

# Course: Applied Machine Learning (CSE3087)

## Module1

### Module-I: Supervised Learning Application]

[14 Sessions] [Blooms Taxonomy Selected-

An overview of Machine Learning (ML); ML workflow; types of ML; Types of features, Feature Engineering - Data Imputation Methods; Regression – introduction; simple linear regression, loss functions; Polynomial Regression; Logistic Regression; Softmax Regression with cross entropy as cost function;

**Bayesian Learning** – Bayes Theorem, estimating conditional probabilities for categorical and continuous features, Naïve Bayes for supervised learning; Bayesian Belief networks; Support Vector Machines – soft margin and kernel tricks.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087

# Machine Learning

Machine learning (ML) is a type of artificial intelligence that enables machines to learn and improve from experience without being explicitly programmed.

ML algorithms can be divided into three main categories: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning involves training a model on a labeled dataset, where the input features are labeled with the corresponding output values.

Unsupervised learning involves training a model on an unlabeled dataset, where the goal is to discover patterns and structure in the data.

Reinforcement learning involves training a model to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties.

- The ML workflow typically consists of the following steps: data collection, data cleaning, feature engineering, model selection, model training, model evaluation, and model deployment.
- Data collection involves gathering data from various sources, such as databases, APIs, and web scraping.
- Data cleaning involves preprocessing the data to remove missing values, outliers, and other anomalies that could affect the model's performance.
- Feature engineering involves selecting and transforming the input features to improve the model's performance. This can include techniques such as feature scaling, one-hot encoding, and dimensionality reduction.



# ML Workflow

- Model selection involves choosing the appropriate algorithm for the task at hand. This can involve experimenting with different algorithms and comparing their performance.
- Model training involves fitting the model to the training data using an optimization algorithm such as gradient descent.
- Model evaluation involves testing the model on a held-out validation set to assess its performance.
- Model deployment involves integrating the trained model into a larger system or application.



# Knowledge Representation

- Wine quality is a subjective measure of the overall quality of a wine, based on factors such as its taste, aroma, and appearance.
- There are many different factors that can affect wine quality, including the grape variety, the region where the grapes were grown, the weather conditions during the growing season, and the winemaking process.
- Wine quality is typically rated on a scale of 0 to 100, with higher scores indicating better quality.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087



# Knowledge Representation

## Wine's Five Key Structural Elements



### **SWEETNESS**

The amount of residual sugar.  
Different from fruity!

### **TANNIN**

That drying feeling in your mouth  
that sucks your cheeks in.

### **ACIDITY**

Mouthwatering sensation that  
makes you crave another sip.

### **ALCOHOL**

Cool-climate wines tend to have  
less than warm-climate wines.

### **BODY**

Residual sugar, alcohol and tannin  
determine the weight of a wine.

@wineenthusiast



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013

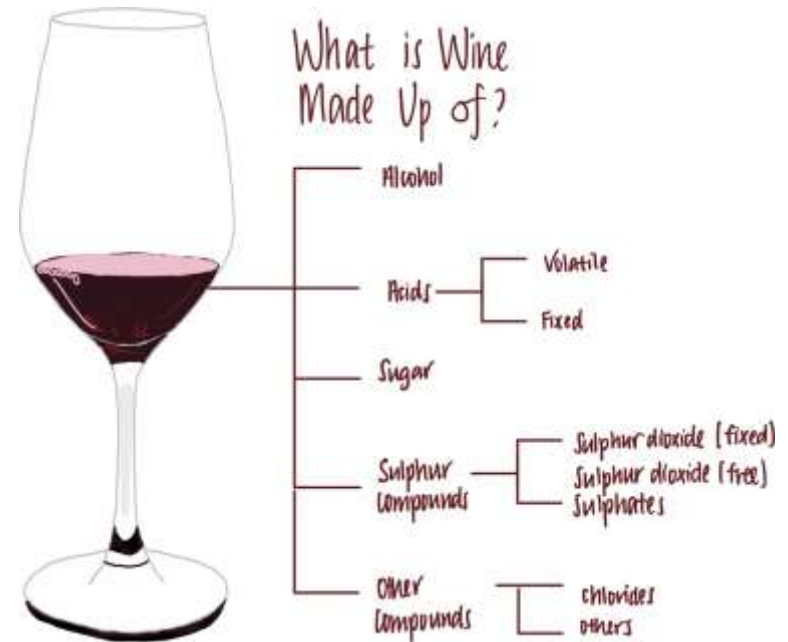


CSE3087

5 /  
51

# Knowledge Representation

- Color: deep red
- Aroma: ripe blackberries and vanilla
- Taste: full-bodied with firm tannins and a long finish
- Ratings: expert- 93 rated by a wine expert and consumer: 85 rated by a consumer





# Types of Features

- Features can be divided into two main categories: categorical and numerical.
- Categorical features represent discrete values, such as colors, types, and categories.
- Numerical features represent continuous values, such as measurements, counts, and percentages.
- Features can also be divided into binary, ordinal, and interval/ratio types, depending on their characteristics and properties.





# Data Transformation



## Natural Language Processing

by National Research University Higher School of Economics



**We can count token pairs, triplets, etc.**

- Also known as n-grams
  - 1-grams for tokens
  - 2-grams for token pairs
  - ...



good movie
not a good movie
did not like



good movie	movie	did not	a	...
1	1	0	0	...
1	1	0	1	...
0	0	1	0	...



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087

## Data Transformation

- Image data transformation: Image data is typically represented as a matrix of pixel values, and a common transformation is to normalize the pixel values so they have a mean of zero and a standard deviation of one.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087

9 /  
51

## Data Transformation

- Text data transformation: A common technique is to represent the text as a bag-of-words, which involves creating a dictionary of all the unique words in the text and representing each document as a vector of word frequencies. The text as a sequence of n-grams (i.e. contiguous sequences of words or characters) and using word embeddings (i.e. dense vectors that represent the meaning of a word) to capture the semantic relationships between words.



## Data Transformation

Numerical data transformation: Numerical data typically requires scaling and normalization to ensure that all the input features have a similar scale and range.

Common techniques include:

- min-max scaling (i.e. scaling the values to be between 0 and 1)

- z-score normalization (i.e. scaling the values to have a mean of 0 and a standard deviation of 1)

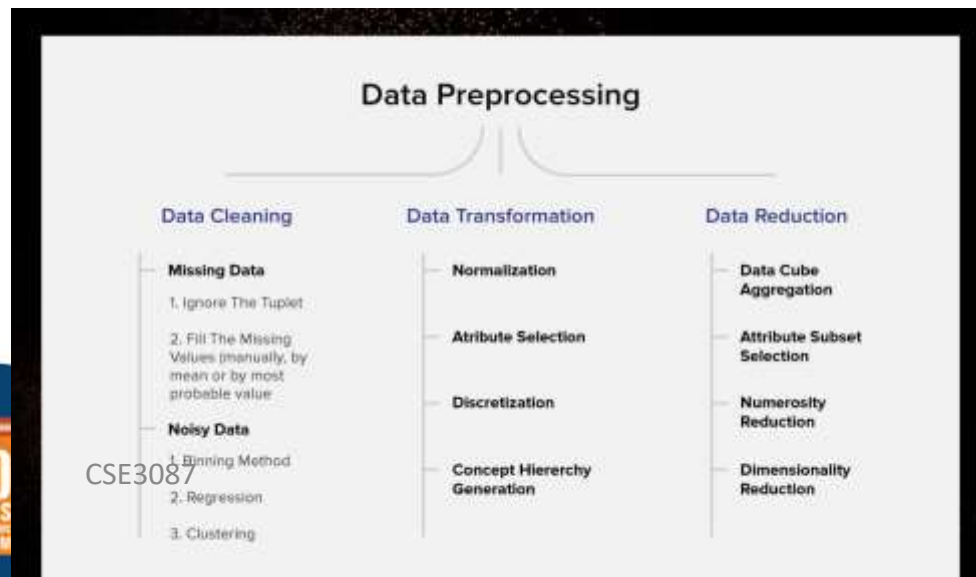
- logarithmic transformation (i.e. transforming the values to their logarithm to reduce skewness in the distribution).

Other techniques include feature selection (i.e. selecting a subset of the most relevant input features) and feature engineering (i.e. creating new features by combining or transforming existing features).



# Feature Engineering

- Feature engineering is the process of selecting and transforming the input features to improve the model's performance.
- Some common feature engineering techniques include feature scaling, one-hot encoding, and dimensionality reduction.
- Feature scaling involves scaling the numerical features to a common scale to prevent bias in the model.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087

12 /  
51

## Methods of Data Imputation

Mean Imputation: Replace missing values with the mean of the non-missing values in the variable.

Regression Imputation: Predict the missing values using a regression model based on the other variables in the dataset.

Multiple Imputation: Create multiple imputed datasets and combine the results for analysis.

K-Nearest Neighbor (KNN) Imputation: Predict the missing values using the values of the nearest neighbors in the dataset.



## Data Imputation

Data Imputation is a statistical technique used to fill in missing data points in a dataset.

There are several methods for data imputation, but one common approach is to use a statistical model to estimate the missing values based on the available data.

### Example

Suppose we have a dataset with five data points with NaN values:

We want to impute the missing value using a simple linear regression model. The linear regression model is given by:

$$y = bx + a$$



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087

14 /

51



- where  $y$  is the dependent variable (the missing value we want to impute),  $x$  is the independent variable (in this case, the index of the data point),  $b$  is the slope of the line, and  $a$  is the  $y$ -intercept.

To estimate the missing value, we first need to fit the linear regression model to the available data. We can do this by using the least squares method, which finds the values of  $a$  and  $b$  that minimize the sum of the squared differences between the predicted values and the actual values of the dependent variable. The formula for the slope  $b$  and  $y$ -intercept  $a$  are given by:



Formula for linear regression equation is given by:

$$y = a + bx$$

$a$  and  $b$  are given by the following formulas:

$$a (\text{intercept}) = \frac{\sum y \sum x^2 - \sum x \sum xy}{(\sum x^2) - (\sum x)^2}$$

$$b (\text{slope}) = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Where,

$x$  and  $y$  are two variables on the regression line.

$b$  = Slope of the line.

$a$  =  $y$ -intercept of the line.

$x$  = Values of the first data set.

$y$  = Values of the second data set.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



**Solution:**

Construct the following table:

x	y	$x^2$	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
$\sum x = 20$	$\sum y = 25$	$\sum x^2 = 120$	$\sum xy = 144$

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{4 \times 144 - 20 \times 25}{4 \times 120 - 400}$$

$$b = 0.95$$

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{25 \times 120 - 20 \times 144}{4(120) - 400}$$

$$a = 1.5$$

Linear regression is given by:

$$y = a + bx$$

$$y = 1.5 + 0.95x$$

find the value of Y for x=12

$$Y = 1.5 + (0.95 * 12)$$

$$Y = 12.5$$



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



## K-Nearest Neighbor (KNN) Imputation

- The KNN algorithm selects the  $k$  most similar observations to the one with the missing value and then takes the average or weighted average of the values of these observations to fill in the missing value.

It is referred to as multivariate because it considers multiple variables or features in the dataset to estimate the missing values. By leveraging the values of other variables, KNN imputation takes into account the relationships and patterns present in the data to impute missing values.



## Calculating K Nearest Neighbors with NaN Euclidean Distance:

When calculating K nearest neighbors using Euclidean distance and dealing with missing values (NaN), special handling is required. Here's how it can be performed:

1. Identify the subset of data points that have non-missing values for the target feature.
2. Calculate the Euclidean distance between the data point with the missing value and each data point in the subset.
3. Exclude data points with missing values (NaN) in the features being compared.
4. Select the K data points with the smallest Euclidean distances as the nearest neighbors.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



```
array([[ 1.         ,  7.25        , 22.         ],
       [ 1.         , 71.2833     , 38.         ],
       [ 0.         ,  7.925      , 26.         ],
       [ 1.         , 53.1        , 35.         ],
       [ 0.         ,  8.05       , 35.         ],
       [ 0.         ,  8.4583     , 28.50639495]])
```

```
# new in 0.22
from sklearn.impute import KNNImputer

impute_knn = KNNImputer(n_neighbors=2)
impute_knn.fit_transform(X)
```

Which features KNN imputer will take into account?

Whatever columns are passed as X

Because of Euclidean distance calculation, KNN imputer is not applicable on categorical variables



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



The formula for Euclidean distance between two data points X and Y is:

Distance =  $\sqrt{\text{weight} * \sum((X_i - Y_i)^2)}$ , for  $i = 1$  to number of dimensions

Where  $X_i$  and  $Y_i$  represent the values of the  $i$ -th feature or dimension of data points X and Y, respectively and  
weight = Total # of coordinates / # of present coordinates  
Euclidean.

An example to illustrate the calculation of Euclidean distance when dealing with NaN values:

Consider two data points, X and Y, with three features: A, B, and C.



**PRESIDENCY  
UNIVERSITY**

Private University Enrolled in Karnataka State by Act No. 41 of 2013



[2, 4, 5]

X: [1, NaN, 3] Y:



To calculate the Euclidean distance between X and Y, we ignore the missing value (NaN) in feature B and compute the distance using the available features:

$$\text{Euclidean Distance} = \sqrt{\text{weight} * (X\_A - Y\_A)^2 + (X\_C - Y\_C)^2}$$

Applying the values from X and Y:

$$\text{Euclidean Distance} = \sqrt{3/2 * (1-2)^2 + (3-5)^2} = \sqrt{1.5 * (1 + 4)}$$

$$\sqrt{1.5 * 5} = 2.74$$

In this case, the missing value in feature B does not contribute to the Euclidean distance calculation between X and Y. The distance is determined based on the available features A and C.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# What is Regression?

Regression is a statistical modeling technique used to explore the relationship between a dependent variable and one or more independent variables. It is used to model and predict the value of the dependent variable based on the values of the independent variables.

- The goal of regression is to find a function that can accurately predict the value of the dependent variable based on the values of the independent variables.



## Simple Linear Regression:

Simple linear regression is a type of regression analysis that models the relationship between a dependent variable and a single independent variable.

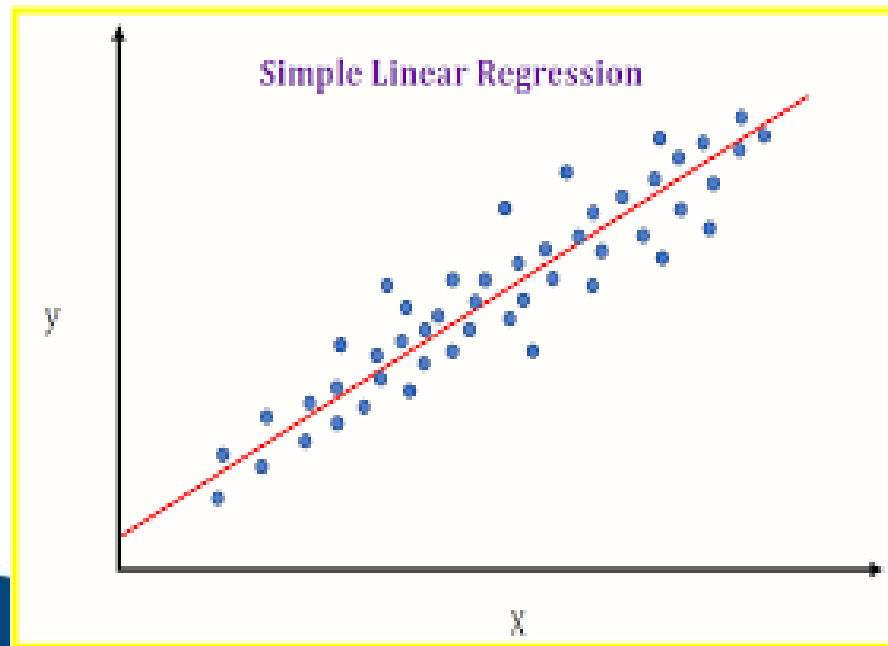
It assumes a linear relationship between the two variables and uses a line to model the relationship.

The line is fitted to the data using a least squares regression approach, which minimizes the sum of the squared differences between the predicted values and the actual values of the dependent variable.



# Simple Linear Regression

Simple linear regression is a type of regression where there is only one independent variable. The goal of simple linear regression is to find the line of best fit that minimizes the difference between the predicted values and the actual values of the dependent variable.



# Loss Functions

In order to find the line of best fit in simple linear regression, we need to define a loss function that measures the difference between the predicted values and the actual values of the dependent variable.

Two commonly used loss functions in simple linear regression are the Mean Squared Error (MSE) and the Mean Absolute Error (MAE).

- $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$



**Loss Functions**

Loss functions are used in regression to quantify the difference between the predicted values and the actual values of the dependent variable.

The goal of regression is to minimize the loss function by adjusting the parameters of the model.

The most commonly used loss function in simple linear regression is the mean squared error (MSE), which is the average of the squared differences between the predicted values and the actual values.

Other loss functions include mean absolute error (MAE) and root mean squared error (RMSE).



# Polynomial Regression

Polynomial regression is a type of regression where the relationship between the dependent variable and the independent variable(s) is modeled as an  $n$ th degree polynomial. This allows for a more complex relationship between the variables than simple linear regression.

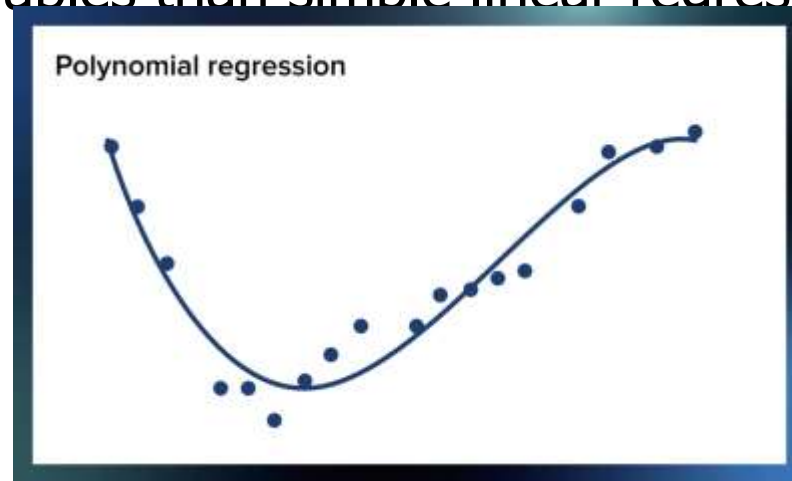


Figure: Polynomial Regression

CSE3087



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013





# Polynomial Regression:

## What is Polynomial Regression?

In polynomial regression, we describe the relationship between the independent variable  $x$  and the dependent variable  $y$  using an  $n$ th-degree polynomial in  $x$ .

Polynomial regression is a type of regression analysis that models the relationship between a dependent variable and one or more independent variables using a polynomial function.

It allows for more complex relationships between the variables by fitting a curve to the data instead of a straight line.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087

28 /

51

## Types of Polynomial Regression

A quadratic equation is a general term for a second-degree polynomial equation. This degree, on the other hand, can go up to  $n$ th values. Here is the categorization of Polynomial Regression:

Linear – if degree as 1

Quadratic – if degree as 2

Cubic – if degree as 3 and goes on, on the basis of degree.

When the Linear Regression Model fails to capture the points in the data and the Linear Regression fails to adequately represent the optimum, then we use Polynomial Regression



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



Polynomials	Form	Degree	Examples
Linear Polynomial	$p(x): ax+b, a \neq 0$	Polynomial with Degree 1	$x + 8$
Quadratic Polynomial	$p(x): ax^2+bx+c, a \neq 0$	Polynomial with Degree 2	$3x^2-4x+7$
Cubic Polynomial	$p(x): ax^3+bx^2+cx, a \neq 0$	Polynomial with Degree 3	$2x^3+3x^2+4x+6$



Simple  
Linear  
Regression

$$y = b_0 + b_1x_1$$

Multiple  
Linear  
Regression

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Polynomial  
Linear  
Regression

$$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$$



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# Introduction to Logistic Regression

- Logistic regression is a popular method for binary classification, where the goal is to predict the probability of an input belonging to a particular class.
- The logistic regression model uses a logistic function to model the probability of an input belonging to the positive class.
- The logistic function maps any input to a value between 0 and 1, which can be interpreted as a probability.

$$p(y = 1|x) = \frac{1}{1 + e^{-z}}$$

Where  $z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$  is the linear combination of the input features  $x_1, x_2, \dots, x_n$  and their corresponding coefficients

$\theta_1, \theta_2, \dots, \theta_n$ .



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



CSE3087

37 /

51

# Softmax Regression

- Softmax regression is a generalization of logistic regression that can be used for multi-class classification problems.
- The softmax function is used to compute the probability of each class, and the class with the highest probability is chosen as the predicted class.
- The softmax function outputs a probability distribution over the classes, and its outputs sum to 1.

$$p(y = i | x) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

where  $K$  is the number of classes,  $z_i$  is the linear combination of input features for class  $i$ , and the denominator sums over all classes.



# Cross Entropy Loss

- To train a logistic regression or softmax regression model, we need a cost function that measures the difference between the predicted probabilities and the true labels.
- The cross entropy loss is a popular choice for such a cost function.





# Solving classification problems with naïve bayes

## How does naïve bayes algorithm work?

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Naive Bayes classifier algorithm is based on a famous theorem called "**Bayes theorem**".
- It can help us find simple yet powerful solutions to many problems ranging from **text analysis** to **spam detection** and much more.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# Probability to describe how likely an event is to happen

A value between 0 and 1 represents the possibility of an event



**Less likely to happen**



**Most likely to happen**



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# Bayes theorem is centered on conditional probability

## What is conditional probability?

**Conditional probability** is the probability of an event 'A' happening given that another event 'B' has already happened.

- ▶ The Bayes theorem is **an extension of conditional probability**. It allows us in a sense to use reverse reasoning.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

CONDITIONAL PROBABILITY



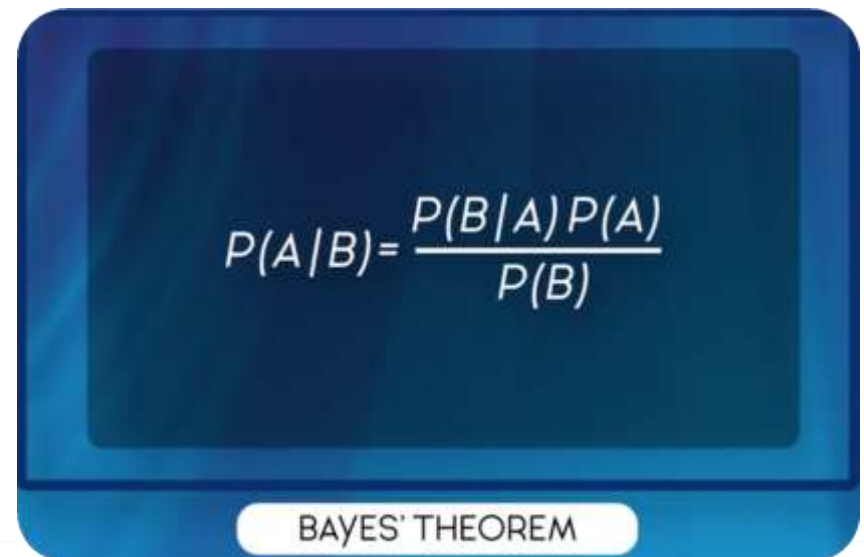
**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# Understanding the Bayes theorem formula

- **Prior probability  $P(A)$**  –  
The probability of just 'A' occurring
- **Posterior probability  $P(A|B)$**  –  
The probability of event 'A' given that event 'B' occurs
- **$P(B|A)$**  - The probability of event B happening given that event A has occurred
- **$P(B)$**  - The probability of just B


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

BAYES' THEOREM



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# What makes Naïve bayes algorithm naïve?

**When the model calculates the conditional probability of one feature given a class,**



...it doesn't take into account the effect of any other feature.



...it assumes that features are independent from each other.



...it gives us the flexibility to describe the probability of each feature.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# The algorithm's naivety has some advantages & limitations

## Advantages



### Quick & simple

- Produce good results with small amount of training data
- Used for benchmarking of a model
- Works well with continuous data by discretizing



### Disadvantages

- In most real-world situations some of the features are likely to be dependent on each other, which might cause wrong results.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



# Three types of naïve bayes classifiers in sklearn

## Bernoulli

Used when data is binary like true or false, yes or no etc.

## Multinomial Naïve Bayes

Used when there are discrete values such as number of family members or pages in a book.

## Gaussian Naïve Bayes

Used when all features are continuous variables, like temperature or height.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



## Reversing the condition

Example: Rahul's favorite breakfast is bagels and his favorite lunch is pizza. The probability of Rahul having bagels for breakfast is 0.6. The probability of him having pizza for lunch is 0.5. The probability of him, having a bagel for breakfast given that he eats a pizza for lunch is 0.7.

Let's define event A as Rahul having a bagel for breakfast, Event B as Rahul having a pizza for lunch.

$$P(A) = 0.6$$

$$P(B) = 0.5$$

If we look at the numbers, the probability of having a bagel is different than the probability of having a bagel given he has a pizza for lunch. This means that the probability of having a bagel is dependent on having a pizza for lunch.

$$P\left(\frac{A}{B}\right) = 0.7$$

Now what if we need to know the probability of having a pizza given you had a bagel for breakfast. i.e. we need to know

$$P\left(\frac{B}{A}\right)$$

. Bayes theorem now comes into the picture.



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013





The Bayes theorem describes the probability of an event based on the prior knowledge of the conditions that  $P\left(\frac{A}{B}\right)$  it be related to the event. If we know the conditional probability  $P\left(\frac{B}{A}\right)$ , we can use the bayes  $P\left(\frac{B}{A}\right)$  to find out the reverse probabilities  $P\left(\frac{A}{B}\right)$ .

For the previous example – if we now wish to calculate the probability of having a pizza for lunch provided you had a bagel for breakfast would be  $= 0.7 * 0.5 / 0.6$ .

We can generalize the formula further.

If multiple events  $A_i$  form an exhaustive set with another event  $B$ .

We can write the equation as

$$P(A_i/B) = \frac{P(B|A_i) * P(A_i)}{\sum_{i=1 \text{ to } n} P(B|A_i) * P(A_i)}$$



Naïve Bayes is a popular algorithm for supervised learning, particularly for text classification problems. It's based on Bayes' theorem, which is a formula for calculating conditional probabilities.

- The basic idea behind Naïve Bayes is to calculate the probability of a particular class given some input features.
- This is done by calculating the conditional probabilities of each feature given the class, and then using Bayes' theorem to calculate the probability of the class given the features.



Bayesian belief networks are a type of graphical model that can be used to represent probabilistic relationships between variables.

- Each variable is represented by a node in the network, and the edges between the nodes represent conditional dependencies.
- The basic idea behind Bayesian belief networks is to use Bayes' theorem to calculate the probabilities of each variable given the values of its parents in the network.
- This can be done by factorizing the joint probability distribution of all the variables into a product of conditional probabilities.



One common application of Bayesian belief networks is in medical diagnosis.

- The variables in the network might represent symptoms and diseases, and the edges represent the conditional dependencies between them.
- Given a set of observed symptoms, the network can be used to calculate the probability of each possible disease.



Naïve Bayes and Bayesian belief networks are powerful tools for probabilistic modeling and inference.

- Naïve Bayes is particularly useful for text classification problems, while Bayesian belief networks are useful for modeling complex dependencies between variables.
- With these tools, we can make more accurate predictions and better understand the relationships between different variables in our data.



- Support Vector Machines (SVMs) are a popular machine learning algorithm for classification and regression. They work by finding the hyperplane that maximally separates the data points in a high-dimensional feature space.
- - The basic idea behind SVMs is to find the hyperplane that maximizes the margin between the positive and negative data points. The margin is the distance between the hyperplane and the closest data points from either class. The hyperplane that maximizes the margin is called the maximum margin hyperplane.
  - - SVMs can be extended to handle non-linearly separable data using a technique called the **kernel trick**. The kernel trick involves mapping the original feature space to a higher-dimensional space using a non-linear function, and then finding the maximum margin hyperplane in this new space.



In practice, data is often not perfectly separable, and finding the maximum margin hyperplane is not always possible. In these cases, we can use a variant of SVM called the **soft margin SVM**.

- The soft margin SVM allows for some misclassifications, and introduces a slack variable that penalizes data points that are on the wrong side of the margin.
- The objective function for the soft margin SVM includes a regularization term that controls the tradeoff between maximizing the margin and minimizing the classification error.



The kernel trick is a powerful technique for extending SVMs to handle non-linearly separable data. The basic idea is to map the data into a higher-dimensional feature space using a non-linear function, and then apply the linear SVM algorithm to this new feature space.

- • The key insight behind the kernel trick is that we can compute the dot product between the mapped data points without explicitly computing the mapping. This is done by defining a kernel function that takes two data points as input and returns the dot product of their mapped features.
- • Some common kernel functions include the polynomial kernel, which computes the dot product of two vectors raised to a certain power, and the radial basis function kernel, which measures the similarity between two data points based on their distance in the feature space.





Support Vector Machines are a powerful machine learning algorithm that can handle both linear and non-linearly separable data.

- The kernel trick allows us to extend SVMs to handle complex data, and the soft margin SVM allows us to handle data that is not perfectly separable.
- With these techniques, we can build powerful models for classification and regression tasks.



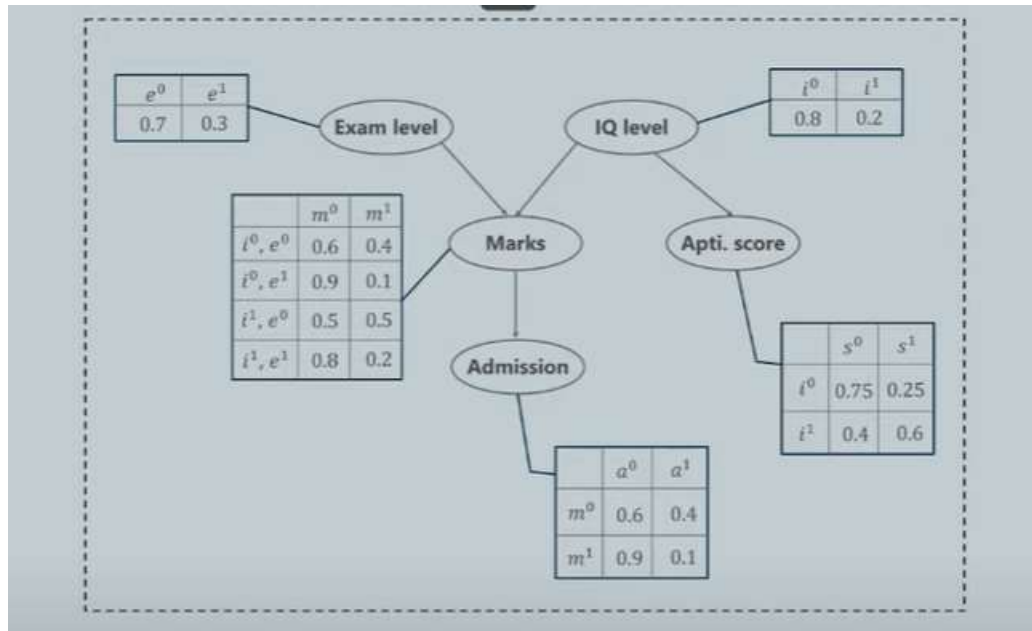
Thank  
you!



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013

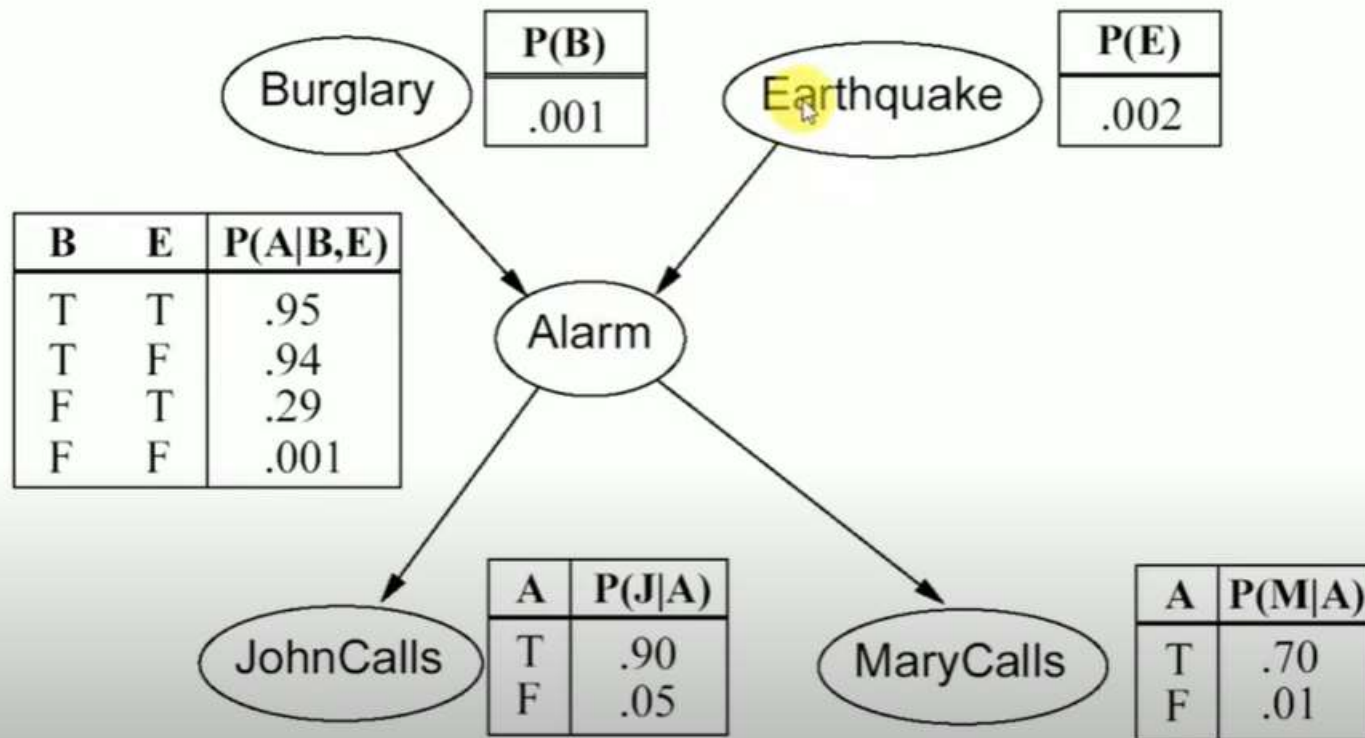




**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013





What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Merry call?

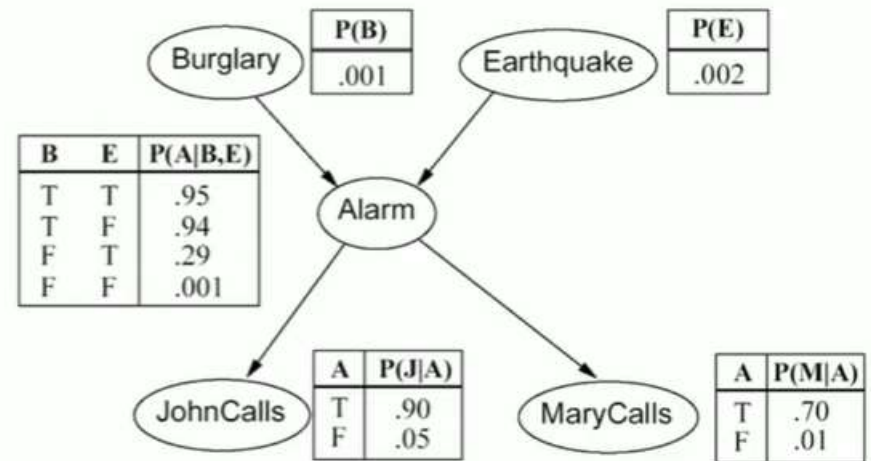


**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



- What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Merry call?



**Solution:**

$$\begin{aligned}
 P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) &= P(j \mid a) P(m \mid a) P(a \mid \neg b, \neg e) P(\neg b) P(\neg e) \\
 &= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 \\
 &= 0.00062
 \end{aligned}$$

This is the Joint Probability Distribution



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013



2. What is the probability that John call?

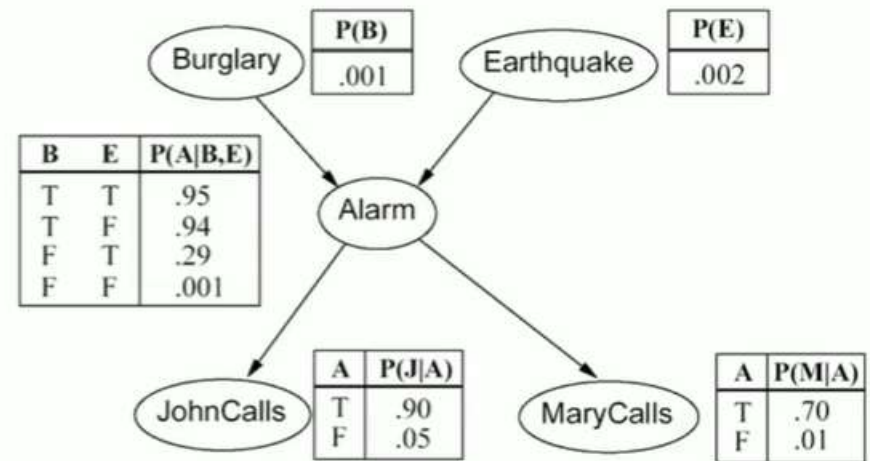
Solution:

$$P(j) = P(j | a) P(a) + P(j | \neg a) P(\neg a)$$

$$= P(j | a) \{P(a | b, e) P(b, e) + P(a | \neg b, e) P(\neg b, e) + P(a | b, \neg e) P(b, \neg e) + P(a | \neg b, \neg e) P(\neg b, \neg e)\}$$

$$+ P(j | \neg a) \{P(\neg a | b, e) P(b, e) + P(\neg a | \neg b, e) P(\neg b, e) + P(\neg a | b, \neg e) P(b, \neg e) + P(\neg a | \neg b, \neg e) P(\neg b, \neg e)\}$$

$$= 0.90 * 0.00252 + 0.05 * 0.9974 = 0.0521$$



**PRESIDENCY  
UNIVERSITY**

Private University Estd. in Karnataka State by Act No. 41 of 2013

