

Introduction to Data Science

Dr. Irfan Yousuf

Department of Computer Science (New Campus)

UET, Lahore

(Lecture # 20; October 21, 2022)

Outline

- Train-Test Split
- Confusion Matrix

Machine Learning Algorithms

Machine Learning

Supervised learning: Train a model with known input and output data to predict future outputs to new data.

Classification

Support vector machine (SVM)

K-nearest-neighbors

Discriminant analysis

Neural Networks

Naive Bayes

Regression

Linear Regression

Assembly Methods

Decision trees

Neural Networks

Unsupervised Learning: Segment a collection of elements with the same attributes (clustering).

Clustering

K-means, k-medoids fuzzy C-means

Hidden Markov models

Neural Networks

Gaussian mixture

Supervised Machine Learning

Supervised Learning

x_1	x_2	x_3	x_p	y

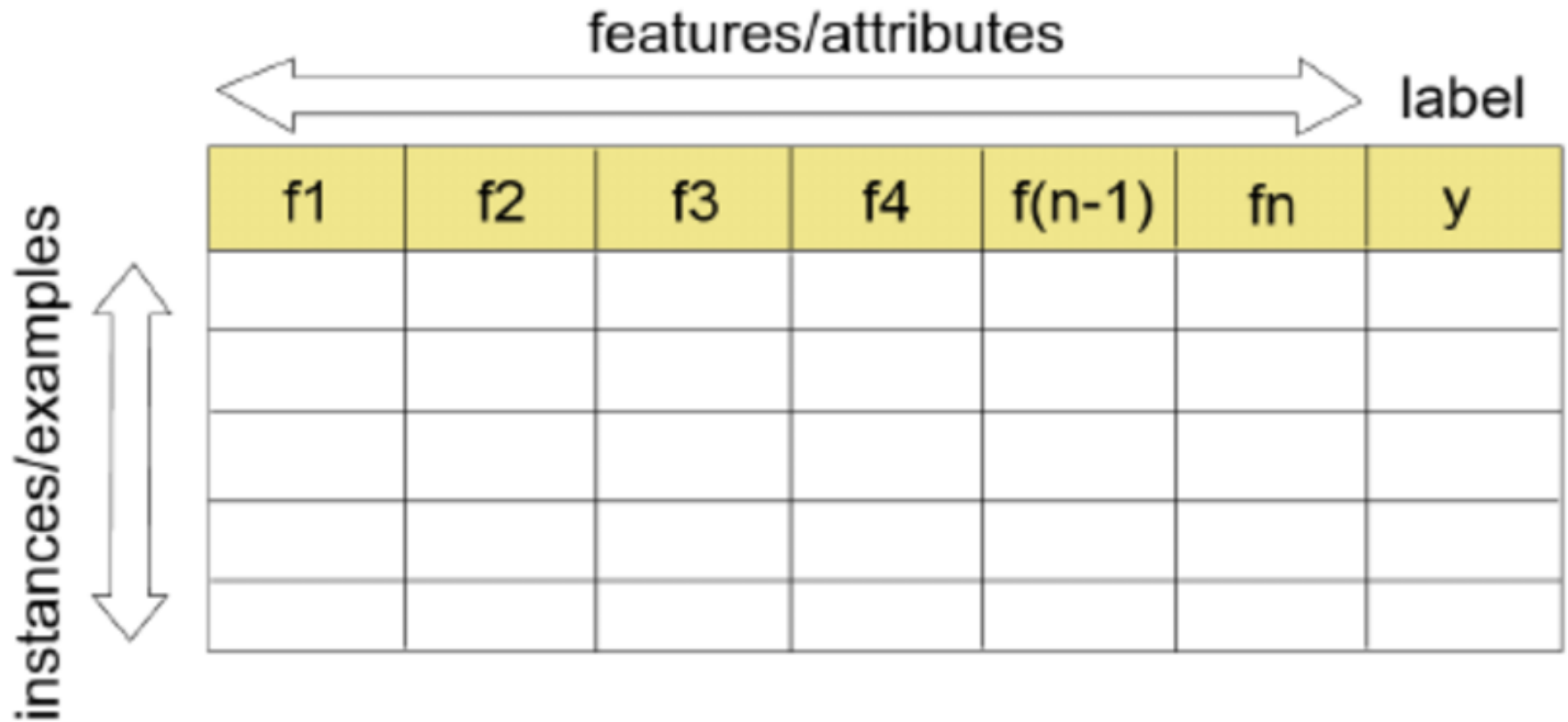
Target

Un-Supervised Learning

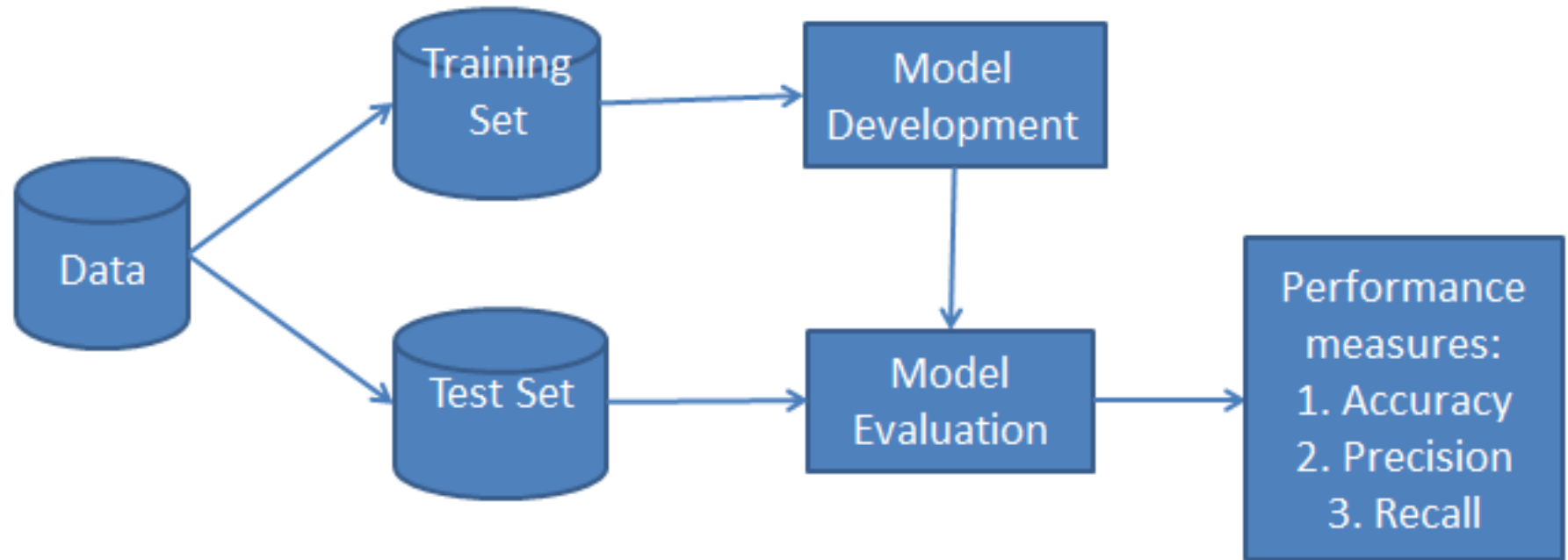
x_1	x_2	x_3	x_p	y

No
Target

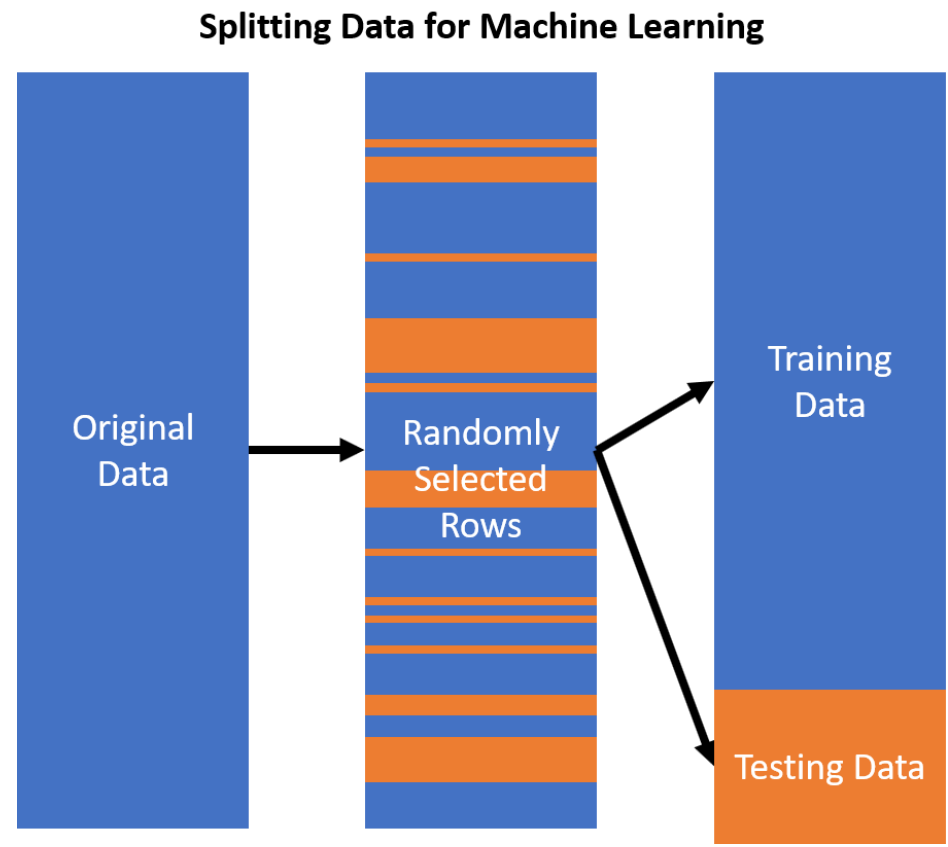
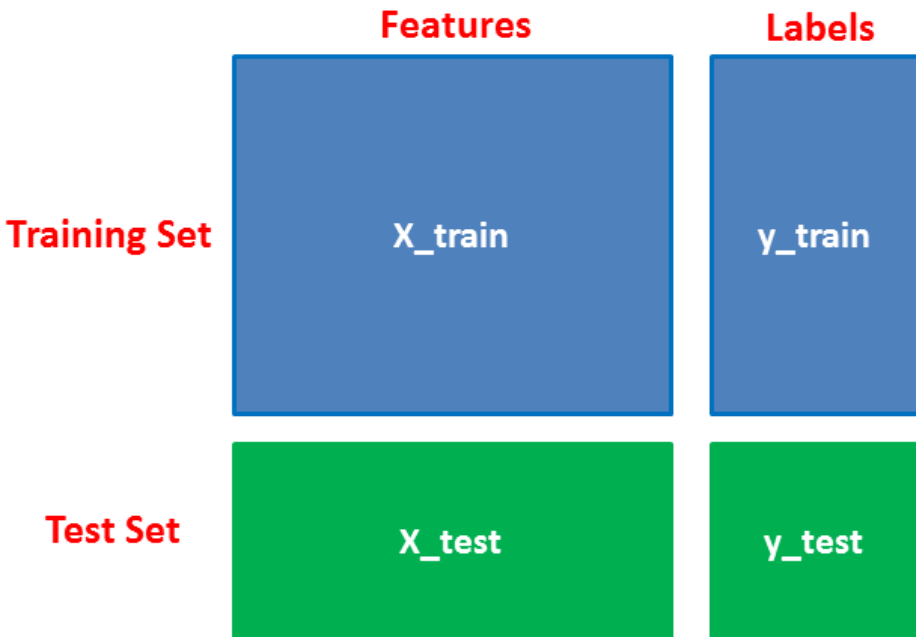
Supervised Machine Learning



Train-Test-Split



Train-Test-Split



TRAIN_TEST_SPLIT SPLITS DATA INTO TRAINING DATA AND TEST DATA

Original Data

X ₁	X ₂	X _p	Y

`train_test_split()`



X_{train}

X ₁	X ₂	X _p

y_{train}

Y

X_{test}

X ₁	X ₂	X _p

y_{test}

Y

Classification Model

- Classification is a technique where we categorize data into a given number of classes.
- The main goal of a classification problem is to identify the category/class to which a new data will fall under.

Classification Model

- **Classifier:** An algorithm that maps the input data to a specific category.
- **Classification model:** A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.
- **Feature:** A feature is an individual measurable property of a phenomenon being observed.
- **Binary Classification:** Classification task with two possible outcomes. Eg: Gender classification (Male / Female)
- **Multi-class classification:** Classification with more than two classes. In multi class classification each sample is assigned to one and only one target label. Eg: An animal can be cat or dog but not both at the same time
- **Multi-label classification:** Classification task where each sample is mapped to a set of target labels (more than one class). Eg: A news article can be about sports, a person, and location at the same time.

Classification Model

The following are the steps involved in building a classification model:

- **Initialize** the classifier to be used.
- **Train the classifier:** All classifiers in scikit-learn uses a `fit(X, y)` method to fit the model(training) for the given train data X and train label y.
- **Predict the target:** Given an unlabeled observation X, the `predict(X)` returns the predicted label y.
- **Evaluate** the classifier model

Evaluating a Classification Model

```
> source('E:/Spring2021/RProgs/SpamFilter.R')
Loading required package: RColorBrewer
Loading required package: NLP
ham --> ham
spam --> spam
ham --> ham
ham --> spam
ham --> ham
ham --> ham
ham --> ham
ham --> spam
ham --> ham
ham --> ham
[1] "Done"
```

Actual

Predicted

Confusion Matrix

- A confusion matrix is a table that is often used to describe the **performance of a classification model** (or "classifier") on a set of test data for which the true values are known.

Confusion Matrix

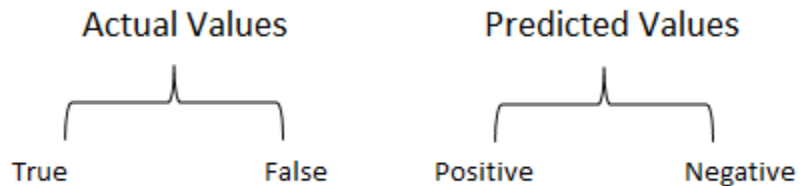
		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

True Positive (TP): You predicted positive and it's true.

True Negative (TN): You predicted negative and it's true.

False Positive (FP): You predicted positive and it's false.

False Negative (FN): You predicted negative and it's false.



Predicted Labels

Actual Labels

Person has Coronavirus

Yes

No

Positive

True Positive (TP):
Person with coronavirus
tested positive

False Positive (FP):
Person without
coronavirus tested
positive

Test Results

Negative

False Negative (FN):
Person with coronavirus
tested negative

True Negative (TN):
Person without
coronavirus tested
negative

Number of **Positive (P)**
predictions that are correct
or **True (T)**

		Actual	
		Spam (+ve)	Not Spam (-ve)
Predictions	Spam (+ve)	TP	FP
	Not Spam (-ve)	FN	TN

Number of **Positive (P)**
predictions that are wrong
or **False (F)**

Number of **Negative (N)**
predictions that are wrong
or **False (F)**

Number of **Negative (N)**
predictions that are correct
or **True (T)**

Sick people correctly
predicted as sick by the
model

Healthy people
incorrectly predicted as
sick by the model

ACTUAL VALUES

PREDICTED VALUES

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP (30)	FP (30)
	NEGATIVE	FN (10)	TN (930)

Sick people incorrectly
predicted as not sick by
the model

Healthy people correctly
predicted as not sick by
the model

Confusion Matrix

		Actual		
		1	0	
Predicted	1	TP	FP	Type 1 Error
	0	FN	TN	

Type 2 Error

The diagram illustrates a confusion matrix with four cells: TP (yellow), FP (orange), FN (green), and TN (blue). The columns represent 'Actual' values (1 and 0) and the rows represent 'Predicted' values (1 and 0). An arrow points from the 'Type 1 Error' label to the FP cell, and another arrow points from the 'Type 2 Error' label to the FN cell.

Confusion Matrix Terminology

- Classification **Accuracy** is the ratio of correct predictions to total predictions made.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Confusion Matrix Terminology

- **Precision** is calculated as the number of correct positive predictions divided by the total number of positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** is calculated as the number of correct positive predictions divided by the total number of positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Confusion Matrix Terminology

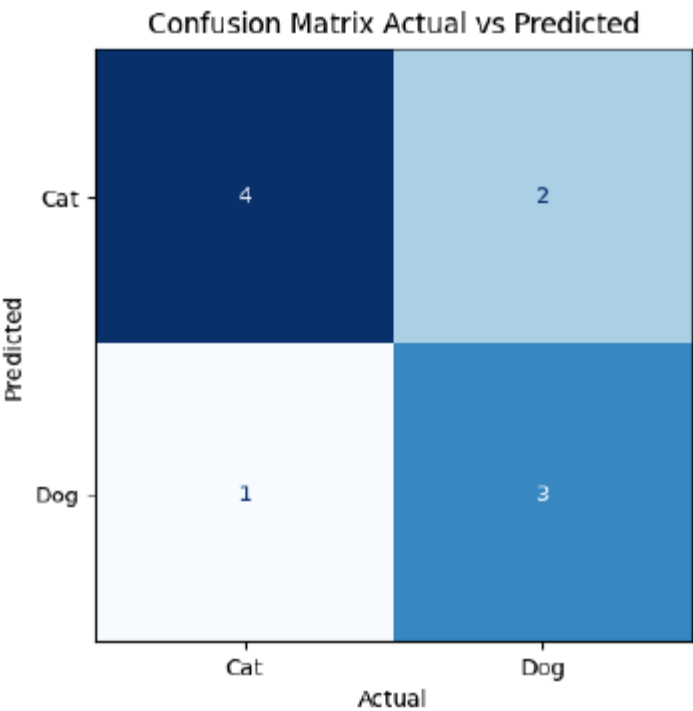
- **F1-score** is the harmonic mean of precision and recall and is a better measure than accuracy.

$$\mathbf{F1\text{-}score} = \frac{2 * \textit{Recall} * \textit{Precision}}{\textit{Recall} + \textit{Precision}}$$

Confusion Matrix Example

a) The output of a machine learning classifier is given below in the form of actual and predicted data. Draw the Confusion Matrix of this classifier and calculate its accuracy.

Actual	Dog	Dog	Cat	Dog	Cat	Cat	Cat	Dog	Dog	Cat
Predicted	Cat	Dog	Cat	Dog	Dog	Cat	Cat	Dog	Cat	Cat



Confusion Matrix Implementation

- Implementation of Confusion Matrix

Summary

- Train Test Split
- Confusion Matrix