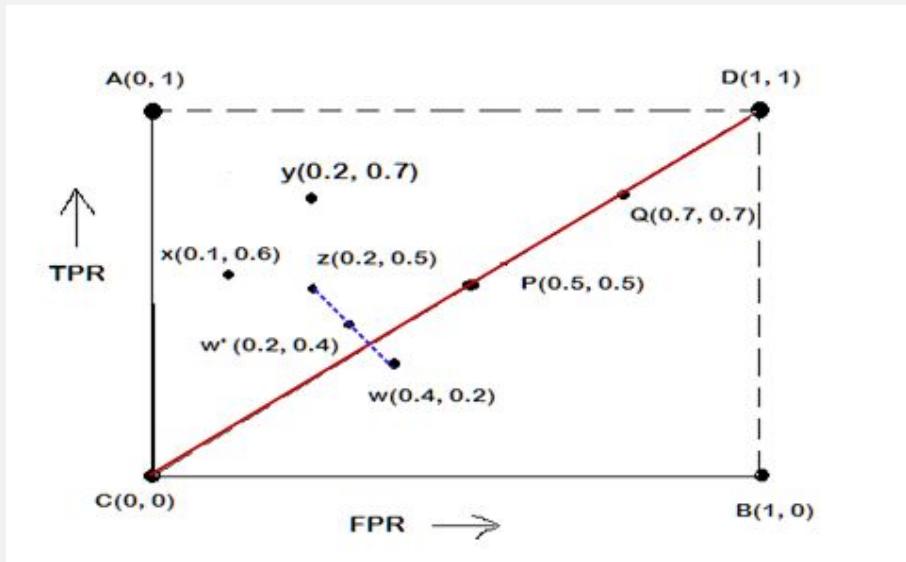


- Note the four cornered points are the four extreme cases of classifiers

Identify the four extreme classifiers.

INTERPRETATION OF DIFFERENT POINTS IN ROC PLOT

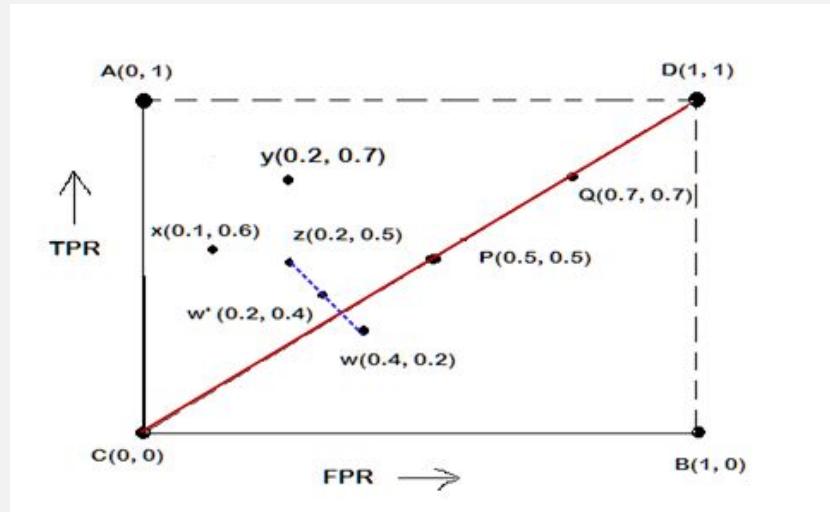
- Let us interpret the different points in the ROC plot.



- The four points (A, B, C, and D)
 - A**: $\text{TPR} = 1, \text{FPR} = 0$, the ideal model, i.e., the **perfect classifier**, no false results
 - B**: $\text{TPR} = 0, \text{FPR} = 1$, the **worst classifier**, not able to predict a single instance
 - C**: $\text{TPR} = 0, \text{FPR} = 0$, the model predicts every instance to be a **Negative** class, i.e., it is an **ultra-conservative classifier**
 - D**: $\text{TPR} = 1, \text{FPR} = 1$, the model predicts every instance to be a **Positive** class, i.e., it is an **ultra-liberal classifier**

INTERPRETATION OF DIFFERENT POINTS IN ROC PLOT

- Let us interpret the different points in the ROC plot.

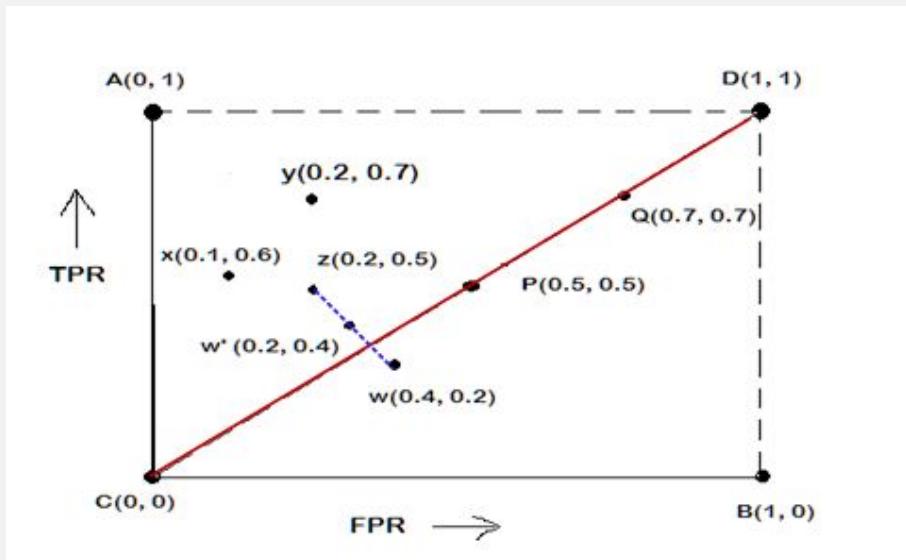


The points on diagonals

- The diagonal line joining point C(0,0) and D(1,1) corresponds to **random guessing**
 - Random guessing means that a record is classified as positive (or negative) with a certain probability
 - Suppose, a test set containing N_+ positive and N_- negative instances. Suppose, the classifier guesses any instances with probability p
 - Thus, the random classifier is expected to correctly classify $p.N_+$ of the positive instances and $p.N_-$ of the negative instances
 - Hence, $TPR = FPR = p$
 - Since $TPR = FPR$, the random classifier results reside on the main diagonals

INTERPRETATION OF DIFFERENT POINTS IN ROC PLOT

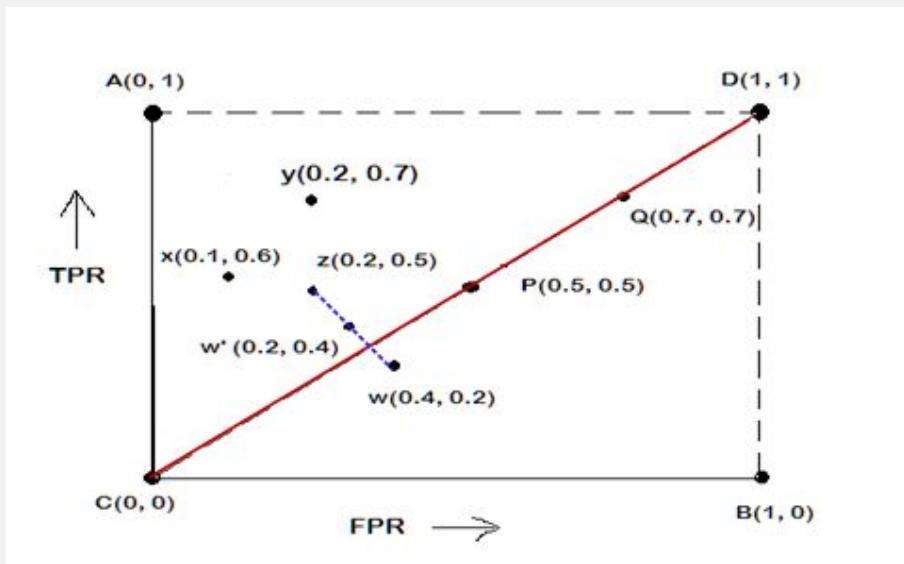
- Let us interpret the different points in the ROC plot.



- The points on the upper diagonal region
 - All points, which reside on upper-diagonal region are corresponding to classifiers “good” as their TPR is as good as FPR (i.e., FPRs are lower than TPRs)
 - Here, X is better than Z as X has higher TPR and lower FPR than Z.
 - If we compare X and Y, neither classifier is superior to the other

INTERPRETATION OF DIFFERENT POINTS IN ROC PLOT

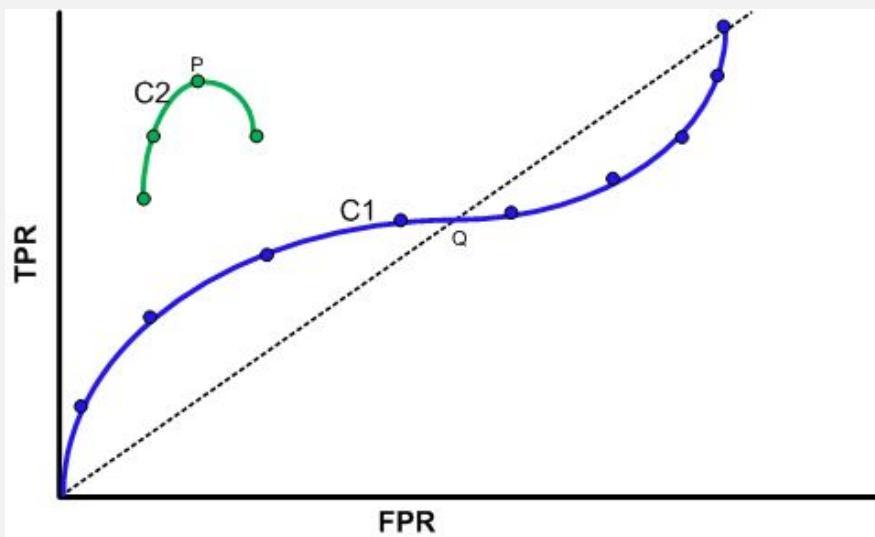
- Let us interpret the different points in the ROC plot.



- The points on the lower diagonal region
 - The Lower-diagonal triangle corresponds to the classifiers that are worst than random classifiers
 - Note: A classifier that is worst than random guessing, simply by reversing its prediction, we can get good results.
 - W'(0.2, 0.4) is the better version than W(0.4, 0.2), W' is a mirror reflection of W

TUNING A CLASSIFIER THROUGH ROC PLOT

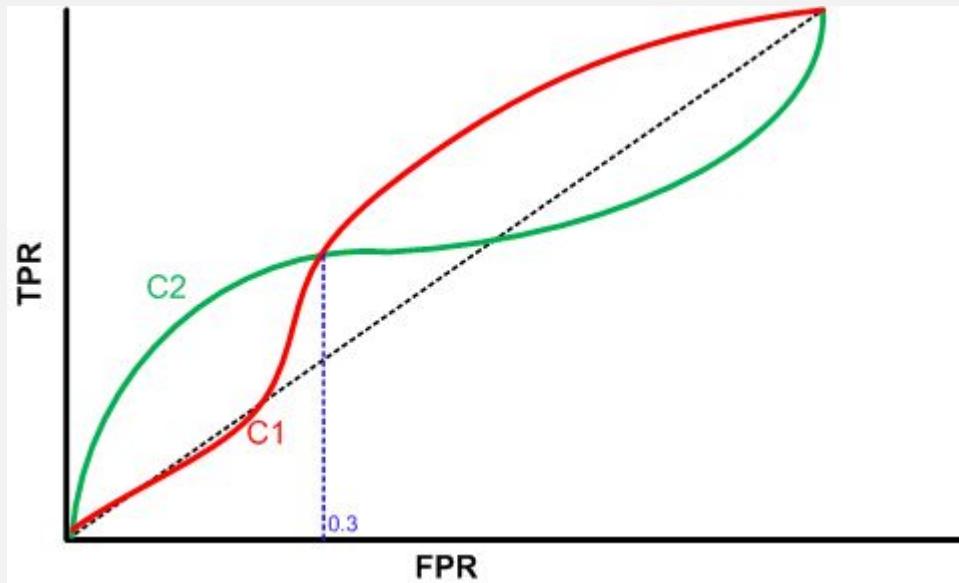
- Using ROC plot, we can compare two or more classifiers by their TPR and FPR values and this plot also depicts the trade-off between TPR and FPR of a classifier.



- Examining ROC curves can give insights into the best way of tuning parameters of classifier.
- For example, in the curve C2, the result is degraded after the point P. Similarly for the observation C1, beyond Q the settings are not acceptable.

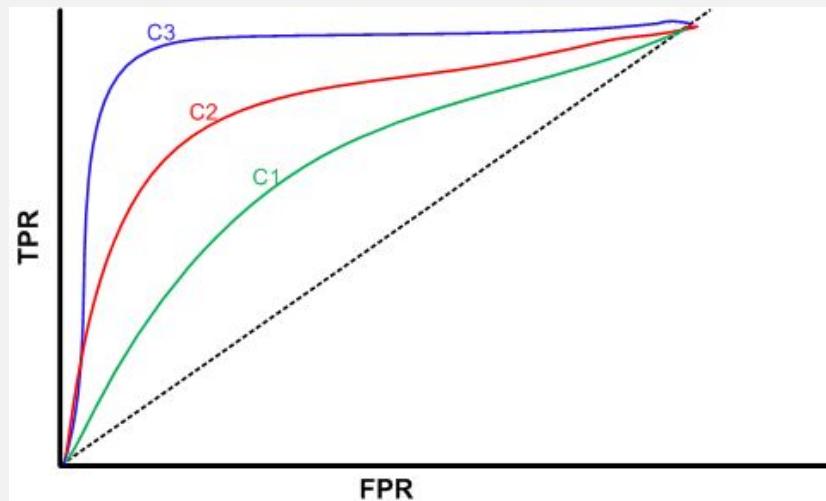
COMPARING CLASSIFIERS THROUGH ROC PLOT

- Two curves C1 and C2 are corresponding to the experiments to choose two classifiers with their parameters.
- Here, C2 is better than C1 when FPR is less than 0.3.
- However, C1 is better, when FPR is greater than 0.3.
- Clearly, neither of these two classifiers dominates the other.



COMPARING CLASSIFIERS THROUGH ROC PLOT

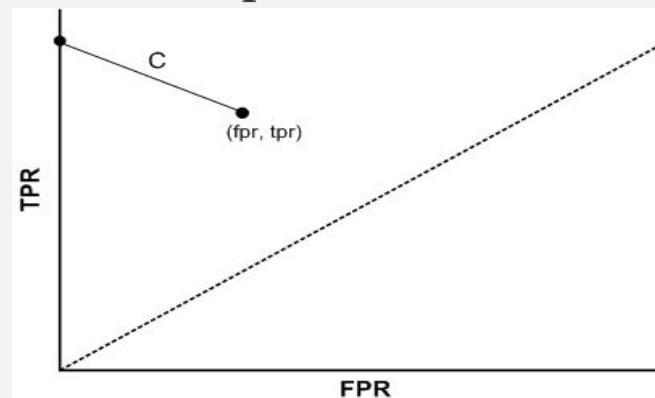
- We can use the concept of “**area under curve**” (AUC) as a better method to compare two or more classifiers.
- If a model is perfect, then its AUC = 1.
- If a model simply performs random guessing, then its AUC = 0.5
- A model that is strictly better than other, would have a larger value of AUC than the other.



- Here, C3 is best, and C2 is better than C1 as $AUC(C3) > AUC(C2) > AUC(C1)$.

A QUANTITATIVE MEASURE OF A CLASSIFIER

- The concept of ROC plot can be extended to compare quantitatively using Euclidean distance measure.
- See the following figure for an explanation.



- Here, $C(fpr, tpr)$ is a classifier and δ denotes the Euclidean distance between the best classifier $(0, 1)$ and C . That is,

$$\bullet \quad \delta = \sqrt{fpr^2 + (1 - tpr)^2}$$

- The smallest possible value of δ is 0
- The largest possible values of δ is $\sqrt{2}$ (when $(fpr = 1$ and $tpr = 0)$.

REFERENCE

- The detail material related to this lecture can be found in

Data Mining: Concepts and Techniques, (3rd Edn.), Jiawei Han, Micheline Kamber, Morgan Kaufmann, 2015.

Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Addison-Wesley, 2014

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 24

Clustering Techniques: Similarity Measures

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

Topics to be covered...

- Introduction to clustering
- Similarity and dissimilarity measures
- Clustering techniques
 - Partitioning algorithms
 - Hierarchical algorithms
 - Density-based algorithm

Introduction to Clustering

- Classification consists of assigning a class label to a set of unclassified cases.
- **Supervised Classification**
 - The set of possible classes is known in advance.
- **Unsupervised Classification**
 - Set of possible classes is not known. After classification we can try to assign a name to that class.
 - Unsupervised classification is called **clustering**.

Introduction to Clustering

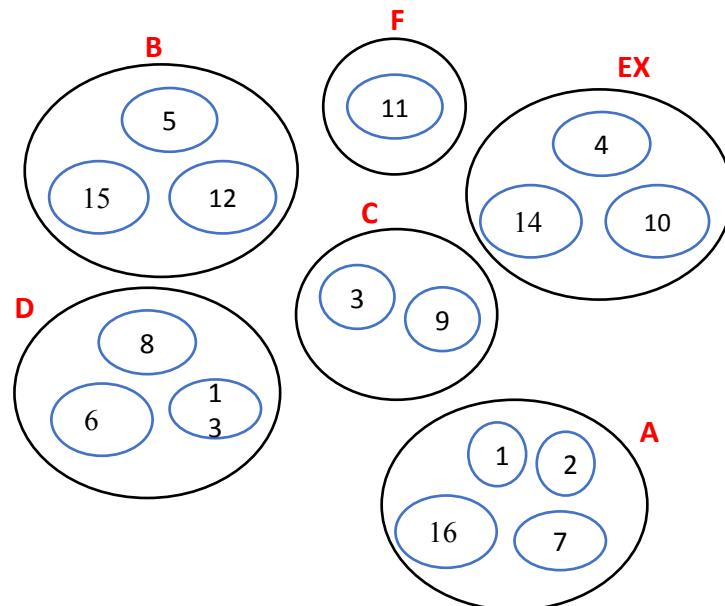
- Clustering is somewhat related to classification in the sense that in both cases data are grouped.
 -
- However, there is a major difference between these two techniques.
- In order to understand the difference between the two, consider a sample dataset containing marks obtained by a set of students and corresponding grades as shown in Table 24.1.

Introduction to Clustering

Table 24.1: Tabulation of Marks

| Roll No | Mark | Grade |
|---------|------|-------|
| 1 | 80 | A |
| 2 | 70 | A |
| 3 | 55 | C |
| 4 | 91 | EX |
| 5 | 65 | B |
| 6 | 35 | D |
| 7 | 76 | A |
| 8 | 40 | D |
| 9 | 50 | C |
| 10 | 85 | EX |
| 11 | 25 | F |
| 12 | 60 | B |
| 13 | 45 | D |
| 14 | 95 | EX |
| 15 | 63 | B |
| 16 | 88 | A |

Figure 24.1: Group representation of dataset in Table 24.1

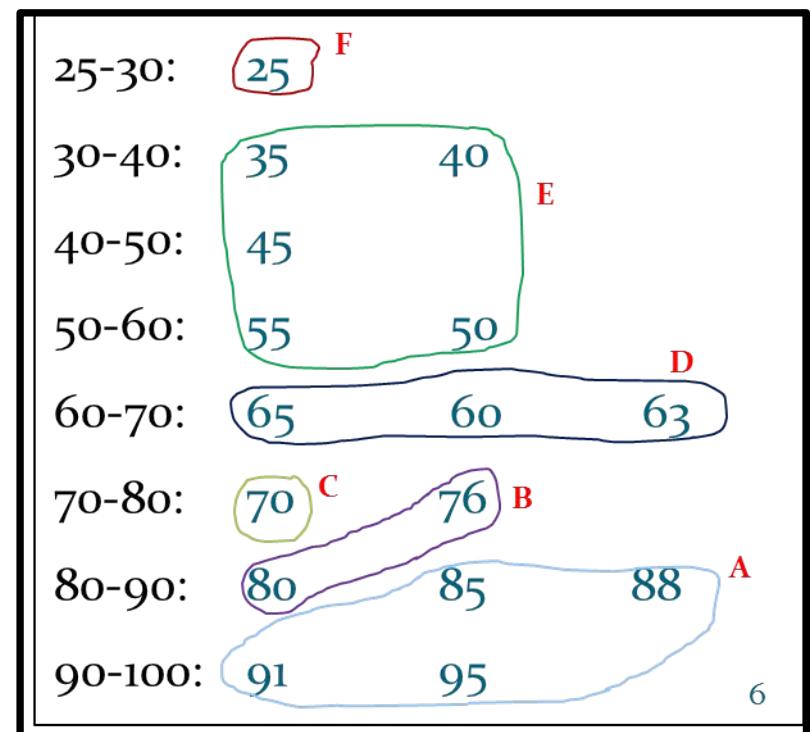


Introduction to Clustering

- It is evident that there is a simple mapping between Table 24.1 and Fig 24.1.
- The fact is that groups in Fig 24.1 are already predefined in Table 24.1. This is similar to classification, where we have given a dataset where **groups of data are predefined**.
- Consider another situation, where ‘Grade’ is not known, but we have to make a grouping.
- Put all the marks into a group if any other mark in that group does not exceed by 5 or more.
- This is similar to “**Relative grading**” concept and grade may range from A to Z.

Introduction to Clustering

- Figure 24.2 shows another grouping by means of another simple mapping, but the difference is **this mapping does not based on predefined classes**.
- In other words, this grouping is accomplished by finding **similarities** between data according to characteristics found in the actual data.
- Such a group making is called **clustering**.



Introduction to Clustering

Example 24.1 : The task of clustering

In order to elaborate the clustering task, consider the following dataset.

Table 24.2: Life Insurance database

| Marital Status | Age | Income | Education | Number of children |
|----------------|-----|--------|----------------|--------------------|
| Single | 35 | 25000 | Under Graduate | 3 |
| Married | 25 | 15000 | Graduate | 1 |
| Single | 40 | 20000 | Under Graduate | 0 |
| Divorced | 20 | 30000 | Post-Graduate | 0 |
| Divorced | 25 | 20000 | Under Graduate | 3 |
| Married | 60 | 70000 | Graduate | 0 |
| Married | 30 | 90000 | Post-Graduate | 0 |
| Married | 45 | 60000 | Graduate | 5 |
| Divorced | 50 | 80000 | Under Graduate | 2 |

With certain similarity or likeliness defined, we can classify the records to one or group of more attributes (and thus mapping being non-trivial).

Introduction to Clustering

- Clustering has been used in many application domains:
 - Image analysis
 - Document retrieval
 - Machine learning, etc.
- When clustering is applied to real-world database, many problems may arise.
 1. The (best) number of cluster is not known.
 - There is no correct answer to a clustering problem.
 - In fact, many answers may be found.
 - The exact number of cluster required is not easy to determine.

Introduction to Clustering

2. There may not be any a priori knowledge concerning the clusters.

- This is an issue that what data should be used for clustering.
- Unlike classification, in clustering, we have no supervisory learning to aid the process.
- Clustering can be viewed as similar to [unsupervised learning](#).

3. Interpreting the semantic meaning of each cluster may be difficult.

- With classification, the labeling of classes is known ahead of time. In contrast, with clustering, this may not be the case.
- Thus, when the clustering process is finished yielding a set of clusters, the exact meaning of each cluster may not be obvious.

Definition of Clustering Problem

Definition 24.1: Clustering

Given a database $D = \{t_1, t_2, \dots, t_n\}$ of n tuples, the clustering problem is to define a mapping $f : D \rightarrow C$, where each $t_i \in D$ is assigned to one cluster $c_i \in C$. Here, $C = \{c_1, c_2, \dots, c_k\}$ denotes a set of clusters.

- Solution to a clustering problem is devising a mapping formulation.
- The formulation behind such a mapping is to establish that a tuple within one cluster is **more like** tuples within that cluster and not similar to tuples outside it.

Definition of Clustering Problem

- Hence, mapping function f in Definition 24.1 may be explicitly stated as

$$f : D \rightarrow \{c_1, c_2, \dots, c_k\}$$

- where
- i) each $t_i \in D$ is assigned to one cluster $c_i \in C$.
 - ii) for each cluster $c_i \in C$, and for all $t_{ip}, t_{iq} \in c_i$ and there exist $t_j \notin c_i$ such that
similarity (t_{ip}, t_{iq}) > similarity (t_{ip}, t_j) AND similarity (t_{iq}, t_j)

- In the field of cluster analysis, this **similarity** plays an important part.
- Now, we shall learn how similarity (this is also alternatively judged as “dissimilarity”) between any two data can be measured.

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

Webster's Dictionary



Similarity is hard
to define, but...

*"We know it when
we see it"*

The real meaning
of similarity is a
philosophical
question. We will
take a more
pragmatic
approach.

Similarity and Dissimilarity Measures

- In clustering techniques, similarity (or dissimilarity) is an important measurement.
- Informally, **similarity** between two objects (e.g., two images, two documents, two records, etc.) is a numerical measure of the degree to which two objects are **alike**.
- The **dissimilarity** on the other hand, is another alternative (or opposite) measure of the degree to which two objects are **different**.
- Both similarity and dissimilarity also termed as **proximity**.
- Usually, similarity and dissimilarity are **non-negative numbers** and may range from **zero** (highly dissimilar (no similar)) to some finite/infinite value (highly similar (no dissimilar)).

Note:

- Frequently, the term **distance** is used as a synonym for dissimilarity
- In fact, it is used to refer as a special case of dissimilarity.

Proximity Measures: Single-Attribute

- Consider an object, which is defined by a single attribute A (e.g., length) and the attribute A has n -distinct values a_1, a_2, \dots, a_n .
- A data structure called “Dissimilarity matrix” is used to store a collection of proximities that are available for all pair of n attribute values.
 - In other words, the Dissimilarity matrix for an attribute A with n values is represented by an $n \times n$ matrix as shown below.

$$\begin{bmatrix} 0 & & & \\ p_{(2,1)} & 0 & & \\ p_{(3,1)} & p_{(3,2)} & 0 & \\ \vdots & \vdots & \vdots & \\ p_{(n,1)} & p_{(n,2)} & \dots & 0 \end{bmatrix}_{n \times n}$$

- Here, $p_{(i,j)}$ denotes the proximity measure between two objects with attribute values a_i and a_j .
- Note: The proximity measure is symmetric, that is, $p_{(i,j)} = p_{(j,i)}$

Proximity Calculation

- Proximity calculation to compute $p_{(i,j)}$ is different for different types of attributes according to NOIR topology.

Proximity calculation for Nominal attributes:

- For example, binary attribute, **Gender** = {Male, female} where **Male** is equivalent to **binary 1** and **female** is equivalent to **binary 0**.
- Similarity value is 1 if the two objects contains the same attribute value, while similarity value is 0 implies objects are not at all similar.

| Object | Gender |
|--------|--------|
| Ram | Male |
| Sita | Female |
| Laxman | Male |

- Here, Similarity value let it be denoted by p , among different objects are as follows.

$$p(Ram, sita) = 0$$

$$p(Ram, Laxman) = 1$$

Note : In this case, if q denotes the dissimilarity between two objects i and j with single binary attributes, then $q_{(i,j)} = 1 - p_{(i,j)}$

Proximity Calculation

- Now, let us focus on how to calculate proximity measures between objects which are defined by two or more binary attributes.
- Suppose, the number of attributes be b . We can define the contingency table summarizing the different matches and mismatches between any two objects x and y , which are as follows.

Table 24.3: Contingency table with binary attributes

| Object x | | |
|------------|----------|----------|
| | 1 | 0 |
| 1 | f_{11} | f_{10} |
| 0 | f_{01} | f_{00} |

Here, f_{11} = the number of attributes where $x=1$ and $y=1$.

f_{10} = the number of attributes where $x=1$ and $y=0$.

f_{01} = the number of attributes where $x=0$ and $y=1$.

f_{00} = the number of attributes where $x=0$ and $y=0$.

Note : $f_{00} + f_{01} + f_{10} + f_{11} = b$, the total number of binary attributes.

Now, two cases may arise: symmetric and asymmetric binary attributes.

Similarity Measure with Symmetric Binary

- To measure the similarity between two objects defined by symmetric binary attributes using a measure called **symmetric binary coefficient** and denoted as \mathcal{S} and defined below

$$\mathcal{S} = \frac{\text{Number of matching attribute values}}{\text{Total number of attributes}}$$

or

$$\mathcal{S} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

The **dissimilarity measure**, likewise can be denoted as \mathcal{D} and defined as

$$\mathcal{D} = \frac{\text{Number of mismatched attribute values}}{\text{Total number of attributes}}$$

or

$$\mathcal{D} = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Note that, $\mathcal{D} = 1 - \mathcal{S}$

Similarity Measure with Symmetric Binary

Example 24.2: Proximity measures with symmetric binary attributes

Consider the following two dataset, where objects are defined with symmetric binary attributes.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I},
Hobby = {T, C}, Job = {Y, N}

| Object | Gender | Food | Caste | Education | Hobby | Job |
|--------|--------|------|-------|-----------|-------|-----|
| Hari | M | V | M | L | C | N |
| Ram | M | N | M | I | T | N |
| Tomi | F | N | H | L | C | Y |

$$S(Hari, Ram) = \frac{1+2}{1+2+1+2} = 0.5$$

Proximity Measure with Asymmetric Binary

- Such a similarity measure between two objects defined by asymmetric binary attributes is done by **Jaccard Coefficient** and which is often symbolized by \mathcal{J} is given by the following equation

$$\mathcal{J} = \frac{\text{Number of matching presence}}{\text{Number of attributes not involved in 00 matching}}$$

or

$$\mathcal{J} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Proximity Measure with Asymmetric Binary

Example 24.3: Jaccard Coefficient

Consider the following two dataset.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I},
Hobby = {T, C}, Job = {Y, N}

Calculate the Jaccard coefficient between Ram and Hari assuming that all binary attributes are asymmetric and for each pair values for an attribute, first one is more frequent than the second.

| Object | Gender | Food | Caste | Education | Hobby | Job |
|--------|--------|------|-------|-----------|-------|-----|
| Hari | M | V | M | L | C | N |
| Ram | M | N | M | I | T | N |
| Tomi | F | N | H | L | C | Y |

$$J(Hari, Ram) = \frac{1}{2+1+1} = 0.25$$

Note: $J(Ram, Tomi) = 0$ and $J(Hari, Ram) = J(Ram, Hari)$, etc.

Example 24.4:

Consider the following two dataset.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I},
Hobby = {T, C}, Job = {Y, N}

| Object | Gender | Food | Caste | Education | Hobby | Job |
|--------|--------|------|-------|-----------|-------|-----|
| Hari | M | V | M | L | C | N |
| Ram | M | N | M | I | T | N |
| Tomi | F | N | H | L | C | Y |



How you can calculate similarity if Gender, Hobby and Job are symmetric binary attributes and Food, Caste, Education are asymmetric binary attributes?

Obtain the similarity matrix with Jaccard coefficient of objects for the above, e.g.

$$\mathcal{J} = \begin{matrix} & H & R & T \\ H & \left[\begin{matrix} 0 & 0 & 0 \\ \mathcal{J}(R, H) & 0 & 0 \\ \mathcal{J}(T, H) & \mathcal{J}(T, R) & 0 \end{matrix} \right] \\ R & & & \\ T & & & \end{matrix}$$

Proximity Measure with Categorical Attribute

- Binary attribute is a special kind of nominal attribute where the attribute has values with two states only.
- On the other hand, categorical attribute is another kind of nominal attribute where it has values with three or more states (e.g. color = {Red, Green, Blue}).
- If $s(x, y)$ denotes the similarity between two objects x and y , then

$$s(x, y) = \frac{\text{Number of matches}}{\text{Total number of attributes}}$$

and the dissimilarity $d(x, y)$ is

$$d(x, y) = \frac{\text{Number of mismatches}}{\text{Total number of attributes}}$$

- If m = number of matches and a = number of categorical attributes with which objects are defined as

$$s(x, y) = \frac{m}{a} \quad \text{and} \quad d(x, y) = \frac{a-m}{a}$$

Proximity Measure with Categorical Attribute

Example 24.4:

| Object | Color | Position | Distance |
|--------|-------|----------|----------|
| 1 | R | L | L |
| 2 | B | C | M |
| 3 | G | R | M |
| 4 | R | L | H |

The similarity matrix considering only color attribute is shown below

$$s = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

Dissimilarity matrix, $d = ?$

Obtain the dissimilarity matrix considering both the categorical attributes (i.e. color and position).

Proximity Measure with Ordinal Attribute

- Ordinal attribute is a special kind of categorical attribute, where the values of attribute follows a sequence (ordering) e.g. Grade = {Ex, A, B, C} where Ex > A > B > C.
- Suppose, A is an attribute of type ordinal and the set of values of $A = \{a_1, a_2, \dots, a_n\}$. Let n values of A are ordered in ascending order as $a_1 < a_2 < \dots < a_n$. Let i -th attribute value a_i be ranked as i , $i=1, 2, \dots, n$.
- The normalized value of a_i can be expressed as

$$\hat{a}_i = \frac{i - 1}{n - 1}$$

- Thus, normalized values lie in the range [0..1].
- As a_i is a numerical value, the similarity measure, then can be calculated using any similarity measurement method for numerical attribute.
- For example, the similarity measure between two objects x and y with attribute values a_i and a_j , then can be expressed as

$$s(x, y) = \sqrt{(\hat{a}_i - \hat{a}_j)^2}$$

where \hat{a}_i and \hat{a}_j are the normalized values of a_i and a_j , respectively.

Proximity Measure with Ordinal Attribute

Example 24.5:

Consider the following set of records, where each record is defined by two ordinal attributes $\text{size} = \{\text{S, M, L}\}$ and $\text{Quality} = \{\text{Ex, A, B, C}\}$ such that $\text{S} < \text{M} < \text{L}$ and $\text{Ex} > \text{A} > \text{B} > \text{C}$.

| Object | Size | Quality |
|--------|---------|----------|
| A | S (0.0) | A (0.66) |
| B | L (1.0) | Ex (1.0) |
| C | L (1.0) | C (0.0) |
| D | M (0.5) | B (0.33) |

- Normalized values are shown in brackets.
- Their similarity measures are shown in the similarity matrix below.

$$\begin{array}{ccccc} & A & B & C & D \\ A & \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \right] \\ B & \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \right] \\ C & \left[\begin{array}{cccc} ? & 0 & 0 & 0 \end{array} \right] \\ D & \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \end{array} \right] \end{array}$$

Find the dissimilarity matrix, when each object is defined by only one ordinal attribute say size (or quality).

Proximity Measure with Interval Scale

- The measure called **distance** is usually referred to estimate the similarity between two objects defined with interval-scaled attributes.
- We first present a generic formula to express distance d between two objects x and y in n -dimensional space. Suppose, x_i and y_i denote the values of i^{th} attribute of the objects x and y respectively.

$$d(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^r \right)^{\frac{1}{r}}$$

- Here, r is any integer value.
- This distance metric most popularly known as **Minkowski metric**.
- This distance measure follows some well-known properties. These are mentioned in the next slide.

Proximity Measure with Interval Scale

Properties of Minkowski metrics:

1. Non-negativity:

a. $d(x, y) \geq 0$ for all x and y

b. $d(x, y) = 0$ only if $x = y$. This is also called identity condition.

2. Symmetry:

$d(x, y) = d(y, x)$ for all x and y

This condition ensures that the order in which objects are considered is not important.

3. Transitivity:

$d(x, z) \leq d(x, y) + d(y, z)$ for all x, y and z .

- This condition has the interpretation that the least distance $d(x, z)$ between objects x and z is always less than or equal to the sum of the distance between the objects x and y , and between y and z .
- This property is also termed as Triangle Inequality.

Proximity Measure with Interval Scale

Depending on the value of r , the distance measure is renamed accordingly.

1. Manhattan distance (L_1 Norm: $r = 1$)

The Manhattan distance is expressed as

$$d = \sum_{i=1}^n |x_i - y_i|$$

where $|...|$ denotes the absolute value. This metric is also alternatively termed as **Taxicals metric, city-block metric**.

Example: $x = [7, 3, 5]$ and $y = [3, 2, 6]$.

The Manhattan distance is $|7 - 3| + |3 - 2| + |5 - 6| = 6$.

- As a special instance of Manhattan distance, when attribute values $\in [0, 1]$ is called **Hamming distance**.
- Alternatively, Hamming distance is the number of bits that are different between two objects that have only binary values (i.e. between two binary vectors).

Proximity Measure with Interval Scale

2. Euclidean Distance (L_2 Norm: $r = 2$)

This metric is same as Euclidean distance between any two points x and y in \mathcal{R}^n .

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Example: $x = [7, 3, 5]$ and $y = [3, 2, 6]$.

The Euclidean distance between x and y is

$$d(x, y) = \sqrt{(7 - 3)^2 + (3 - 2)^2 + (5 - 6)^2} = \sqrt{18} \approx 2.426$$

Proximity Measure with Interval Scale

3. Chebychev Distance (L_∞ Norm: $r \in \mathcal{R}$)

This metric is defined as

$$d(x, y) = \max_{\forall i} \{|x_i - y_i|\}$$

- We may clearly note the difference between Chebychev metric and Manhattan distance. That is, instead of summing up the absolute difference (in Manhattan distance), we simply take the maximum of the absolute differences (in Chebychev distance). Hence, $L_\infty < L_1$

Example: $x = [7, 3, 5]$ and $y = [3, 2, 6]$.

The Manhattan distance = $|7 - 3| + |3 - 2| + |5 - 6| = 6$.

The chebychev distance = Max $\{|7 - 3|, |3 - 2|, |5 - 6|\} = 4$.

Proximity Measure with Interval Scale

4. Other metrics:

a. Canberra metric:

$$d(x, y) = \sum_{i=1}^n \frac{|x_i - y_i|}{(|x_i| + |y_i|)^q}$$

- where q is a real number. Usually $q = 1$, because numerator of the ratio is always \leq denominator, the ratio ≤ 1 , that is, the sum is always bounded and small.
- If $q \neq 1$, it is called Fractional Canberra metric.
- If $q > 1$, the opposite relationship holds.

b. Hellinger metric:

$$d(x, y) = \sum_{i=1}^n (\sqrt{x_i} - \sqrt{y_i})^2$$

This metric is then used as either squared or transformed into an acceptable range [-1, +1] using the following transformations.

$$\begin{aligned} i. \quad d(x, y) &= (1 - r(x, y))/2 \\ ii. \quad d(x, y) &= 1 - r(x, y) \end{aligned}$$

Where $r(x, y)$ is correlation coefficient between x and y .

Note: Dissimilarity measurement is not relevant with distance measurement.

Proximity Measure for Ratio-Scale

The proximity between the objects with ratio-scaled variable can be carried with the following steps:

1. Apply the appropriate transformation to the data to bring it into a linear scale. (e.g. logarithmic transformation to data of the form $X = Ae^B$).
2. The transformed values can be treated as interval-scaled values. Any distance measure discussed for interval-scaled variable can be applied to measure the similarity.

Note:

There are two concerns on proximity measures:

- Normalization of the measured values.
- Intra-transformation from similarity to dissimilarity measure and vice-versa.

Proximity Measure for Ratio-Scale

Normalization:

- A major problem when using the similarity (or dissimilarity) measures (such as Euclidean distance) is that the large values frequently swamp the small ones.
- For example, consider the following data.

| Make | Cost 1 | Cost 2 | Cost 3 |
|------|----------|--------|--------|
| X | 2,00,000 | 70 | 10 |
| Y | 2,50,000 | 100 | 5 |

Here, the contribution of Cost 2 and Cost 3 is insignificant compared to Cost 1 so far the Euclidean distance is concerned.

- This problem can be avoided if we consider the normalized values of all numerical attributes.
- Another normalization may be to take the estimated values in a normalized range say [0, 1]. Note that, if a measure varies in the range, then it can be normalized as

$$s' = \frac{1}{1+s} \text{ where } s \in [0.. \infty]$$

Proximity Measure for Ratio-Scale

Intra-transformation:

- Transforming similarities to dissimilarities and vice-versa is also relatively straightforward.
- If the similarity (or dissimilarity) falls in the interval [0..1], the dissimilarity (or similarity) can be obtained as

$$\begin{aligned}d &= 1 - s \\ \text{or} \\ s &= 1 - d\end{aligned}$$

- Another approach is to define similarity as the negative of dissimilarity (or vice-versa).

Proximity Measure with Mixed Attributes

- The previous metrics on similarity measures assume that all the attributes were of the same type. Thus, a **general approach is needed when the attributes are of different types.**
- One straightforward approach is to compute the similarity between each attribute separately and then combine these attribute using a method that results in a similarity between 0 and 1.
- Typically, the overall similarity is defined as the average of all the individual attribute similarities.
- See the algorithm in the next slide for doing this.

Similarity Measure with Mixed Attributes

Example 24.6:

Consider the following set of objects. Obtain the similarity matrix.

| Object | A (Binary) | B (Categorical) | C (Ordinal) | D (Numeric) | E (Numeric) |
|--------|---------------|--------------------|----------------|----------------|----------------|
| 1 | Y | R | X | 475 | 10^8 |
| 2 | N | R | A | 10 | 10^{-2} |
| 3 | N | B | C | 1000 | 10^5 |
| 4 | Y | G | B | 500 | 10^3 |
| 5 | Y | B | A | 80 | 1 |

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \left[\begin{matrix} 0 & 0 & 0 & 0 & 0 \\ ? & 0 & 0 & 0 & 0 \\ ? & ? & 0 & 0 & 0 \\ ? & ? & ? & 0 & 0 \\ ? & ? & ? & ? & 0 \end{matrix} \right] \end{matrix}$$

[For C X>A>B>S]

How cosine similarity can be applied to this?

Non-Metric similarity

- In many applications (such as information retrieval) objects are complex and contains a large number of symbolic entities (such as keywords, phrases, etc.).
- To measure the distance between complex objects, it is often desirable to introduce a non-metric similarity function.
- Here, we discuss few such non-metric similarity measurements.

Cosine similarity

Suppose, x and y denote two vectors representing two complex objects. The cosine similarity denoted as $\cos(x, y)$ and defined as

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

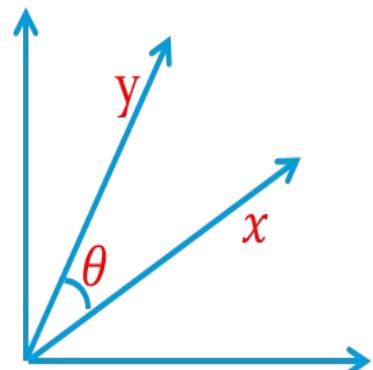
- where $x \cdot y$ denotes the vector dot product, namely $x \cdot y = \sum_{i=1}^n x_i \cdot y_i$ such that $x = [x_1, x_2, \dots, x_n]$ and $y = [y_1, y_2, \dots, y_n]$.
- $\|x\|$ and $\|y\|$ denote the Euclidean norms of vector x and y , respectively (essentially the length of vectors x and y), that is

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \text{ and } \|y\| = \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}$$

Cosine Similarity

- In fact, cosine similarity essentially is a measure of the (cosine of the) angle between x and y .
- Thus if the cosine similarity is 1, then the angle between x and y is 0° and in this case, x and y are the same except for magnitude.
- On the other hand, if cosine similarity is 0, then the angle between x and y is 90° and they do not share any terms.
- Considering, this cosine similarity can be written equivalently

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} = \hat{x} \cdot \hat{y}$$



where $\hat{x} = \frac{x}{\|x\|}$ and $\hat{y} = \frac{y}{\|y\|}$. This means that cosine similarity does not take the magnitude of the two vectors into account, when computing similarity.

- It is thus, one way normalized measurement.

Non-Metric Similarity

Example 24.7: Cosine Similarity

Suppose, we are given two documents with count of 10 words in each are shown in the form of vectors x and y as below.

$$x = [3, 2, 0, 5, 0, 0, 0, 2, 0, 0] \text{ and } y = [1, 0, 0, 0, 0, 0, 0, 1, 0, 2]$$

$$\begin{aligned} \text{Thus, } x \cdot y &= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 \\ &= 5 \end{aligned}$$

$$\|x\| = \sqrt{3^2 + 2^2 + 0 + 5^2 + 0 + 0 + 0 + 2^2 + 0 + 0} = 6.48$$

$$\|y\| = \sqrt{1^2 + 0 + 0 + 0 + 0 + 0 + 0 + 1^2 + 0 + 2^2} = 2.24$$

$$\therefore \cos(x, y) = 0.31$$

Extended Jaccard Coefficient

The extended Jaccard coefficient is denoted as EJ and defined as

$$EJ = \frac{x \cdot y}{\|x\|^2 \cdot \|y\|^2 - x \cdot y}$$

- This is also alternatively termed as [Tanimoto coefficient](#) and can be used to measure like document similarity.

Compute Extended Jaccard coefficient (EJ) for the above example 24.7.

Pearson's Correlation

- The correlation between two objects x and y gives a measure of the linear relationship between the attributes of the objects.
- More precisely, Pearson's correlation coefficient between two objects x and y is defined in the following.

$$P(x, y) = \frac{S_{xy}}{S_x \cdot S_y}$$

where $S_{xy} = \text{covariance } (x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

$S_x = \text{Standard deviation } (x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

$S_y = \text{Standard deviation } (y) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$

$\bar{x} = \text{mean } (x) = \frac{1}{n} \sum_{i=1}^n x_i$

$\bar{y} = \text{mean } (y) = \frac{1}{n} \sum_{i=1}^n y_i$

and n is the number of attributes in x and y .

Note 1: Correlation is always in the range of -1 to 1. A correlation of 1(-1) means that x and y have a perfect positive (negative) linear relationship, that is, $x_i = a \cdot y_i + b$ for some a and b .

Example 24.8: Pearson's correlation

Calculate the Pearson's correlation of the two vectors x and y as given below.

$$x = [3, 6, 0, 3, 6]$$

$$y = [1, 2, 0, 1, 2]$$

Note: Vector components can be negative values as well.

Note:

If the correlation is 0, then there is no linear relationship between the attribute of the object.

Example 24.9: Non-linear correlation

Verify that there is no linear relationship among attributes in the objects x and y given below.

$$x = [-3, -2, -1, 0, 1, 2, 3]$$

$$y = [9, 4, 1, 0, 1, 4, 9]$$

$P(x, y) = 0$, and also note $x_i = y_i^2$ for all attributes here.

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 25

Clustering techniques

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

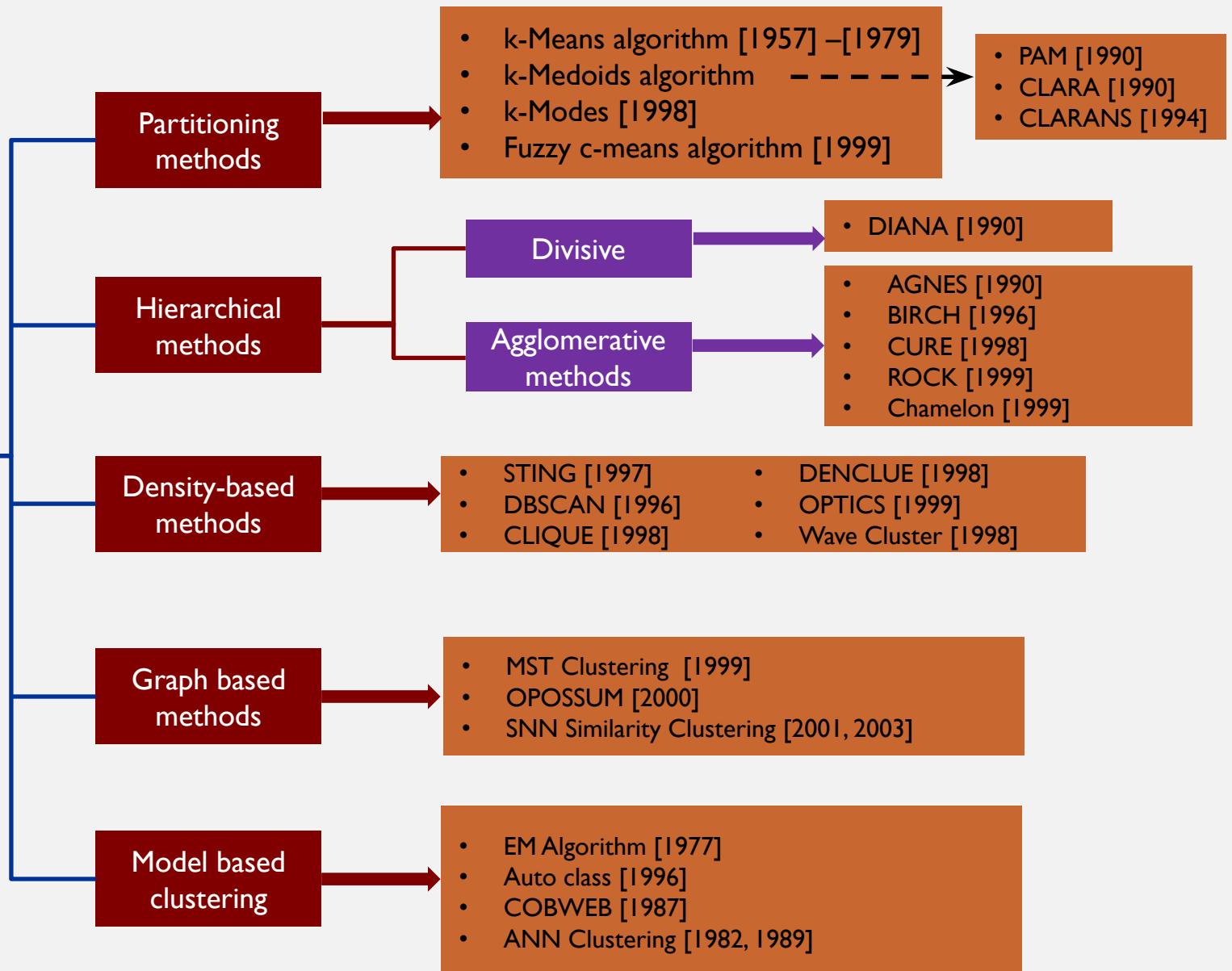
TOPICS TO BE COVERED...

- Introduction to clustering
- Similarity and dissimilarity measures
- ~~Clustering techniques~~
- ~~Partitioning algorithms~~
- Hierarchical algorithms
- Density-based algorithm

CLUSTERING TECHNIQUES

- Clustering has been studied extensively for more than 40 years and across many disciplines due to its broad applications.
- As a result, many clustering techniques have been reported in the literature.
- Let us categorize the clustering methods. In fact, it is difficult to provide a crisp categorization because many techniques overlap to each other in terms of clustering paradigms or features.
- A broad taxonomy of existing clustering methods is shown in the next slide.
- It is not possible to cover all the techniques in this lecture series. We emphasize on major techniques belong to partitioning and hierarchical algorithms.

Clustering Techniques



CLUSTERING TECHNIQUES

- In this lecture, we shall cover the following clustering techniques only.
 - Partitioning
 - k-Means algorithm
 - PAM (k-Medoids algorithm)
 - Hierarchical
 - DIANA (divisive algorithm)
 - AGNES
 - ROCK
 - Density – Based
 - DBSCAN

K-MEANS ALGORITHM

- k-Means clustering algorithm proposed by J. Hartigan and M. A. Wong [1979].
- Given a set of n distinct objects, the k-Means clustering algorithm partitions the objects into k number of clusters such that intracluster similarity is high but the intercluster similarity is low.
- In this algorithm, user has to specify k , the number of clusters and consider the objects are defined with numeric attributes and thus using any one of the distance metric to demarcate the clusters.

K-MEANS ALGORITHM

The algorithm can be stated as follows.

- First it selects k number of objects at random from the set of n objects. These k objects are treated as the **centroids or center of gravities** of k clusters.
- For each of the **remaining objects**, it is assigned to one of the **closest centroid**. Thus, it forms a **collection of objects assigned to each centroid** and is called a **cluster**.
- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).
- The assignment and update procedure is until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no assignment, etc.)

K-MEANS ALGORITHM

Algorithm 24.1: k-Means clustering

Input: D is a dataset containing n objects, k is the number of cluster

Output: A set of k clusters

Steps:

1. Randomly choose k objects from D as the initial cluster centroids.
2. **For** each of the objects in D **do**
 - Compute distance between the current objects and k cluster centroids
 - Assign the current object to that cluster to which it is closest.
3. Compute the “cluster centers” of each cluster. These become the new cluster centroids.
4. Repeat step 2-3 until the convergence criterion is satisfied
5. Stop

K-MEANS ALGORITHM

Note:

- 1) Objects are defined in terms of set of attributes. $A = \{A_1, A_2, \dots, A_m\}$ where each A_i is continuous data type.
- 2) Distance computation: Any distance such as L_1, L_2, L_3 or cosine similarity.
- 3) Minimum distance is the measure of closeness between an object and centroid.
- 4) Mean Calculation: It is the mean value of each attribute values of all objects.
- 5) Convergence criteria: Any one of the following are termination condition of the algorithm.
 - Number of maximum iteration permissible.
 - No change of centroid values in any cluster.
 - Zero (or no significant) movement(s) of object from one cluster to another.
 - Cluster quality reaches to a certain level of acceptance.

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- Table 24.1: 16 objects with two attributes A_1 and A_2 .

| A_1 | A_2 |
|-------|-------|
| 6.8 | 12.6 |
| 0.8 | 9.8 |
| 1.2 | 11.6 |
| 2.8 | 9.6 |
| 3.8 | 9.9 |
| 4.4 | 6.5 |
| 4.8 | 1.1 |
| 6.0 | 19.9 |
| 6.2 | 18.5 |
| 7.6 | 17.4 |
| 7.8 | 12.2 |
| 6.6 | 7.7 |
| 8.2 | 4.5 |
| 8.4 | 6.9 |
| 9.0 | 3.4 |
| 9.6 | 11.1 |

Fig 24.1: Plotting data of Table 24.1

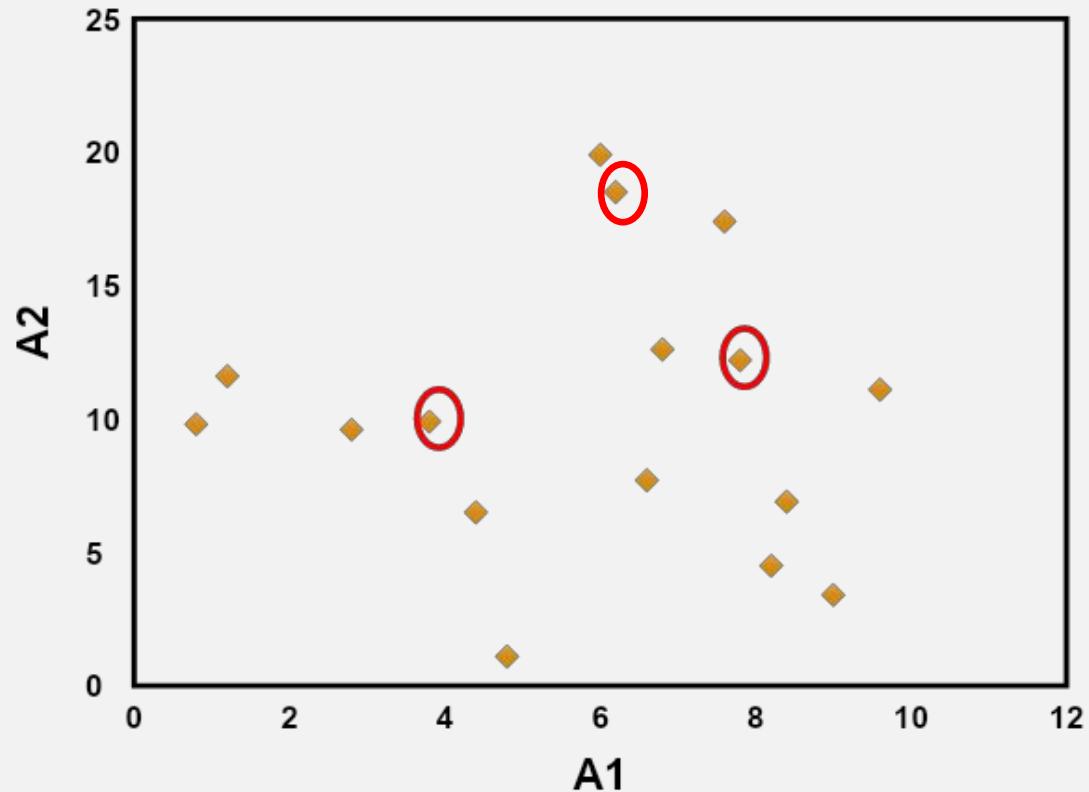


ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- Suppose, $k=3$. Three objects are chosen at random shown as circled (see Fig 24.1). These three centroids are shown below.

Initial Centroids chosen randomly

| Centroid | Objects | |
|----------|---------|------|
| | A1 | A2 |
| c_1 | 3.8 | 9.9 |
| c_2 | 7.8 | 12.2 |
| c_3 | 6.2 | 18.5 |

- Let us consider the Euclidean distance measure (L_2 Norm) as the distance measurement in our illustration.
- Let d_1 , d_2 and d_3 denote the distance from an object to c_1 , c_2 and c_3 respectively. The distance calculations are shown in Table 24.2.
- Assignment of each object to the respective centroid is shown in the right-most column and the clustering so obtained is shown in Fig 24.2.

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

Table 24.2: Distance calculation

| A ₁ | A ₂ | d ₁ | d ₂ | d ₃ | cluster |
|----------------|----------------|----------------|----------------|----------------|---------|
| 6.8 | 12.6 | 4.0 | 1.1 | 5.9 | 2 |
| 0.8 | 9.8 | 3.0 | 7.4 | 10.2 | 1 |
| 1.2 | 11.6 | 3.1 | 6.6 | 8.5 | 1 |
| 2.8 | 9.6 | 1.0 | 5.6 | 9.5 | 1 |
| 3.8 | 9.9 | 0.0 | 4.6 | 8.9 | 1 |
| 4.4 | 6.5 | 3.5 | 6.6 | 12.1 | 1 |
| 4.8 | 1.1 | 8.9 | 11.5 | 17.5 | 1 |
| 6.0 | 19.9 | 10.2 | 7.9 | 1.4 | 3 |
| 6.2 | 18.5 | 8.9 | 6.5 | 0.0 | 3 |
| 7.6 | 17.4 | 8.4 | 5.2 | 1.8 | 3 |
| 7.8 | 12.2 | 4.6 | 0.0 | 6.5 | 2 |
| 6.6 | 7.7 | 3.6 | 4.7 | 10.8 | 1 |
| 8.2 | 4.5 | 7.0 | 7.7 | 14.1 | 1 |
| 8.4 | 6.9 | 5.5 | 5.3 | 11.8 | 2 |
| 9.0 | 3.4 | 8.3 | 8.9 | 15.4 | 1 |
| 9.6 | 11.1 | 5.9 | 2.1 | 8.1 | 2 |

Fig 24.2: Initial cluster with respect to Table 24.2

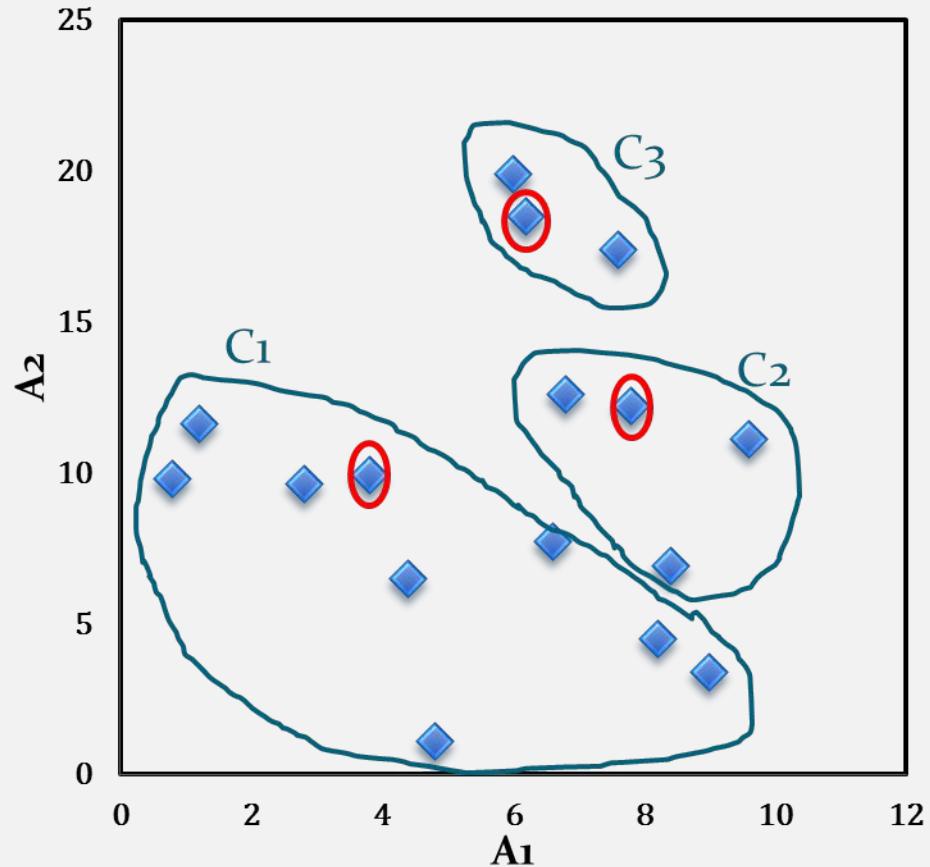


ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

The calculation new centroids of the three cluster using the mean of attribute values of A_1 and A_2 is shown in the Table below. The cluster with new centroids are shown in Fig 24.3.

Calculation of new centroids

| New Centroid | Objects | |
|--------------|---------|-------|
| | A_1 | A_2 |
| c_1 | 4.6 | 7.1 |
| c_2 | 8.2 | 10.7 |
| c_3 | 6.6 | 18.6 |

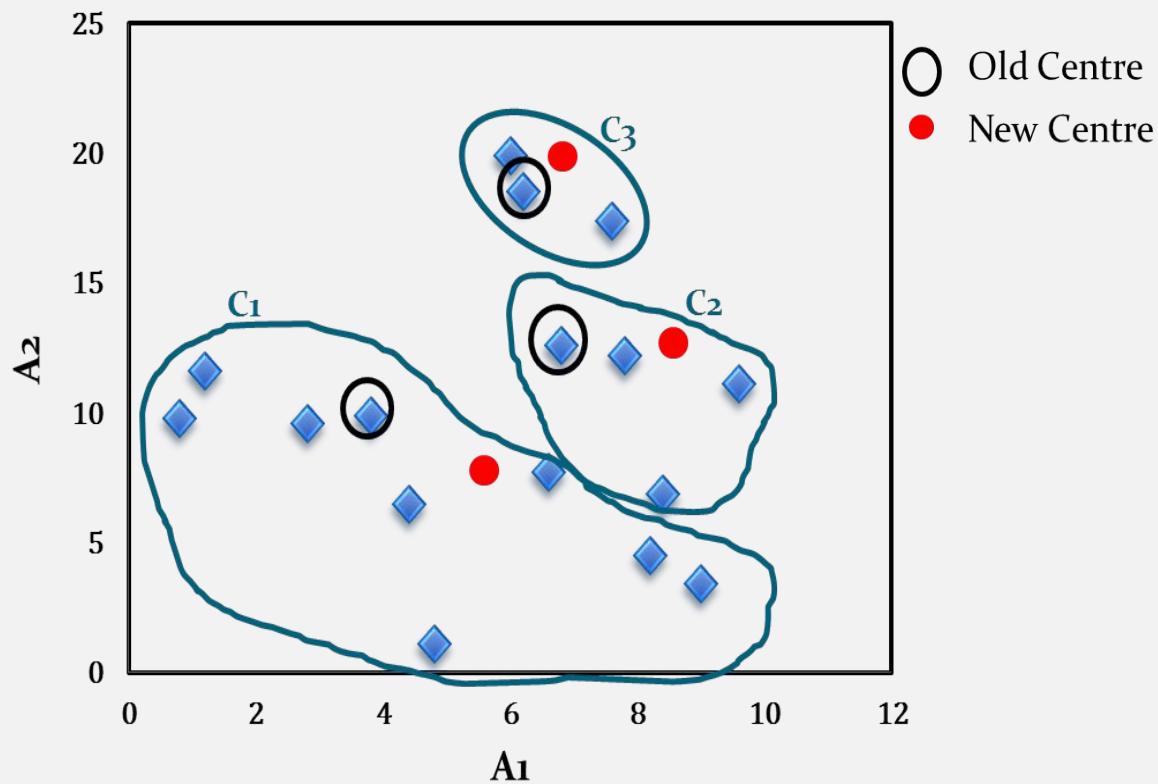


Fig 24.3: Initial cluster with new centroids

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

We next reassign the 16 objects to three clusters by determining which centroid is closest to each one. This gives the revised set of clusters shown in Fig 24.4.

Note that point p moves from cluster C_2 to cluster C_1 .

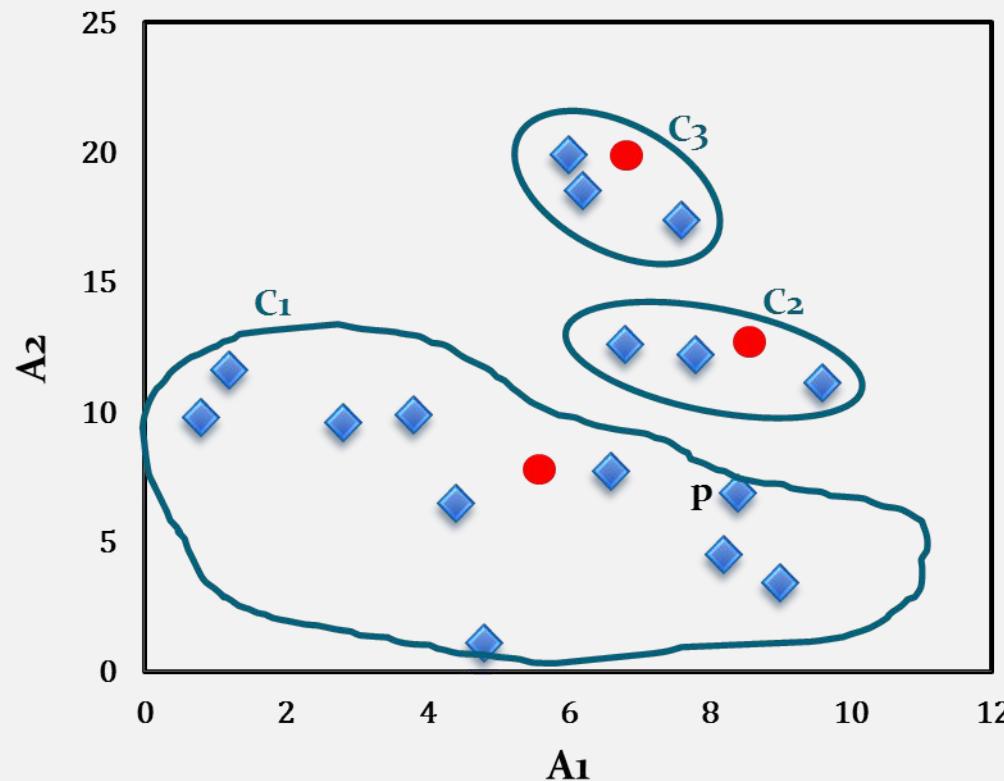


Fig 24.4: Cluster after first iteration

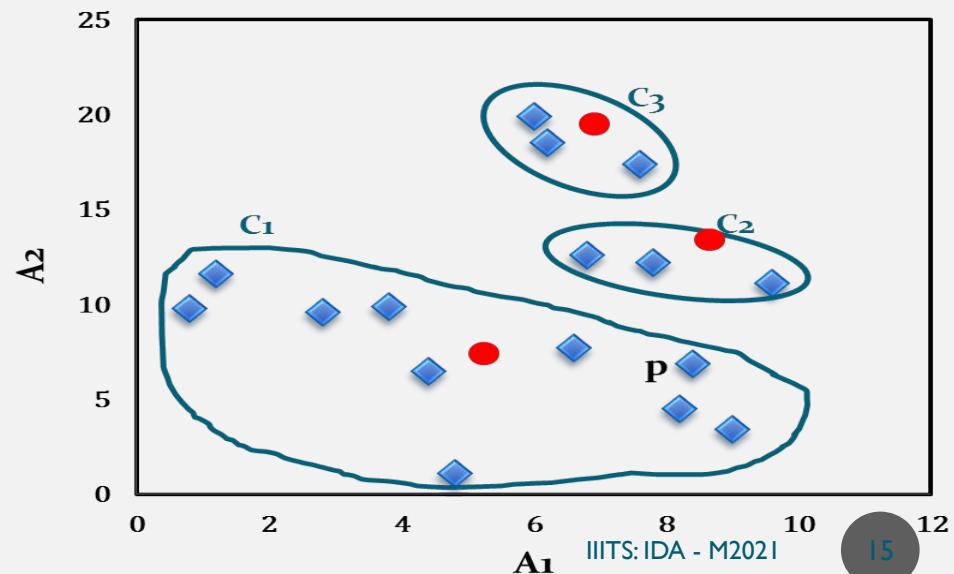
ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- The newly obtained centroids after second iteration are given in the table below. Note that the centroid c_3 remains unchanged, where c_2 and c_1 changed a little.
- With respect to newly obtained cluster centres, 16 points are reassigned again. These are the same clusters as before. Hence, their centroids also remain unchanged.
- Considering this as the termination criteria, the k-means algorithm stops here. Hence, the final cluster in Fig 24.5 is same as Fig 24.4.

Cluster centres after second iteration

| Centroid | Revised Centroids | |
|----------|-------------------|------|
| | A1 | A2 |
| c_1 | 5.0 | 7.1 |
| c_2 | 8.1 | 12.0 |
| c_3 | 6.6 | 18.6 |

Fig 24.5: Cluster after Second iteration



Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 26

Clustering techniques

Dr. Sreeja S R

Assistant Professor

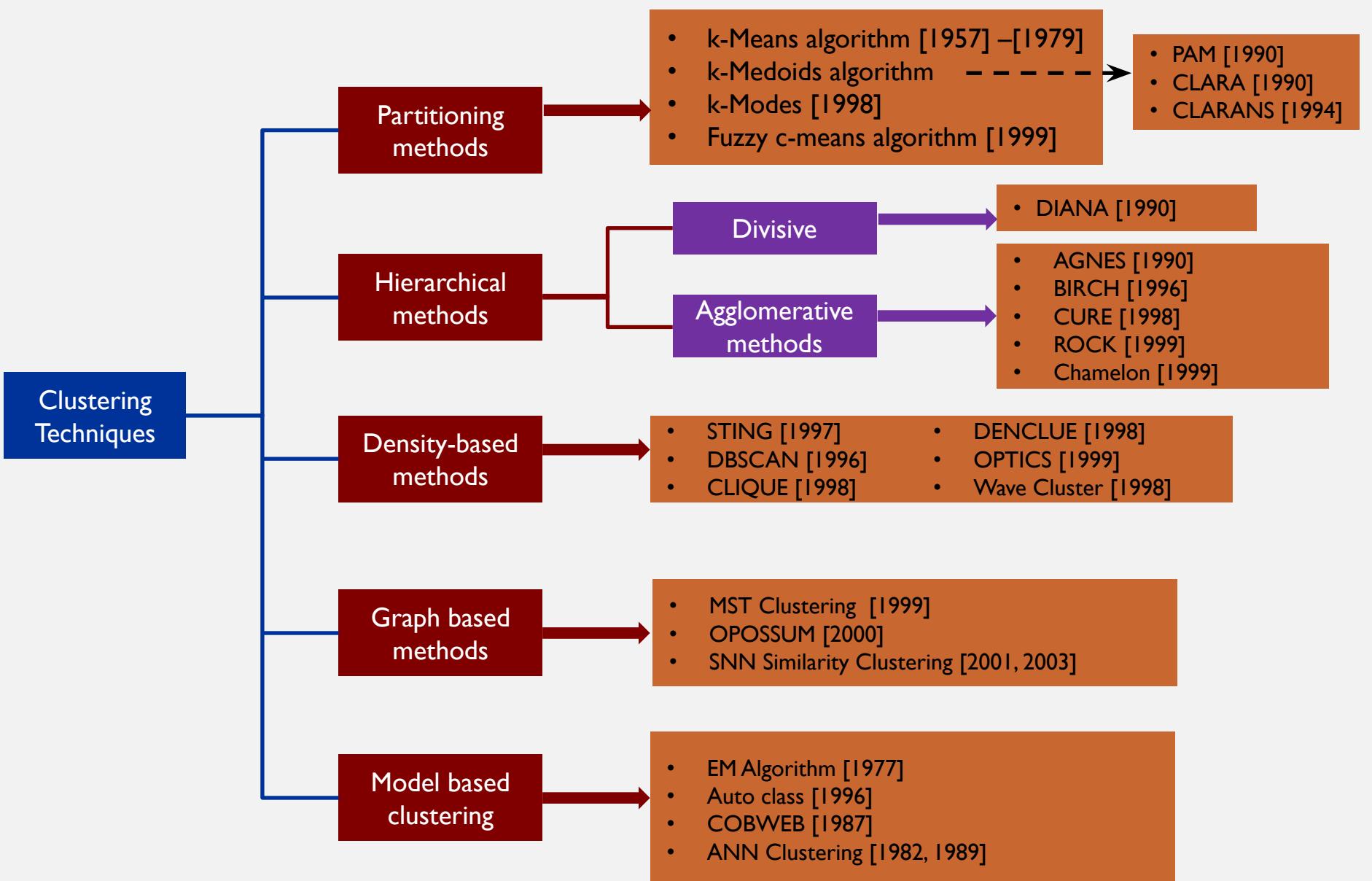
**Indian Institute of Information Technology
IIIT Sri City**

TOPICS TO BE COVERED...

- Introduction to clustering
- Similarity and dissimilarity measures
- ~~Clustering techniques~~
- ~~Partitioning algorithms~~
- Hierarchical algorithms
- Density-based algorithm

CLUSTERING TECHNIQUES

- Clustering has been studied extensively for more than 40 years and across many disciplines due to its broad applications.
- As a result, many clustering techniques have been reported in the literature.
- Let us categorize the clustering methods. In fact, it is difficult to provide a crisp categorization because many techniques overlap to each other in terms of clustering paradigms or features.
- A broad taxonomy of existing clustering methods is shown in the next slide.
- It is not possible to cover all the techniques in this lecture series. We emphasize on major techniques belong to partitioning and hierarchical algorithms.



CLUSTERING TECHNIQUES

- In this lecture, we shall cover the following clustering techniques only.
 - Partitioning
 - k-Means algorithm
 - PAM (k-Medoids algorithm)
 - Hierarchical
 - DIANA (divisive algorithm)
 - AGNES
 - ROCK
 - Density – Based
 - DBSCAN

K-MEANS ALGORITHM

- k-Means clustering algorithm proposed by J. Hartigan and M. A. Wong [1979].
- Given a set of n distinct objects, the k-Means clustering algorithm partitions the objects into k number of clusters such that intracluster similarity is high but the intercluster similarity is low.
- In this algorithm, user has to specify k , the number of clusters and consider the objects are defined with numeric attributes and thus using any one of the distance metric to demarcate the clusters.

K-MEANS ALGORITHM

The algorithm can be stated as follows.

- First it selects k number of objects at random from the set of n objects. These k objects are treated as the **centroids or center of gravities** of k clusters.
- For each of the **remaining objects**, it is assigned to one of the **closest centroid**. Thus, it forms a **collection of objects assigned to each centroid** and is called a **cluster**.
- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).
- The assignment and update procedure is until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no assignment, etc.)

K-MEANS ALGORITHM

Algorithm 24.1: k-Means clustering

Input: D is a dataset containing n objects, k is the number of cluster

Output: A set of k clusters

Steps:

1. Randomly choose k objects from D as the initial cluster centroids.
2. **For** each of the objects in D **do**
 - Compute distance between the current objects and k cluster centroids
 - Assign the current object to that cluster to which it is closest.
3. Compute the “cluster centers” of each cluster. These become the new cluster centroids.
4. Repeat step 2-3 until the convergence criterion is satisfied
5. Stop

K-MEANS ALGORITHM

Note:

- 1) Objects are defined in terms of set of attributes. $A = \{A_1, A_2, \dots, A_m\}$ where each A_i is continuous data type.
- 2) Distance computation: Any distance such as L_1, L_2, L_3 or cosine similarity.
- 3) Minimum distance is the measure of closeness between an object and centroid.
- 4) Mean Calculation: It is the mean value of each attribute values of all objects.
- 5) Convergence criteria: Any one of the following are termination condition of the algorithm.
 - Number of maximum iteration permissible.
 - No change of centroid values in any cluster.
 - Zero (or no significant) movement(s) of object from one cluster to another.
 - Cluster quality reaches to a certain level of acceptance.

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- Table 24.1: 16 objects with two attributes A_1 and A_2 .

| A_1 | A_2 |
|-------|-------|
| 6.8 | 12.6 |
| 0.8 | 9.8 |
| 1.2 | 11.6 |
| 2.8 | 9.6 |
| 3.8 | 9.9 |
| 4.4 | 6.5 |
| 4.8 | 1.1 |
| 6.0 | 19.9 |
| 6.2 | 18.5 |
| 7.6 | 17.4 |
| 7.8 | 12.2 |
| 6.6 | 7.7 |
| 8.2 | 4.5 |
| 8.4 | 6.9 |
| 9.0 | 3.4 |
| 9.6 | 11.1 |

Fig 24.1: Plotting data of Table 24.1

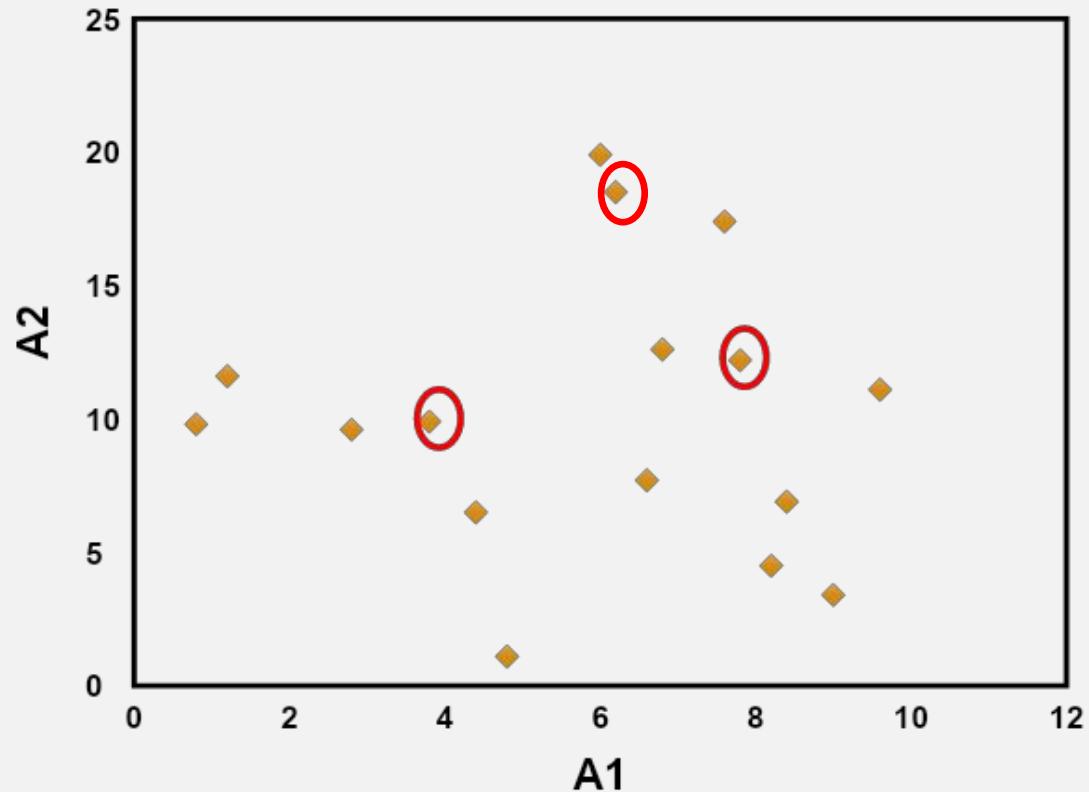


ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- Suppose, $k=3$. Three objects are chosen at random shown as circled (see Fig 24.1). These three centroids are shown below.

Initial Centroids chosen randomly

| Centroid | Objects | |
|----------|---------|------|
| | A1 | A2 |
| c_1 | 3.8 | 9.9 |
| c_2 | 7.8 | 12.2 |
| c_3 | 6.2 | 18.5 |

- Let us consider the Euclidean distance measure (L_2 Norm) as the distance measurement in our illustration.
- Let d_1 , d_2 and d_3 denote the distance from an object to c_1 , c_2 and c_3 respectively. The distance calculations are shown in Table 24.2.
- Assignment of each object to the respective centroid is shown in the right-most column and the clustering so obtained is shown in Fig 24.2.

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

Table 24.2: Distance calculation

| A ₁ | A ₂ | d ₁ | d ₂ | d ₃ | cluster |
|----------------|----------------|----------------|----------------|----------------|---------|
| 6.8 | 12.6 | 4.0 | 1.1 | 5.9 | 2 |
| 0.8 | 9.8 | 3.0 | 7.4 | 10.2 | 1 |
| 1.2 | 11.6 | 3.1 | 6.6 | 8.5 | 1 |
| 2.8 | 9.6 | 1.0 | 5.6 | 9.5 | 1 |
| 3.8 | 9.9 | 0.0 | 4.6 | 8.9 | 1 |
| 4.4 | 6.5 | 3.5 | 6.6 | 12.1 | 1 |
| 4.8 | 1.1 | 8.9 | 11.5 | 17.5 | 1 |
| 6.0 | 19.9 | 10.2 | 7.9 | 1.4 | 3 |
| 6.2 | 18.5 | 8.9 | 6.5 | 0.0 | 3 |
| 7.6 | 17.4 | 8.4 | 5.2 | 1.8 | 3 |
| 7.8 | 12.2 | 4.6 | 0.0 | 6.5 | 2 |
| 6.6 | 7.7 | 3.6 | 4.7 | 10.8 | 1 |
| 8.2 | 4.5 | 7.0 | 7.7 | 14.1 | 1 |
| 8.4 | 6.9 | 5.5 | 5.3 | 11.8 | 2 |
| 9.0 | 3.4 | 8.3 | 8.9 | 15.4 | 1 |
| 9.6 | 11.1 | 5.9 | 2.1 | 8.1 | 2 |

Fig 24.2: Initial cluster with respect to Table 24.2

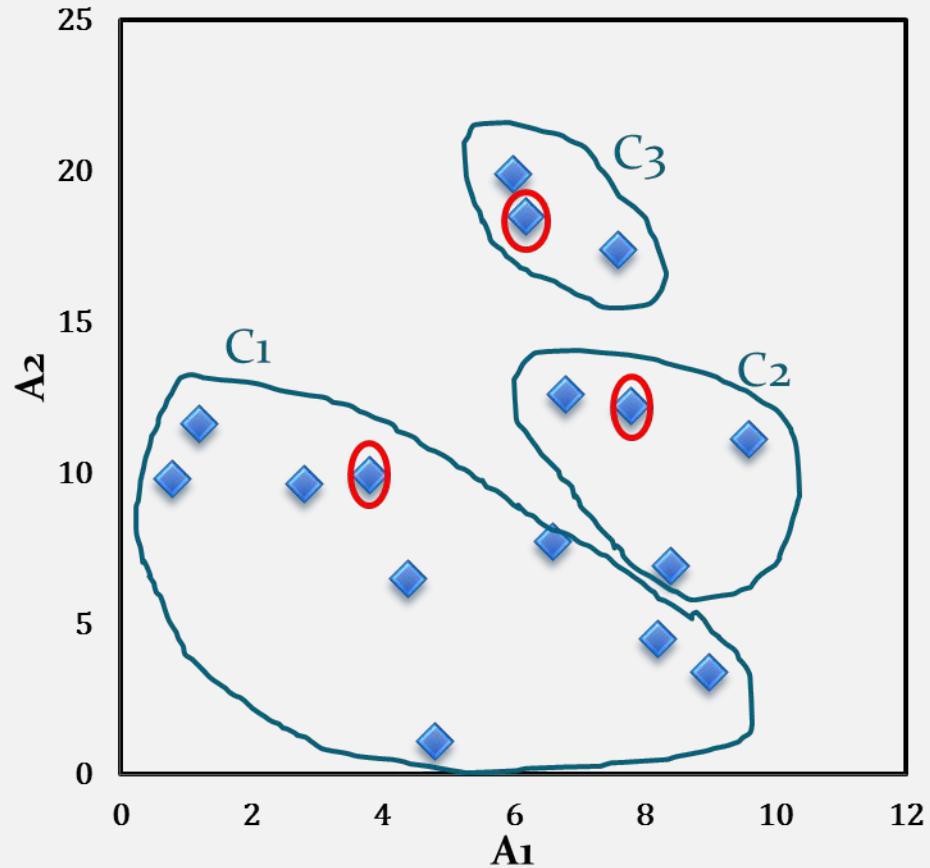


ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

The calculation new centroids of the three cluster using the mean of attribute values of A_1 and A_2 is shown in the Table below. The cluster with new centroids are shown in Fig 24.3.

Calculation of new centroids

| New Centroid | Objects | |
|--------------|---------|-------|
| | A_1 | A_2 |
| c_1 | 4.6 | 7.1 |
| c_2 | 8.2 | 10.7 |
| c_3 | 6.6 | 18.6 |

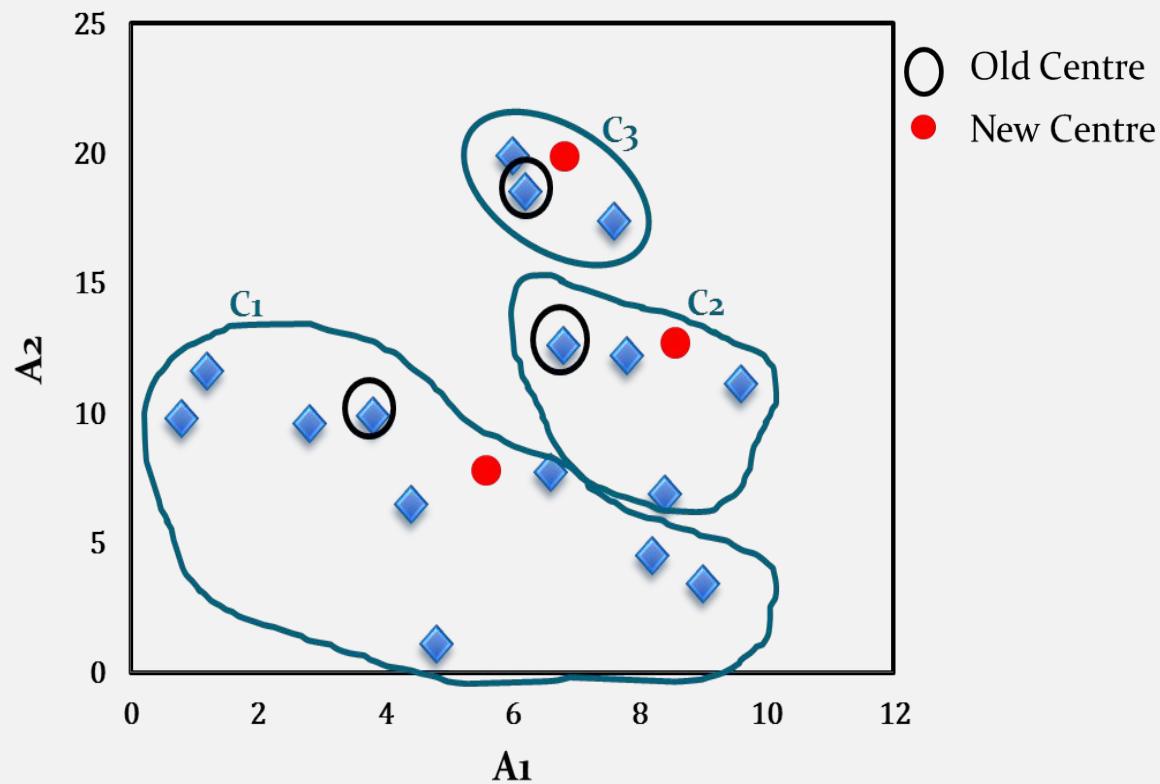


Fig 24.3: Initial cluster with new centroids

ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

We next reassign the 16 objects to three clusters by determining which centroid is closest to each one. This gives the revised set of clusters shown in Fig 24.4.

Note that point p moves from cluster C_2 to cluster C_1 .

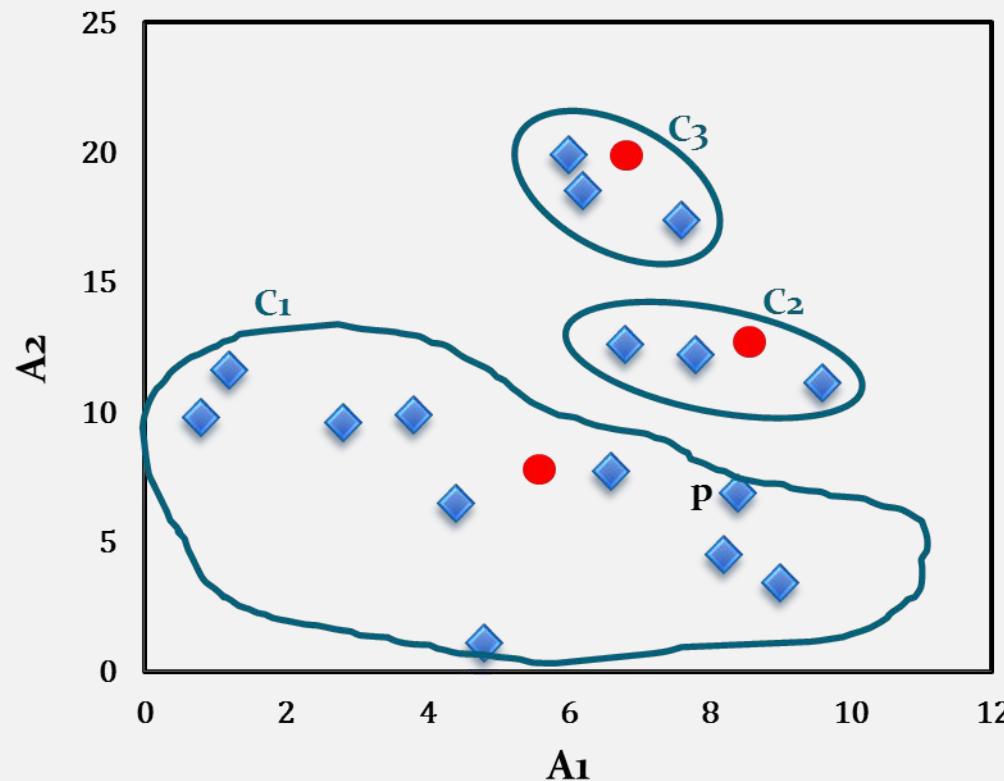


Fig 24.4: Cluster after first iteration

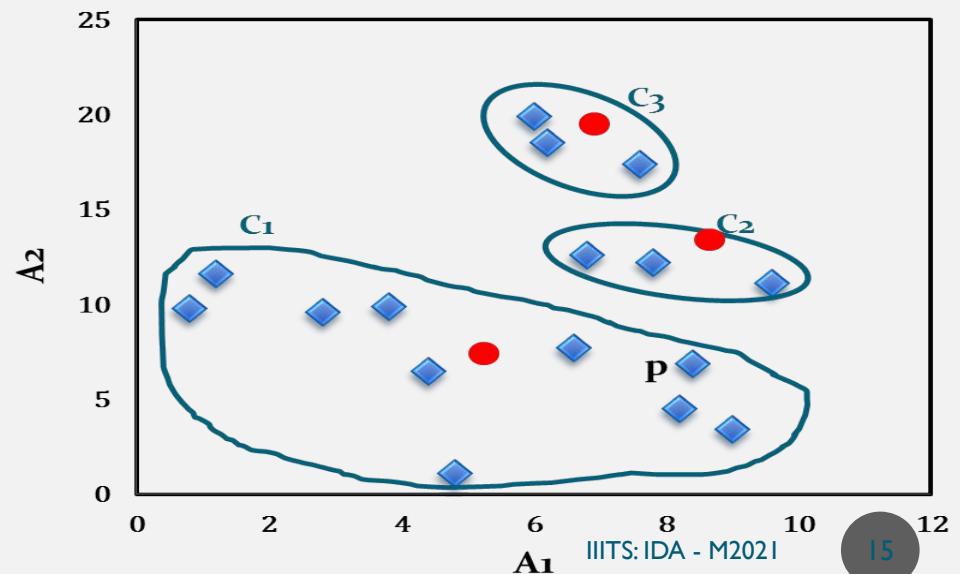
ILLUSTRATION OF K-MEANS CLUSTERING ALGORITHMS

- The newly obtained centroids after second iteration are given in the table below. Note that the centroid c_3 remains unchanged, where c_2 and c_1 changed a little.
- With respect to newly obtained cluster centres, 16 points are reassigned again. These are the same clusters as before. Hence, their centroids also remain unchanged.
- Considering this as the termination criteria, the k-means algorithm stops here. Hence, the final cluster in Fig 24.5 is same as Fig 24.4.

Cluster centres after second iteration

| Centroid | Revised Centroids | |
|----------|-------------------|------|
| | A1 | A2 |
| c_1 | 5.0 | 7.1 |
| c_2 | 8.1 | 12.0 |
| c_3 | 6.6 | 18.6 |

Fig 24.5: Cluster after Second iteration



COMMENTS ON K-MEANS ALGORITHM

Let us analyse the k-Means algorithm and discuss the pros and cons of the algorithm. We shall refer to the following notations in our discussion.

- **Notations:**

- x : an object under clustering
- n : number of objects under clustering
- C_i : the i -th cluster
- c_i : the centroid of cluster C_i
- n_i : number of objects in the cluster C_i
- c : denotes the centroid of all objects
- k : number of clusters

COMMENTS ON K-MEANS ALGORITHM

1. Value of k:

- The k-means algorithm produces only one set of clusters, for which, user must specify the desired number, k of clusters.
- In fact, k should be the **best guess** on the number of clusters present in the given data. Choosing the best value of k for a given dataset is, therefore, an issue.
- We may not have an idea about the possible number of clusters for high dimensional data, and for data that are not scatter-plotted.
- Further, possible number of clusters is hidden or ambiguous in image, audio, video and multimedia clustering applications etc.
- There is no principled way to know what the value of k ought to be. We may try with successive value of k starting with 2.
- The process is stopped when two consecutive k values produce more-or-less identical results (with respect to some cluster quality estimation).

COMMENTS ON K-MEANS ALGORITHM

Example 24.1: k versus cluster quality

- Usually, there is some objective function to be met as a goal of clustering. One such objective function is **sum-square-error** denoted by **SSE** and defined as

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} (x - c_i)^2$$

- Here, $x - c_i$ denotes the error, if x is in cluster C_i with cluster centroid c_i .
- Usually, this error is measured as distance norms like L_1 , L_2 , L_3 or Cosine similarity, etc.

COMMENTS ON K-MEANS ALGORITHM

Example 24.1: k versus cluster quality

- With reference to an arbitrary experiment, suppose the following results are obtained.

| k | SSE |
|---|------|
| 1 | 62.8 |
| 2 | 12.3 |
| 3 | 9.4 |
| 4 | 9.3 |
| 5 | 9.2 |
| 6 | 9.1 |
| 7 | 9.05 |
| 8 | 9.0 |

- With respect to this observation, we can choose the value of $k \approx 3$, as with this smallest value of k it gives reasonably good result.
- Note: If $k = n$, then SSE=0; However, the cluster is useless!

COMMENTS ON K-MEANS ALGORITHM

2. Choosing initial centroids:

- Another requirement in the k-Means algorithm to choose initial cluster centroid for each k would be clusters.
- It is observed that the k-Means algorithm terminate whatever be the initial choice of the cluster centroids.
- It is also observed that initial choice influences the ultimate cluster quality. In other words, the result may be trapped into local optima, if initial centroids are chosen properly.
- One technique that is usually followed to avoid the above problem is to choose initial centroids in multiple runs, each with a different set of randomly chosen initial centroids, and then select the best cluster (with respect to some quality measurement criterion, e.g. SSE).
- However, this strategy suffers from the combinational explosion problem due to the number of all possible solutions.

COMMENTS ON K-MEANS ALGORITHM

2. Choosing initial centroids:

- A detail calculation reveals that there are $c(n, k)$ possible combinations to examine the search of global optima.

$$c(n, k) = \frac{1}{k!} \sum_{i=1}^k (-1)^{k-i} \binom{k}{i} (i)^n$$

- For example, there are $o(10^{10})$ different ways to cluster 20 items into 4 clusters!
- Thus, the strategy having its own limitation is practical only if
 - 1) The sample is negatively small ($\sim 100\text{-}1000$), and
 - 2) k is relatively small compared to n (i.e.. $k \ll n$).

COMMENTS ON K-MEANS ALGORITHM

3. Distance Measurement:

- To assign a point to the closest centroid, we need a proximity measure that should quantify the notion of “closest” for the objects under clustering.
- Usually Euclidean distance (L_2 norm) is the best measure when object points are defined in n-dimensional Euclidean space.
- Other measure namely cosine similarity is more appropriate when objects are of document type.
- Further, there may be other type of proximity measures that appropriate in the context of applications.
- For example, Manhattan distance (L_1 norm), Jaccard measure, etc.

COMMENTS ON K-MEANS ALGORITHM

3. Distance Measurement:

Thus, in the context of different measures, the **sum-of-squared error** (i.e., objective function/convergence criteria) of a clustering can be stated as under.

Data in Euclidean space (L_2 norm):

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} (c_i - x)^2$$

Data in Euclidean space (L_1 norm):

The Manhattan distance (L_1 norm) is used as a proximity measure, where the objective is to minimize the **sum-of-absolute error** denoted as **SAE** and defined as

$$SAE = \sum_{i=1}^k \sum_{x \in C_i} |c_i - x|$$

COMMENTS ON K-MEANS ALGORITHM

Distance with document objects

Suppose a set of n document objects is defined as d document term matrix (DTM) (a typical look is shown in the below form).

| Document | Term | | | |
|----------|-------|-------|-------|-------|
| | t_1 | t_2 | t_3 | t_n |
| D_1 | | | | |
| D_2 | | | | |
| | | | | |
| D_n | | | | |

Here, the objective function, which is called Total cohesion denoted as TC and defined as

$$TC = \sum_{i=1}^k \sum_{x \in C_i} \cos(x, c_i)$$

where $\cos(x, c_i) = \frac{x \cdot c_i}{\|x\| \|c_i\|}$

$$x \cdot c_i = \sum_j x_j c_{ij} \quad \text{and} \quad \|x\| = \sqrt{\sum_j^p x_j^2}$$

$$\hat{x} = \sum_{j=1}^p \hat{x}_j \quad \hat{c}_i = \sum_{j=1}^p \hat{c}_{ij} \quad \|\hat{c}_{ij}\| = \sqrt{\sum_j^p \hat{c}_{ij}^2}$$

COMMENTS ON K-MEANS ALGORITHM

Note: The criteria of objective function with different proximity measures

1. SSE (using L_2 norm) : To **minimize** the SSE.
2. SAE (using L_1 norm) : To **minimize** the SAE.
3. TC(using cosine similarity) : To **maximize** the TC.

COMMENTS ON K-MEANS ALGORITHM

4. Type of objects under clustering:

- The k-Means algorithm can be applied only when the mean of the cluster is defined (hence it named **k-Means**). The cluster mean (also called centroid) of a cluster C_i is defined as

$$c_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

- In other words, the mean calculation assumed that each object is defined with numerical attribute(s). Thus, we cannot apply the k-Means to objects which are defined with categorical attributes.
- More precisely, the k-means algorithm require some definition of cluster mean exists, but not necessarily it does have as defined in the above equation.
- In fact, the k-Means is a very general clustering algorithm and can be used with a wide variety of data types, such as documents, time series, etc.

COMMENTS ON K-MEANS ALGORITHM

Note:

- 1) When SSE (L_2 norm) is used as objective function and the objective is to minimize, then the cluster centroid (i.e. mean) is the mean value of the objects in the cluster.
- 2) When the objective function is defined as SAE (L_1 norm), minimizing the objective function implies the cluster centroid as the median of the cluster.

The above two interpretations can be readily verified as given in the next slide.

COMMENTS ON K-MEANS ALGORITHM

Case 1: SSE

We know,

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} (c_i - x)^2$$

To minimize SSE means, $\frac{\partial(SSE)}{\partial c_i} = 0$

Thus,

$$\frac{\partial}{\partial c_i} \left(\sum_{i=1}^k \sum_{x \in C_i} (c_i - x)^2 \right) = 0$$

Or,

$$\sum_{i=1}^k \sum_{x \in C_i} \frac{\partial}{\partial c_i} (c_i - x)^2 = 0$$

COMMENTS ON K-MEANS ALGORITHM

Or,

$$\sum_{x \in C_i} 2(c_i - x) = 0$$

Or,

$$n_i \cdot c_i = \sum_{x \in C_i} x$$

Or,

$$c_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

- Thus, **the best centroid for minimizing SSE of a cluster is the mean of the objects in the cluster.**

COMMENTS ON K-MEANS ALGORITHM

Case 2: SAE

We know,

$$SAE = \sum_{i=1}^k \sum_{x \in C_i} |c_i - x|$$

To minimize SAE means, $\frac{\partial(SAE)}{\partial c_i} = 0$

Thus,

$$\frac{\partial}{\partial c_i} \left(\sum_{i=1}^k \sum_{x \in C_i} |c_i - x| \right) = 0$$

Or,

$$\sum_{i=1}^k \sum_{x \in C_i} \frac{\partial}{\partial c_i} |c_i - x| = 0$$

COMMENTS ON K-MEANS ALGORITHM

Or,

$$\sum_{x \in C_i} \left\{ (x - c_i) \Big|_{if \ x > c_i} + (c_i - x) \Big|_{if \ c_i > x} \right\} = 0$$

Solving the above equation, we get

$$c_i = median \{x | x \in C_i\}$$

- Thus, the best centroid for minimizing SAE of a cluster is the median of the objects in the cluster.

COMMENTS ON K-MEANS ALGORITHM

5. Complexity analysis of k-Means algorithm

Let us analyse the time and space complexities of k-Means algorithm.

Time complexity:

The time complexity of the k-Means algorithm can be expressed as

$$T(n) = O(n \times m \times k \times l)$$

where n = number of objects

m = number of attributes in the object definition

k = number of clusters

l = number of iterations.

Thus, time requirement is a linear order of number of objects and the algorithm runs in a modest time if $k \ll n$ and $l \ll n$ (the iteration can be moderately controlled to check the value of l).

COMMENTS ON K-MEANS ALGORITHM

5. Complexity analysis of k-Means algorithm

Space complexity: The storage complexity can be expressed as follows.

It requires $n \times m$ space to store the objects and $n \times k$ space to store the proximity measure from n objects to the centroids of k clusters.

Thus the total storage complexity is

$$S(n) = O(n \times (m + k))$$

That is, space requirement is in the linear order of n if $k \ll n$.

COMMENTS ON K-MEANS ALGORITHM

6. Final comments:

Advantages:

- k-Means is simple and can be used for a wide variety of object types.
- It is also efficient both from storage requirement and execution time point of views. By saving distance information from one iteration to the next, the actual number of distance calculations, that must be made can be reduced (specially, as it reaches towards the termination).

Limitations:

- The k-Means is not suitable for all types of data. For example, k-Means does not work on categorical data because mean cannot be defined.
- k-means finds a local optima and may actually minimize the global optimum.

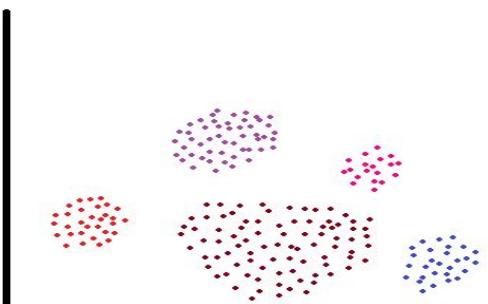
COMMENTS ON K-MEANS ALGORITHM

6. Final comments:

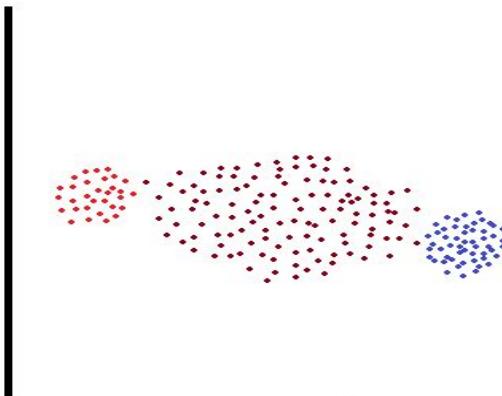
Limitations :

- k-means has trouble clustering data that contains outliers. When the SSE is used as objective function, outliers can unduly influence the cluster that are produced. More precisely, in the presence of outliers, the cluster centroids, in fact, not truly as representative as they would be otherwise. It also influence the SSE measure as well.
- k-Means algorithm cannot handle non-globular clusters, clusters of different sizes and densities (see Fig 24.6 in the next slide).
- k-Means algorithm not really beyond the scalability issue (and not so practical for large databases).

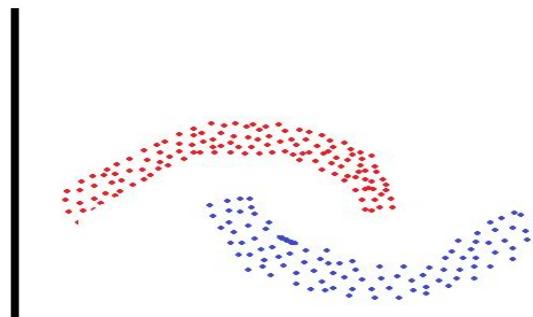
COMMENTS ON K-MEANS ALGORITHM



Cluster with different sizes



Cluster with different densities



Non-convex shaped clusters

Fig 24.6: Some failure instance of k-Means algorithm

DIFFERENT VARIANTS OF K-MEANS ALGORITHM

There are quite a few variants of the k-Means algorithm. These can differ in the procedure of selecting the initial k means, the calculation of proximity and strategy for calculating cluster means. Another variants of k-means to cluster categorical data.

Few variant of k-Means algorithm includes

- Bisecting k-Means (addressing the issue of initial choice of cluster means).
 - I. M. Steinbach, G. Karypis and V. Kumar “A comparison of document clustering techniques”, *Proceedings of KDD workshop on Text mining*, 2000.
- Mean of clusters (Proposing various strategies to define means and variants of means).
 - B. zhan “Generalised k-Harmonic means – Dynamic weighting of data in unsupervised learning”, *Technical report, HP Labs*, 2000.
 - A. D. Chaturvedi, P. E. Green, J. D. Carroll, “k-Modes clustering”, *Journal of classification*, Vol. 18, PP. 35-36, 2001.
 - D. Pelleg, A. Moore, “x-Means: Extending k-Means with efficient estimation of the number of clusters”, *17th International conference on Machine Learning*, 2000.

DIFFERENT VARIANTS OF K-MEANS ALGORITHM

- N. B. Karayiannis, M. M. Randolph, “Non-Euclidean c-Means clustering algorithm”, *Intelligent data analysis journal*, Vol 7(5), PP 405-425, 2003.
- V. J. Olivera, W. Pedrycy, “Advances in Fuzzy clustering and its applications”, Edited book. John Wiley [2007]. (Fuzzy c-Means algorithm).
- A. K. Jain and R. C. Dubes, “Algorithms for clustering Data”, Prentice Hall, 1988.

Online book at http://www.cse.msu.edu/~jain/clustering_Jain_Dubes.pdf

- A. K. Jain, M. N. Munty and P. J. Flynn, “Data clustering: A Review”, *ACM computing surveys*, 31(3), 264-323 [1999]. Also available online.

THE K-MEDOIDS ALGORITHM

Now, we shall study a variant of partitioning algorithm called k-Medoids algorithm.

Motivation: We have learnt that the k-Means algorithm is sensitive to outliers because an object with an “extremely large value” may substantially distort the distribution. The effect is particularly exacerbated due to the use of the SSE (sum-of-squared error) objective function. The k-Medoids algorithm aims to diminish the effect of outliers.

Basic concepts:

- The basic concepts of this algorithm is to **select an object as a cluster center** (one representative object per cluster) instead of taking the mean value of the objects in a cluster (as in k-Means algorithm).
- We call this cluster representative as a **cluster medoid** or simply **medoid**.
 1. Initially, it selects a random set of k objects as the set of medoids.
 2. Then at each step, all objects from the set of objects, which are not currently medoids are examined one by one to see if they should be medoids.

THE K-MEDOIDS ALGORITHM

- That is, the k-Medoids algorithm **determines** whether there is an object that should replace one of the current medoids.
- This is accomplished by looking all pair of medoid, non-medoid objects, and then choosing a pair that improves the objective function of clustering the best and exchange them.
- The sum-of-absolute error (SAE) function is used as the objective function.

$$SAE = \sum_{i=1}^k \sum_{x \in C_i, x \notin M \text{ and } c_m \in M} |x - c_m|$$

Where c_m denotes a medoid

M is the set of all medoids at any instant

x is an object belongs to set of non-medoid object, that is, x belongs to some cluster and is not a medoid. i.e. $x \in C_i, x \notin M$

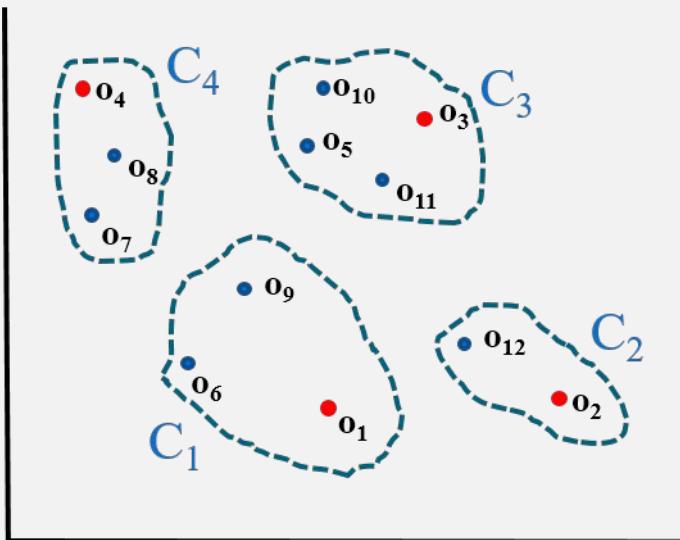
PAM (PARTITIONING AROUND MEDOIDS)

- For a given set of medoids, at any iteration, it select and exchange which has minimum SAE.
- The procedure terminates, if there is no change in SAE in successive iteration (i.e. there is no change in medoid).
- This k-Medoids algorithm is also known as PAM (Partitioning around Medoids).

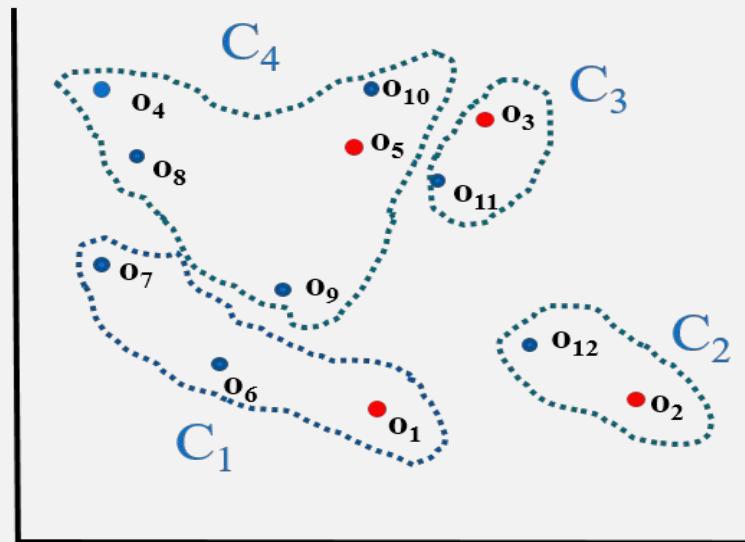
Illustration of PAM

- Suppose, there are set of 12 objects $O(o_1, o_2, \dots, o_{12})$ and we are to cluster them into four clusters. At any instant, the four cluster C_1, C_2, C_3 and C_4 are shown in Fig. 24.7 (a). Also assume that o_1, o_2, o_3 , and o_4 are the medoids in the clusters C_1, C_2, C_3 and C_4 , respectively. For this clustering we can calculate SAE.
- There are many ways to choose a non-medoid object to be replaced by any one medoid object. Out of these, suppose, if o_5 is considered as candidate medoid instead of o_4 , then it gives the lowest SAE. Thus, the new set of medoids would be o_1, o_2, o_3 and o_5 . The new cluster is shown in Fig. 24.7 (b).

PAM (PARTITIONING AROUND MEDOIDS)



(a) Cluster with o_1, o_2, o_3 , and o_4 as medoids



(b) Cluster after swapping o_4 and o_5 (o_5 becomes the new medoid).

Fig 24.7: Illustration of PAM

PAM (PARTITIONING AROUND MEDOIDS)

PAM algorithm is thus a procedure of iterative selection of medoids and it is precisely stated in Algorithm 24.2.

Algorithm 24.2: PAM

Input: Database of objects D.

k, the number of desired clusters.

Output: Set of k clusters

Steps:

1. Arbitrarily select k medoids from D.
2. **For** each object o_i not a medoid **do**
3. **For** each medoid o_j **do**
4. Let $M = \{o_1, o_2, \dots, o_{i-1}, o_i, o_{i+1}, o_k\}$ //Set of current medoids
 $M' = \{o_1, o_2, \dots, o_{j-1}, o_j, o_{j+1}, o_k\}$ //set of medoids but swap with non-medoids o_j
5. Calculate $\text{cost}(o_i, o_j) = SAE|_M - SAE|_{M'}$.

PAM (PARTITIONING AROUND MEDOIDS)

Algorithm 24.2: PAM

7. Find o_i, o_j for which the $\text{cost}(o_i, o_j)$ is the smallest.
8. Replace o_i with o_j and accordingly update the set M .
9. Repeat step 2 - step 8 until $\text{cost}(o_i, o_i) \leq 0$.
10. Return the cluster with M as the set of cluster centers.
11. Stop

COMMENTS ON PAM

1. Comparing k-Means with k-Medoids:

- Both algorithms needs to fix k , the number of cluster prior to the algorithms. Also, both algorithm arbitrarily choose the initial cluster centroids.
- The k-Medoid method is more robust than k-Means in the presence of outliers, because a medoid is less influenced by outliers than a mean.

2. Time complexity of PAM:

- For each iteration, PAM consider $k(n - k)$ pairs of object o_i, o_j for which a cost $\text{cost}(o_i, o_j)$ determines. Calculating the cost during each iteration requires that the cost be calculated for all other non-medoids o_j . There are $n - k$ of these. Thus, the total time complexity per iteration is $n(n - k)^2$. The total number of iterations may be quite large.

3. Applicability of PAM:

- PAM does not scale well to large database because of its computation

OTHER VARIANTS OF K-MEDOIDS ALGORITHMS

- There are some variants of PAM that are targeted mainly large datasets are CLARA (Clustering LARge Applications) and CLARANS (Clustering Large Applications based upon RANdomized Search), it is an improvement of CLARA.

References:

For PAM and CLARA:

- L. kaufman and P. J. Rousseeuw, “Finding Groups in Data: An introduction to cluster analysis”, John and Wiley, 1990.

For CLARANS:

- R. Ng and J. Han, “Efficient and effective clustering method for spatial Data mining”, Proceeding very large databases [VLDB-94], 1994.

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 27
ANOVA

Dr. Sreeja S R

Assistant Professor

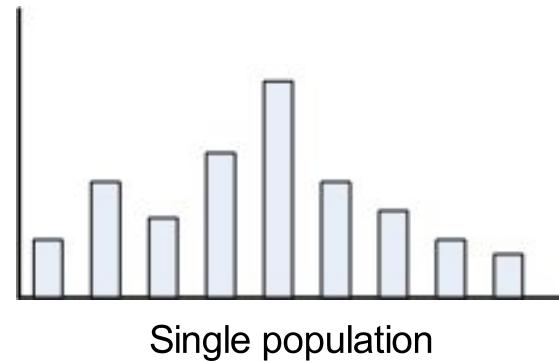
**Indian Institute of Information Technology
IIIT Sri City**

THIS PRESENTATION INCLUDES...

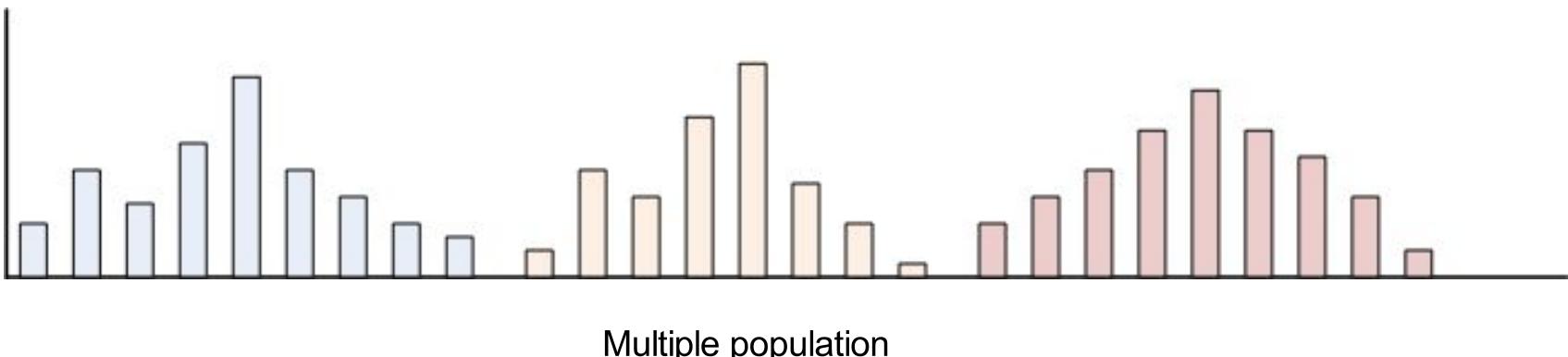
- What is “Analysis of variance”?
- Why ANOVA?
- How ANOVA?
 - *One – way ANOVA*
 - *Two-way ANOVA*

What is Analysis of Variance?

WHAT IS ANALYSIS OF VARIATION?



Single population



Multiple population

EXAMPLE : SINGLE VS. MULTIPLE POPULATION



WHAT IS THE ISSUE?

- Are the statistical inference valid?

μ

σ

EXAMPLE 1: THE ISSUE IN STATISTICAL TESTING

A recent study claims that using music in a class enhances the concentration and consequently helps students absorb more information.

- What if it affected the results of the students in a negative way?

or

- What kind of music would be a good choice for this?

We should have some proof that it actually works or not.

DESIGN OF EXPERIMENT

- The teacher decided to implement it on a smaller group of randomly selected students from **three different** classes.

Three different groups of **ten randomly selected students** from three different classrooms were taken.

Each classroom was provided with **three different environments** for students to study.

- Classroom A had **constant music** being played in the background
 - Classroom B had **variable music** being played in the background
 - Classroom C was a regular class **with no music playing**
-
- A test was conducted after one month for all the three groups and their test scores were collected.

TEST RESULT

| | Test scores of students (out of 10) | | | | | | | | | | Mean |
|--------------------------|-------------------------------------|---|---|---|---|---|---|----|---|---|------|
| Class A (constant music) | 7 | 9 | 5 | 8 | 6 | 8 | 6 | 10 | 7 | 4 | 7 |
| Class B (variable music) | 4 | 3 | 6 | 2 | 7 | 5 | 5 | 4 | 1 | 3 | 4 |
| Class C (no music) | 6 | 1 | 3 | 5 | 3 | 4 | 6 | 5 | 7 | 3 | 4.3 |
| | Grand Mean -> | | | | | | | | | | 5.1 |

OBSERVATIONS FROM THE RESULTS

- It is noticed that the mean score of students from **Group A** is definitely greater than the other two groups, so the treatment must be helpful.
- Maybe it's true, but there is also a slight chance that we happened to select the best students from class A, which resulted in better test scores (remember, the selection was done at random).
- This leads to a few questions:
 - I. How do we decide that these three groups performed differently because of the different situations and **not merely by chance?**
 2. In a statistical sense, how different are these three samples from each other?

ANALYSIS OF VARIANCE (ANOVA)

Definition 16.1

- Analysis of Variance (ANOVA) is derived from a partitioning of total variability into its component parts.
 - ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
 - ANOVA checks the impact of one or more factors by comparing the means of different samples.
-
- This technique was invented by Sir Ronald Aylmer Fisher (1921), and is often referred to as Fisher's ANOVA.

Why ANOVA?

STATISTICAL INFERENCES

- ANOVA is a statistical technique
 - It is similar in application to techniques such as t-test, z-test and χ^2 -test in that it is used to compare means and the relative variance between them.
- Why not use t-test, z-test and χ^2 -test ?
- Why analysis of variance for comparing means?



USING T-TEST

t-test is used to:

- To infer **mean of a single population**
- T-test can be used to compare two populations

However, t-test is not useful to compare mean of more than two populations

EXTENDING THE TWO POPULATION PROCEDURE

- Construct pairwise comparison on all means.

For 5 populations → 10 possible pairs. When all pairwise comparisons are made for n groups, the total number of possible combinations is $n*(n - 1)/2$.

- Considering $\alpha = 0.05$, probability of correctly failing to reject the null hypothesis for all 10 tests is $(0.95)^{10}$, assuming that the tests are independent
- Thus, the true value of α for this set of comparison is 0.4, instead of .05
- **It inflates the Type 1 error.**
- The probability that a Type I error occurs if k comparisons are made is $1-(1-\alpha)^k$; if 10 comparisons are made, the Type I error rate increases to 40%.

Kao, Lillian S. et al., “Analysis of Variance: Is There a Difference in Means and What Does It Mean?”, Journal of Surgical Research, Volume 144, Issue 1, 158 – 170, 2007

EXTENDING THE TWO POPULATION PROCEDURE

- Statistical Inference I
 - A car magazine wishes to compare the average petrol consumption of THREE models for car and has available SIX vehicles of each model.

| Model 1 | Model 2 | Model 3 |
|---------|---------|---------|
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

- There are THREE populations
- There are samples each of size six from each population

EXTENDING THE TWO POPULATION PROCEDURE

- Statistical Inference II
 - A teacher is interested in a comparison of the average percentage marks obtained in the examinations of five different subjects and has available the marks of eight students who all completed each examination.

| Subject 1 | Subject 2 | Subject 3 | Subject 4 | Subject 5 |
|-----------|-----------|-----------|-----------|-----------|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

- What is the number of populations?
- How many samples? What are their sizes?? Are each sample independent to each other?

EXAMPLE 2 : WHY ANOVA?

Consider the two sets of contrived data as shown below:

| Set 1 (Benz) | | | Set 2 (Toyota) | | |
|--------------|--------------|--------------|----------------|--------------|--------------|
| Sample 1 | Sample 2 | Sample 3 | Sample 1 | Sample 2 | Sample 3 |
| 5.7 | 9.4 | 14.2 | 3.0 | 5.0 | 11.0 |
| 5.9 | 9.8 | 14.4 | 4.0 | 7.0 | 13.0 |
| 6.0 | 10.0 | 15.0 | 6.0 | 10.0 | 16.0 |
| 6.1 | 10.2 | 15.6 | 8.0 | 13.0 | 17.0 |
| 6.3 | 10.6 | 15.8 | 9.0 | 15.0 | 18.0 |
| $y^- = 6.0$ | $y^- = 10.0$ | $y^- = 15.0$ | $y^- = 6.0$ | $y^- = 10.0$ | $y^- = 15.0$ |

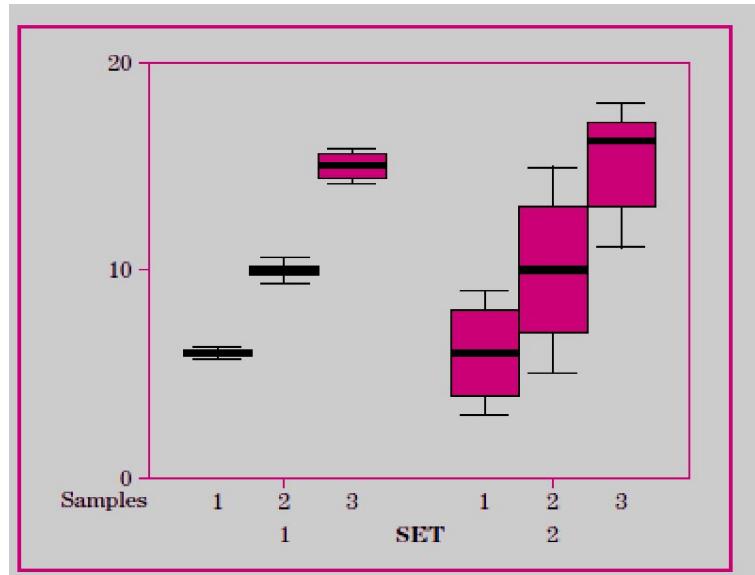
Observations:

- Looking only at the means, we can see that they are identical for the three populations in both the sets.
- Using the means alone, we would state that there is no difference between the two sets.

BOX PLOTS OF THE TWO EXPERIMENTS

Observation from Box plots

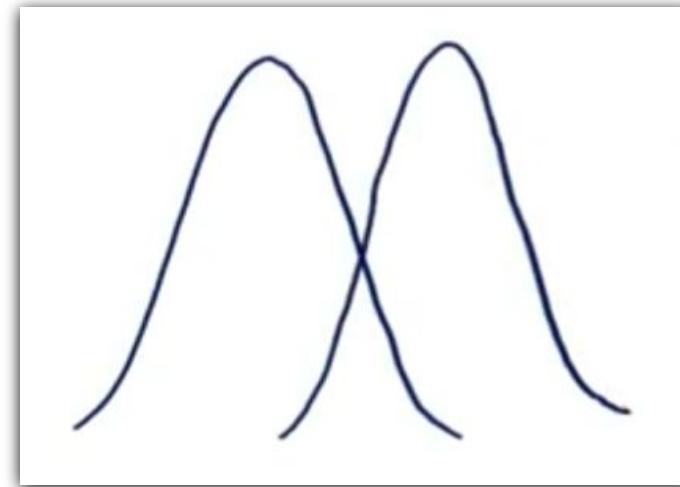
- It appears that there is stronger evidence of differences among means in Set 1 than among means in Set 2.
- The observations *within* the samples are more closely bunched in Set 1 than they are in Set 2.
- We know that **sample means from populations with smaller variances** will also be less variable.
(Central Limit Theorem)
- Thus, although the variances among the means for the two sets are identical, the variance among the observations within the individual samples is smaller for Set 1 and is the reason for the apparently stronger evidence of different means.
- This observation is the basis for using the analysis of variance for making inferences about differences among means.
- The analysis of variance is based on the **comparison of the variance among the means of the populations to the variance among sample observations within the *individual populations*.**



BETWEEN GROUP VARIABILITY

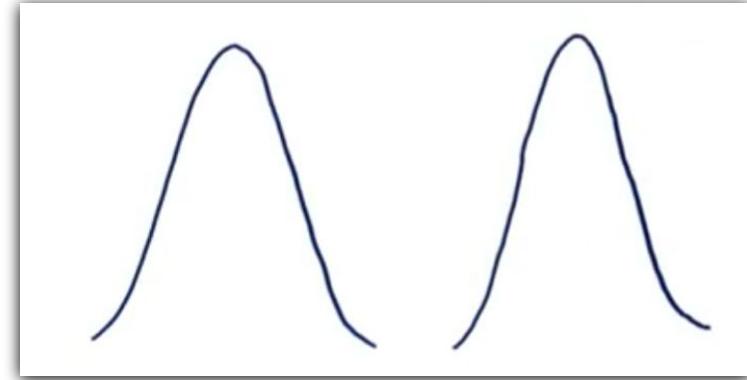
Variance among the means of the populations

- Consider the distributions of the below two samples.
- As these samples overlap, their individual means won't differ by a great margin.
- Hence, the difference between their individual means and grand mean won't be significant enough.
- Mean is a simple or arithmetic average of a range of values. There are two kinds of means that we use in ANOVA calculations, which are separate sample means (μ_1 and μ_2) and the grand mean μ
- The grand mean is the mean of sample means or the mean of all observations combined, irrespective of the sample.



BETWEEN GROUP VARIABILITY

Now consider these two sample distributions. As the samples differ from each other by a big margin, their individual means would also differ. The difference between the individual means and grand mean would therefore also be significant.

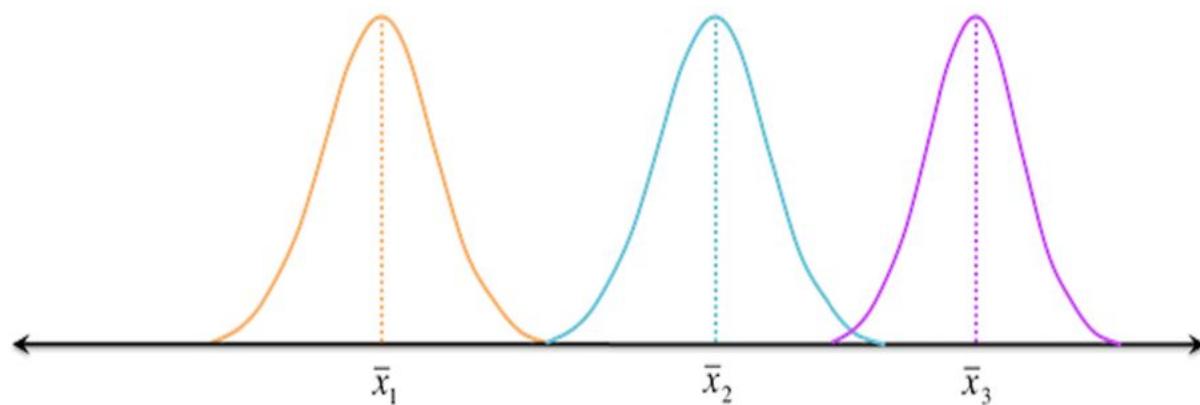
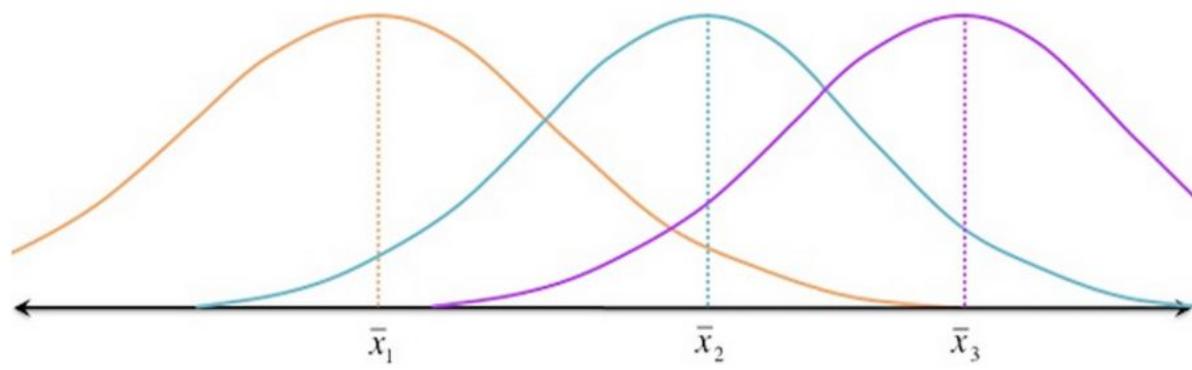


- Such variability between the distributions called *Between-group variability or variance among the means of the populations.*
- Each sample is looked at and the difference between its mean and grand mean is calculated to calculate the variability.
- If the distributions overlap or are close, the grand mean will be similar to the individual means, whereas if the distributions are far apart, difference between means and grand mean would be large.

WITHIN GROUP VARIABILITY

Variance among sample observations

Consider the given distributions of three samples. As the spread (variability) of each sample is increased, their distributions overlap and they become part of a big population.



Now consider another distribution of the same three samples but with less variability. Although the means of samples are similar to the samples in the given image, they seem to belong to different populations.

REFERENCE

- The detail material related to this lecture can be found in
Design and Analysis of Experiments (8th Edition), Douglas C. Montgomery, John Wiley & Sons, 2013.

Any question?



INTRODUCTION TO DATA ANALYTICS

Class # 28

ANOVA

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

How ANOVA?

SOME TERMINOLOGIES

- Factor
 - A characteristic under consideration, thought to influence the measured observations.
- Level (also called treatment)
 - A value of the factor

Typical data for a **Single-Factor Experiment**

| Level | Observations (RBC) | | | | Total | Mean |
|------------------|--------------------|----------|-----|-----------|-------|------|
| 1_Anemic | y_{11} | y_{12} | ... | y_{1n1} | | |
| 2_Normal | y_{21} | y_{22} | ... | y_{2n2} | | |
| 3-Erythrocytosis | ... | ... | ... | ... | | |
| ... | ... | ... | ... | ... | | |
| ... | ... | ... | ... | ... | | |
| k | y_{k1} | y_{k2} | ... | y_{knk} | | |

EXAMPLE 3: SINGLE-FACTOR ANOVA

- Draw a straight line of between 20cm and 25 cm on a sheet of plain white card (only you know its exact length).
- Collect 6 to 10 volunteers from each of Class VII, Class X and Class XII. Ask each volunteer to estimate independently the length of the line.
- Do differences in class means appear to outweigh differences within class?

What is/ are the Factor(s) and Levels here?

Example 4 : Two-Factor ANOVA

- Make a list of 10 food/household items purchased regularly by your family.
- Obtain the current prices of the items in three different shops; preferably a small 'corner' shop, a small supermarket and a large supermarket or hyper market.

What is/ are the Factor(s) and Levels here?

- Compare total shop prices.

VARIANTS OF ANOVA

Based on the number of Independent Variables and Dependent Variables considered for the study, there are different variants of ANOVA

1. **One-way ANOVA:** Only one independent variable (factor) with greater than 2 levels.
2. **Two-way ANOVA:** Two independent variables (i.e., factors).
3. **Three-way ANOVA:** Three independent variables (i.e., factors).
4. **Multivariate ANOVA:** It is used to test the significance of the effect of more independent variables.

One-way ANOVA

One-way ANOVA

- The purpose of the procedure is to compare sample means of k populations.
- In general, One-way ANOVA technique can be used to study the effect of k (> 2) levels of a single factor.
- To determine if different levels of the factor affect measured observations differently, the following hypotheses are tested.

$$H_0: \mu_i = \mu \text{ all } i = 1, 2, \dots, k$$

$$H_1: \mu_i \neq \mu \text{ some } i = 1, 2, \dots, k$$

That is, at least one equality is not satisfied

where μ_i is the population mean for a level i .

Assumptions

- When applying one-way analysis of variance, there are three key assumptions that should be satisfied as follows.
 1. The observations are obtained independently and randomly from the populations defined by the factor levels.
 2. The population at each factor level is (approximately) normally distributed.
 3. These normal populations have a common variance, σ^2 .
- Thus, for factor level i , the population is assumed to have a distribution which is $N(\mu_i, \sigma^2)$.

ONE-WAY ANOVA

| Level (Group) | Observations (One Factor) | | | | Total | Average |
|------------------|---------------------------|--|-------|--|-------|---------|
| 1 | | | | | | |
| 2 | | | | | | |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |
| k | | | | | | |
| | | | | | | |

An entry in the table (e.g., y_{ij}) represents the j^{th} observation taken under the factor at level i .

- There will be, in general, n observations under the i^{th} level.
- $y_{i\cdot}$ represents the total of the observations under the i^{th} level.
- $\bar{y}_{i\cdot}$ represent the average of the observation under the i^{th} level.
- y_g represent the grand total of all the observation under the *factor*.
- \bar{y}_g represent the average grand total of all the observation under the factor.

ONE-WAY ANOVA

Expressed symbolically,

$$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij} \quad i = 1, 2, \dots, k$$

$$\bar{y}_{i..} = \frac{y_{i\cdot}}{n_i}$$

$$y_g = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \quad \bar{y}_g = y_g / N$$

Here, N is the total observations, that is, $N = n_1 + n_2 + \dots + n_k$

OVERALL VARIABILITY IN DATA

The correlated sum of squares for each factor level

$$SS_i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \text{ for } i = 1, 2, \dots, k$$

OVERALL VARIABILITY IN DATA

The corrected sum of squares for each factor level

$$SS_i = \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_g)^2$$

Alternatively, it can be represented as,

$$SS_i = \sum_{j=1}^{n_i} y_{ij}^2 - \frac{(y_{i..})^2}{n_i}$$

OVERALL VARIABILITY IN DATA

We then calculate a pooled sum of squares

$$SS_p = \sum_{i=1}^k SS_i$$

Finally, the pooled sample of variance is

$$s_p = \frac{SS_p}{\text{pooled degree of freedom}} = \frac{SS_p}{(\sum n_i) - k}$$

Note that if the individual variances are available, the same can be computed as

$$s_p = \frac{\sum_{i=1}^k (n_i - 1)s_i^2}{(\sum n_i) - k}$$

where s_i^2 are the variances for each sample. This is also called **variance within samples** and also popularly be denoted as $\hat{\sigma}_W^2$

Example 5: Variance within Samples

- The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours, of samples of 60W electric light bulbs of three different brands.

| Lifetime of bulb | | |
|------------------|--------|--------|
| Blub 1 | Blub 2 | Blub 3 |
| 16 | 18 | 26 |
| 15 | 22 | 31 |
| 13 | 20 | 24 |
| 21 | 16 | 30 |
| 15 | 24 | 24 |

Solution : Variance within Samples

- Here, there is one factor (lifetime of the bulb) at three levels (bulb 1, bulb 2 and bulb3). Also the sample sizes are all equal (to 5).
- The sample mean and variance (divisor ($n - 1$)) for each level are as follows.

| | Lifetime of bulb | | |
|----------------|------------------|--------|--------|
| | Blub 1 | Blub 2 | Blub 3 |
| Sample Size | 5 | 5 | 5 |
| Sum | 80 | 100 | 135 |
| Sum of squares | 1316 | 2040 | 3689 |
| Mean | 16 | 20 | 27 |
| Variance | 9 | 10 | 11 |

Solution : Variance within Samples

- A pooled estimate of variance then can be calculated as follows.

$$\hat{\sigma}_W^2 = \frac{(5 - 1) \times 9 + (5 - 1) \times 10 + (5 - 1) \times 11}{5 + 5 + 5 - 3} = 10$$

- This quantity is called the **variance within samples**.
- It is an estimate of σ^2 based on $v = 5 + 5 + 5 - 3 = 12$ degrees of freedom.

Heuristic Justification of ANOVA

- From the sampling distribution of the mean, we know that a sample mean computed from a random sample of size n from a population with mean μ and variance σ^2 is a random variable with mean μ and variance σ^2/n [Central Limit Theorem].
- Let us see, what we can conclude in case of k (where $k > 1$) populations, which may have different μ_i but have the same variance σ^2 .

Heuristic Justification of ANOVA

- If the null hypothesis is true, that is, each of the μ_i has the same value, say, μ , then the distribution of each of the k sample means, \bar{y}_i will have mean μ and variance σ^2/n .
- It then follows that, if we calculate a variance using the sample means as observations,

$$\hat{\sigma}_B^2 = \sum (\bar{y}_{i\cdot} - \bar{y}_g)^2 / (k - 1)$$

- Then the quantity is an estimate of σ^2/n .
- Hence, $n\hat{\sigma}_B^2$ is an estimate of σ^2 .

Heuristic Justification of ANOVA

- Out of several sampling distributions, the F-distribution describes the ratio of two independent estimates of a common variance.
- The parameters of the distribution are the degrees of freedom of the numerator and denominator variances, respectively.
- If the null hypothesis of equal mean is true, then we can compute the two estimates of σ^2 namely

$$\hat{\sigma}_B^2 = \sum (\bar{y}_i - \bar{y}_g)^2 / (k - 1) \quad \text{and } s_p^2, \text{ the pooled variance}$$

- Therefore, the ratio $\frac{n\hat{\sigma}_B^2}{s_p^2}$ has the F-distribution with degrees of freedom $(k - 1)$ and $n - k$

Heuristic Justification of ANOVA

- Thus, the procedure for testing the hypothesis.

$$H_0: \mu_i = \mu \text{ all } i = 1, 2, \dots, k$$

H_1 : at least one equality is not satisfied

- We are to reject H_0 , if the calculated value of $F = \frac{\widehat{n}\sigma_B^2}{s_p^2}$ exceeds α (confidence level) of the F-distributions with $(k-1)$ and $n - k$ degrees of freedom.

Example 6: F-Test

| Set 1 | | | Set 2 | | |
|-------------|--------------|--------------|-------------|--------------|--------------|
| Sample 1 | Sample 2 | Sample 3 | Sample 1 | Sample 2 | Sample 3 |
| 5.7 | 9.4 | 14.2 | 3.0 | 5.0 | 11.0 |
| 5.9 | 9.8 | 14.4 | 4.0 | 7.0 | 13.0 |
| 6.0 | 10.0 | 15.0 | 6.0 | 10.0 | 16.0 |
| 6.1 | 10.2 | 15.6 | 8.0 | 13.0 | 17.0 |
| 6.3 | 10.6 | 15.8 | 9.0 | 15.0 | 18.0 |
| $y^- = 6.0$ | $y^- = 10.0$ | $y^- = 15.0$ | $y^- = 6.0$ | $y^- = 10.0$ | $y^- = 15.0$ |

- For both sets, the value of $n\hat{\sigma}_B^2$ is 101.67. However, for Set 1, $s_p^2 = 0.250$ while for Set 2, $s_p^2 = 10.67$. Thus, for Set 1, $F = 406.67$ and for Set 2, $F = 9.53$.

Example 7: Variance between Samples

- The table below shows the lifetimes under controlled conditions, in hours in excess of 1000 hours, of samples of 60W electric light bulbs of three different brands.

| Brand | | |
|-------|----|----|
| 1 | 2 | 3 |
| 16 | 18 | 26 |
| 15 | 22 | 31 |
| 13 | 20 | 24 |
| 21 | 16 | 30 |
| 15 | 24 | 24 |

- Assuming all lifetimes to be normally distributed with common variance, test, at the 1% significance level, the hypothesis that there is no difference between the three brands with respect to mean lifetime.

Solution : Variance between Samples

- The variability between samples may be estimated from the three sample means as follows.

| | Brand | | |
|----------------|-------|------|----|
| | 1 | 2 | 3 |
| Sample Mean | 16 | 20 | 27 |
| Sum | | 63 | |
| Sum of squares | | 1385 | |
| Mean | | 21 | |
| Variance | | 31 | |

- This variance (divisor $(n - 1)$), denoted by $\hat{\sigma}_B^2$ is called the **variance between sample means**. Since it calculated using sample means, it is an estimate of

$$\frac{\sigma^2}{5} \text{ (that is } \frac{\sigma^2}{n} \text{ in general)}$$

based upon $(3 - 1) = 2$ degrees of freedom, but only if the null hypothesis is true. If H_0 is false, then the subsequent 'large' differences between the sample means will result in $5\hat{\sigma}_B^2$ being an inflated estimate of σ^2 .

Solution : F-Test

- The two estimates of σ^2 , $\widehat{n\sigma}_B^2$ and $\widehat{\sigma}_W^2$, may be tested for equality using the F -test with

$$F = \frac{5\widehat{\sigma}_B^2}{\widehat{\sigma}_W^2}$$

as lifetimes may be assumed to be normally distributed.

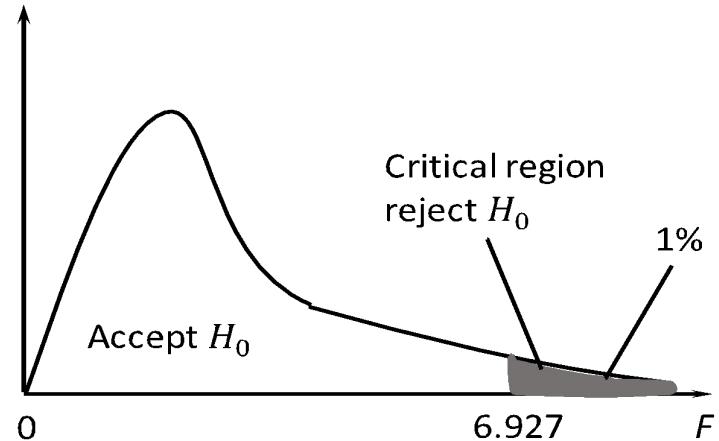
- Recall that the F -test requires the two variances to be independently distributed (from independent samples). Although this is by no means obvious here (both were calculated from the same data), $\widehat{\sigma}_W^2$ and $\widehat{\sigma}_B^2$ are in fact independently distributed.
- The test is always one-sided, upper-tail, since if H_0 is false, $\widehat{\sigma}_W^2$ is inflated whereas $5\widehat{\sigma}_B^2$ is unaffected.
- **Thus in analysis of variance, the convention of placing the larger sample variance in the numerator of the F-statistic is NOT applied.**

Solution

- The solution is thus summarized and completed as follows.

- $H_0: \mu_i = \mu \text{ all } i = 1, 2, 3$
- $H_1: \mu_i \neq \mu \text{ some } i = 1, 2, 3$
- Significance level, $\alpha = 0.01$
- Degrees of freedom, $v_1 = 2, v_2 = 12$
- Critical region is $F > 6.927$
- Test statistic is $F = \frac{5\hat{\sigma}_B^2}{\hat{\sigma}_W^2} = \frac{155}{10} = 15.5$

- This value does lie in the critical region. There is evidence, at the 1% significance level, that **the true mean lifetimes of the three brands of bulb do differ**.



Notation and computational formulae

- In essence, given a population a single factor of k levels, we have to calculate two estimations for σ^2 .
- Sampling variance between groups with $(k-1)$ degree of freedom

$$n\hat{\sigma}_B^2 = n \sum (\bar{y}_{i\cdot} - \bar{y}_g)^2 / (k - 1).$$

- Sampling variance within groups with $(n-k)$ degree of freedom

$$\hat{\sigma}_W^2 = \frac{\sum_{i=1}^k SS_i}{\sum n_i - k}$$

Notation and computational formulae

- The calculations undertaken in the previous example are somewhat cumbersome, and are prone to inaccuracy with non-integer sample means. They also require considerable changes when the sample sizes are unequal. Equivalent computational formulae are available which cater for both equal and unequal sample sizes.
- First, some notation.

Number of samples (or levels)

Total number of observations

Notation and computational formulae

- The computational formulae now follow.

Total sum of squares,

Between samples sum of squares,

Within samples sum of squares,

- A mean square (or unbiased variance estimate) is given by

(sum of squares) \div (degrees of freedom)

$$\text{e.g.} \quad \hat{\sigma}^2 = \frac{(x - \bar{x})^2}{n-1}$$

Hence

Total mean square,

Between samples mean square,

Within samples mean square,

- Note that for the degrees of freedom: $(k - 1) + (n - k) = (n - 1)$**

Example 8: F-Test using Formula

- For the previous example on 60W electric light bulbs, use these computational formulae to show the following.

(a) $SS_T = 430$

(b) $SS_B = 310$

(c) $MS_B = 155 (5\hat{\sigma}_B^2)$

(d) $MS_W = 10 (\hat{\sigma}_W^2)$

Note that $F = \frac{MS_B}{MS_W} = \frac{155}{10} = 15.5$ as previously.

One-way ANOVA Table

- It is convenient to summarize the results of an analysis of variance in a table. For a one factor analysis this takes the following form.

| Source of variation | Sum of squares | Degrees of freedom | Mean square | F ratio |
|---------------------|----------------|--------------------|-------------|---------|
| Between samples | | | | |
| Within samples | | | | |
| Total | | | | |

Example 9: F-Test for unbalanced

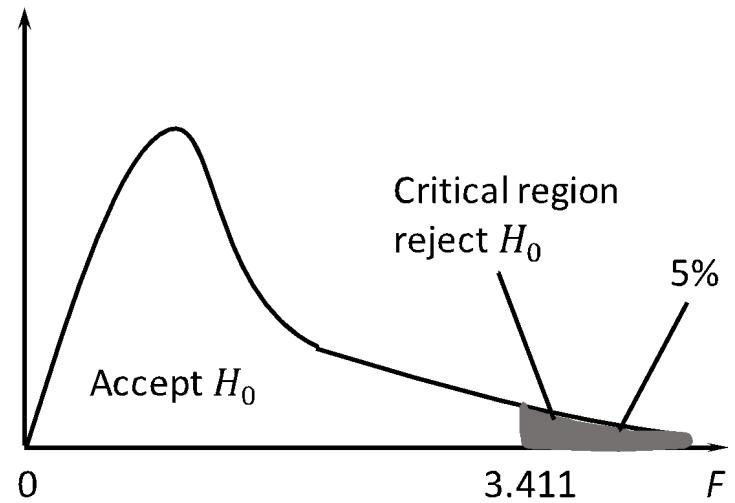
- In a comparison of the cleaning action of four detergents, 20 pieces of white cloth were first soiled with India ink. The cloths were then washed under controlled conditions with 5 pieces washed by each of the detergents. Unfortunately three pieces of cloth were 'lost' in the course of the experiment. Whiteness readings, made on the 17 remaining pieces of cloth, are shown below.

| Detergent | | | |
|-----------|----|----|----|
| A | B | C | D |
| 77 | 74 | 73 | 76 |
| 81 | 66 | 78 | 85 |
| 61 | 58 | 57 | 77 |
| 76 | | 69 | 64 |
| 69 | | 63 | |

- Assuming all whiteness readings to be normally distributed with common variance, test the hypothesis of no difference between the four brands as regards mean whiteness readings after washing.

Solution

- - $H_0: \mu_i = \mu \text{ all } i = 1, 2, 3$
 - $H_1: \mu_i \neq \mu \text{ some } i = 1, 2, 3$
 - Significance level, $\alpha = 0.05$ (say)
 - Degrees of freedom, $v_1 = k - 1 = 3$,
and $v_2 = n - k = 17 - 4 = 13$
 - Critical region is $F > 3.411$



A-16 Appendix Tables

Table A.9 Critical Values for F Distributions (cont.)

| | | $\nu_1 = \text{numerator df}$ | | | | | | | | | |
|----|------|-------------------------------|-------|-------|------|------|------|------|------|------|------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| | | .100 | 3.14 | 2.76 | 2.56 | 2.43 | 2.35 | 2.28 | 2.23 | 2.20 | 2.16 |
| 13 | .050 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | |
| | .010 | 9.07 | 6.70 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | |
| | .001 | 17.82 | 12.31 | 10.21 | 9.07 | 8.35 | 7.86 | 7.49 | 7.21 | 6.98 | |
| | | | | | | | | | | | |
| 14 | .100 | 3.10 | 2.73 | 2.52 | 2.39 | 2.31 | 2.24 | 2.19 | 2.15 | 2.12 | |
| | .050 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | |
| | .010 | 8.86 | 6.51 | 5.56 | 5.04 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 | |
| | .001 | 17.14 | 11.78 | 9.73 | 8.62 | 7.92 | 7.44 | 7.08 | 6.80 | 6.58 | |
| 15 | .100 | 3.07 | 2.70 | 2.49 | 2.36 | 2.27 | 2.21 | 2.16 | 2.12 | 2.09 | |
| | .050 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | |
| | .010 | 8.68 | 6.36 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | |
| | .001 | 16.59 | 11.34 | 9.34 | 8.25 | 7.57 | 7.09 | 6.74 | 6.47 | 6.26 | |
| 16 | .100 | 3.05 | 2.67 | 2.46 | 2.33 | 2.24 | 2.18 | 2.13 | 2.09 | 2.06 | |
| | .050 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | |
| | .010 | 8.53 | 6.23 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | |
| | .001 | 16.12 | 10.97 | 9.01 | 7.94 | 7.27 | 6.80 | 6.46 | 6.19 | 5.98 | |
| 17 | .100 | 3.03 | 2.64 | 2.44 | 2.31 | 2.22 | 2.15 | 2.10 | 2.06 | 2.03 | |
| | .050 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | |
| | .010 | 8.40 | 6.11 | 5.19 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | |
| | .001 | 15.72 | 10.66 | 8.73 | 7.68 | 7.02 | 6.56 | 6.22 | 5.96 | 5.75 | |
| 18 | .100 | 3.01 | 2.62 | 2.42 | 2.29 | 2.20 | 2.13 | 2.08 | 2.04 | 2.00 | |
| | .050 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | |
| | .010 | 8.29 | 6.01 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | |
| | .001 | 15.38 | 10.39 | 8.49 | 7.46 | 6.81 | 6.35 | 6.02 | 5.76 | 5.56 | |
| 19 | .100 | 2.99 | 2.61 | 2.40 | 2.27 | 2.18 | 2.11 | 2.06 | 2.02 | 1.98 | |
| | .050 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | |
| | .010 | 8.18 | 5.93 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | |
| | .001 | 15.08 | 10.16 | 8.28 | 7.27 | 6.62 | 6.18 | 5.85 | 5.59 | 5.39 | |
| | | .000 | 2.87 | 2.59 | 2.28 | 2.05 | 2.16 | 2.00 | 2.04 | 2.00 | 1.96 |

Solution

| | A | B | C | D | Total |
|--|-----|-----|-----|-----|-------|
| | 5 | 3 | 5 | 4 | |
| | 364 | 198 | 340 | 302 | |

$$\sum_i \sum_j y_{ij}^2 = 86362$$

$$SS_T = 86362 - \frac{1204^2}{17} = 1090.47$$

$$SS_B = \left(\frac{364^2}{5} + \frac{198^2}{3} + \frac{340^2}{5} + \frac{302^2}{4} \right) - \frac{1204^2}{17} = 216.67$$

$$SS_W = 1090.47 - 216.67 = 873.80$$

Solution

- The ANOVA table is now as follows.

| Source of variation | Sum of squares | Degrees of freedom | Mean square | F ratio |
|---------------------|----------------|--------------------|-------------|---------|
| Between detergents | | | | |
| Within detergents | | | | |
| Total | | | | |

- The F ratio of 1.07 does not lie in the critical region.
- Thus there is no evidence, at the 5% significance level, to suggest a difference between the four brands as regards mean whiteness after washing. IIITS: IDA - M2021

Two-way ANOVA

Two-way (factor) anova

- This is an extension of the one factor situation to take account of a second factor.
- The levels of this second factor are often determined by groupings of subjects or units used in the investigation. As such it is often called a blocking factor because it places subjects or units into homogeneous groups called blocks. The design itself is then called a randomised block design.

Example 10: Two-factor Analysis

- A computer manufacturer wishes to compare the speed of four of the firm's compilers. The manufacturer can use one of two experimental designs.
 - a) Use 20 similar programs, randomly allocating 5 programs to each compiler.
 - b) Use 4 copies of any 5 programs, allocating 1 copy of each program to each compiler.
- Which of (a) and (b) would you recommend, and why?

Solution

- In (a), although the 20 programs are similar, any differences between them may affect the compilation times and hence perhaps any conclusions. Thus in the 'worst scenario', the 5 programs allocated to what is really the fastest compiler could be the 5 requiring the longest compilation times, resulting in the compiler appearing to be the slowest! If used, the results would require a one factor analysis of variance; the factor being compiler at 4 levels.
- In (b), since all 5 programs are run on each compiler, differences between programs should not affect the results. Indeed it may be advantageous to use 5 programs that differ markedly so that comparisons of compilation times are more general. For this design, there are two factors; compiler (4 levels) and program (5 levels). The factor of principal interest is compiler whereas the other factor, program, may be considered as a blocking factor as it creates 5 blocks each containing 4 copies of the same program.
- **Thus (b) is the better designed investigation.**

SOLUTION

- The actual compilation times, in milliseconds, for this two factor (randomised block) design are shown in the following table.

| | Compiler | | | |
|-----------|----------|-------|-------|-------|
| | 1 | 2 | 3 | 4 |
| Program A | 29.21 | 28.25 | 28.20 | 28.62 |
| Program B | 26.18 | 26.02 | 26.22 | 25.56 |
| Program C | 30.91 | 30.18 | 30.52 | 30.09 |
| Program D | 25.14 | 25.26 | 25.20 | 25.02 |
| Program E | 26.16 | 25.16 | 25.26 | 25.46 |

Assumptions and Interaction

- The three assumptions for a two factor analysis of variance when there is only one observed measurement at each combination of levels of the two factors are as follows.
 1. The population at each factor level combination is (approximately) normally distributed.
 2. These normal populations have a common variance, σ^2 .
 3. The effect of one factor is the same at all levels of the other factor.

Hence from assumptions 1 and 2, when one factor is at level i and the other at level j, the population has a distribution which is

$$N(\mu_{ij}, \sigma^2)$$

- Assumption 3 is equivalent to stating that there is no interaction between the two factors.

ASSUMPTIONS AND INTERACTION

- Now interaction exists when the effect of one factor depends upon the level of the other factor. For example consider the effects of the two factors: sugar (levels none and 2 teaspoons), and stirring (levels none and 1 minute), **on the sweetness of a cup of tea.**
- Stirring has no effect on sweetness if sugar is not added but certainly does have an effect if sugar is added. Similarly, adding sugar has little effect on sweetness unless the tea is stirred.
- Hence factors sugar and stirring are said to interact.
- Interaction can only be assessed if more than one measurement is taken at each combination of the factor levels. Since such situations are beyond the scope of this text, it will always be assumed that interaction between the two factors does not exist.

ASSUMPTIONS AND INTERACTION

- Thus, for example, since it would be most unusual to find one compiler particularly suited to one program, the assumption of no interaction between compilers and programs appears reasonable.

Notation and Computational Formulae

- As illustrated earlier, the data for a two-way ANOVA can be displayed in a two-way table. It is thus convenient, in general, to label the factors as **a row factor** and a **column factor**.
- Notation, similar to that for the one factor case, is then as follows.

| | |
|---|------------------------------------|
| Number of levels of row factor | $= r$ |
| Number of levels of column factor | $= c$ |
| Total number of observations | $= rc$ |
| Observation in (i j-th cell of table) | $= x_{ij}$ |
| (ith level of row factor and jth level of column factor) | $i=1,2,\dots,r$ $j=1,2,\dots,c$ |

Notation and computational formulae

Sum of c observations in i-th row

Sum of r observations in j-th column

Sum of all rc observations

- These lead to the following computational formulae which again are similar to those for one-way ANOVA except that there is an additional sum of squares, etc. for the second factor.

Notation and computational formulae

Total sum of squares,

Between rows sum of squares,

Between columns sum of squares,

Error (residual) sum of squares,

What are the degrees of freedom for SST , SSR and SSC when there are 20 observations in a table of 5 rows and 4 columns?
What is the degrees of freedom of SSE ?

ANOVA Table and Hypothesis Test

For a two factor analysis of variance this takes the following form.

| Source of variation | Sum of squares | Degrees of freedom | Mean square | F ratio |
|---------------------|----------------|--------------------|-------------|---------|
| Between rows | | $r - 1$ | | |
| Between columns | | $c - 1$ | | |
| Error (residual) | | $(r-1)(c-1)$ | | |
| Total | | $rc - 1$ | | |

- **Notes :**

1. The three sums of squares, SS_R , SS_C and SS_E are independently distributed.
2. For the degrees of freedom:
$$(r-1) + (c-1) + (r-1)(c-1) = rc - 1$$

ANOVA Table and Hypothesis Test

- Using the F ratios, tests for significant row effects and for significant column effects can be undertaken.

| H0: no effect due to row factor | H0: no effect due to column factor |
|---------------------------------|------------------------------------|
| H1: an effect due to row factor | H1: an effect due to column factor |
| | |
| | |

Example 11: Two-way ANOVA

- Returning to the compilation times, in milliseconds, for each of five programs, run on four compilers.
- Test, at the 1% significance level, the hypothesis that there is no difference between the performance of the four compilers.
- Has the use of programs as a blocking factor proved worthwhile? Explain.
- The data, given earlier, are reproduced below.

| | Compiler | | | |
|-----------|----------|-------|-------|-------|
| | 1 | 2 | 3 | 4 |
| Program A | 29.21 | 28.25 | 28.20 | 28.62 |
| Program B | 26.18 | 26.02 | 26.22 | 25.56 |
| Program C | 30.91 | 30.18 | 30.52 | 30.09 |
| Program D | 25.14 | 25.26 | 25.20 | 25.02 |
| Program E | 26.16 | 25.16 | 25.26 | 25.46 |

Solution : Dataset

- To ease computations, these data have been transformed (coded) by
 $x = 100 \times (\text{time} - 25)$

to give the following table of values and totals.

| | Compiler | | | | |
|-----------|-------------|------------|-------------|------------|-----------------|
| | 1 | 2 | 3 | 4 | |
| Program A | 421 | 325 | 320 | 362 | 1428 |
| Program B | 118 | 102 | 122 | 56 | 398 |
| Program C | 591 | 518 | 552 | 509 | 2170 |
| Program D | 14 | 26 | 20 | 2 | 62 |
| Program E | 116 | 14 | 26 | 46 | 202 |
| | 1260 | 985 | 1040 | 975 | 4260 = T |
| | | | | | |

Solution : Parameters

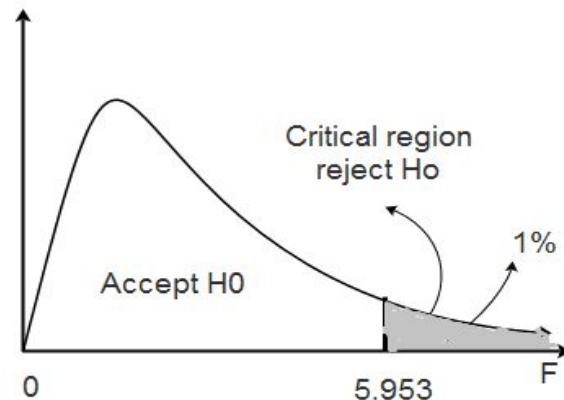
- The sums of squares are now calculated as follows.
(Rows = Programs, Columns = Compilers)
- $SS_T = 1757768 = \frac{4260^2}{20} = 850388$
- $SS_R = \frac{1}{4} (1428^2 + 398^2 + 2170^2 + 62^2 + 202^2) - \frac{4260^2}{20} = 830404$
- $SS_C = \frac{1}{5} (1260^2 + 985^2 + 1040^2 + 975^2) - \frac{4260^2}{20} = 10630$
- $SS_E = 850388 - 830404 - 10630 = 9354$

Solution: ANOVA Table

| Source of variation | Sum of squares | Degrees of freedom | Mean square | F ratio |
|---------------------|----------------|--------------------|-------------|---------|
| Between programs | 830404 | 4 | 207601.0 | 266.33 |
| Between compilers | 10630 | 3 | 3543.3 | 4.55 |
| Error (residual) | 9354 | 12 | 779.5 | |
| Total | 850388 | 19 | | |

Solution : Hypothesis Test

- H_0 : no effect on compilation times due to compilers
- H_1 : an effect on compilation times due to compilers
- Significance level, $\alpha = 0.001$
- Degrees of freedom, $v_1 = c - 1 = 3$
and $v_2 = (r - 1)(c - 1) = 4 \times 3 = 12$
- Critical region is $F > 5.953$
- Test statistic $FC = 4.55$



- This value does not lie in the critical region. Thus there is no evidence, at the 1% significance level, to suggest a difference in compilation times between the four compilers.

REFERENCE

- The detail material related to this lecture can be found in
Design and Analysis of Experiments (8th Edition), Douglas C. Montgomery, John Wiley & Sons, 2013.

Any question?