Group 5
Team Batak (TB)

Meet the Team

Edward Vincent **“VINCE”** Duero

Likes: Ice cream

Dislikes: Seafoods

Rosiel Jazmine **“ROSE”** Villareal

Likes: Sinigang, Sinampalukan; All kinds of seafood

Dislikes: Fatty part in meat

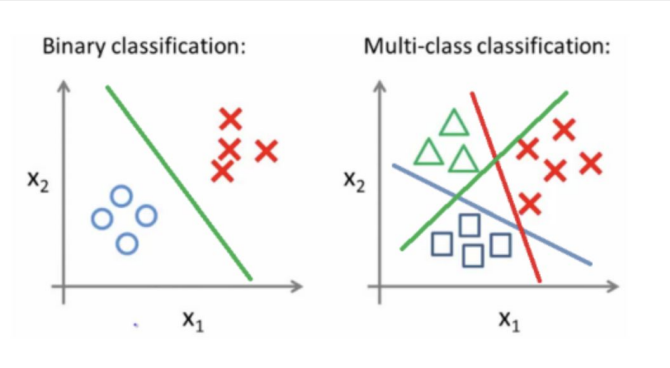
Jericho Carlo **“ECHO”** Agudo

Likes: Sushi, Curry, Silogs

Dislikes: Grape juice, Cakes

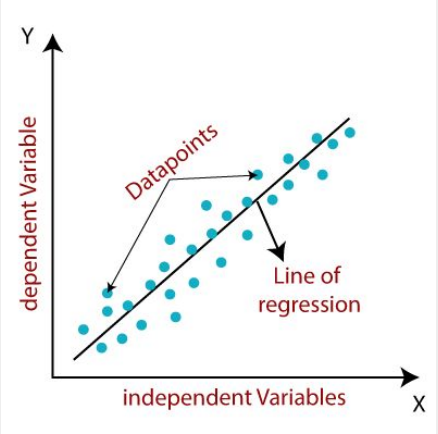
Algorithm 1: Linear Learner

- supervised learning
- classification or regression



APPLICATIONS:

- Binary Classification - Is this email spam or not? Will the hospital patient survive or not?
- Multi-class Classification - Is this animal a dog, horse, or cat? Is the player's position guard, forward, or center?
- Regression - What will the temperature be in Manila tomorrow? What will be the price of this diamond?



Linear Learner: Sample Datasets

dataset.csv (31.41 MiB)

Detail | Compact | Column

patient_id	age	bmi	elective_s...	gender	icu_admit_...	hospital_d...
25312	68	22.73	0	M	Floor	0
59342	77	27.42	0	F	Floor	0
50777	25	31.95	0	F	Accident & Emergency	0
46918	81	22.64	1	F	Operating Room / Recovery	0
34377	19		0	M	Accident & Emergency	0
74489	67	27.56	0	M	Accident & Emergency	0
49526	59	57.45	0	F	Accident & Emergency	0
50129	70		0	M	Accident & Emergency	0
10577	45		0	M	Other Hospital	1
90749	50	25.71	0	M	Accident & Emergency	0
125898	72	28.25705249	1	F	Operating Room / Recovery	0
78266	80	27.3828125	1	F	Operating Room / Recovery	0
41311	48		0	M	Accident & Emergency	0
103766	65		1	M	Operating Room / Recovery	0
98174	81	38.18906706	1	M	Operating Room / Recovery	0
124688	78		0	F	Accident & Emergency	0

Patient Survival Prediction Dataset

diamonds.csv (3.19 MiB)

Detail | Compact | Column

#	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55	326
2	0.21	Premium	E	SI1	59.8	61	326
3	0.23	Good	E	VS1	56.9	65	327
4	0.29	Premium	I	VS2	62.4	58	334
5	0.31	Good	J	SI2	63.3	58	335
6	0.24	Very Good	J	VVS2	62.8	57	336
7	0.24	Very Good	I	VVS1	62.3	57	336
8	0.26	Very Good	H	SI1	61.9	55	337
9	0.22	Fair	E	VS2	65.1	61	337
10	0.23	Very Good	H	VS1	59.4	61	338
11	0.3	Good	J	SI1	64	55	339
12	0.23	Ideal	J	VS1	62.8	56	340
13	0.22	Premium	F	SI1	60.4	61	342
14	0.31	Ideal	J	SI2	62.2	54	344
15	0.2	Premium	E	SI2	60.2	62	345
16	0.32	Premium	E	I1	60.9	58	345
17	0.3	Ideal	I	SI2	62	54	348
18	0.3	Good	J	SI1	63.4	54	351
19	0.3	Good	J	SI1	63.8	56	351
20	0.3	Very Good	J	SI1	62.7	59	351
21	0.3	Good	I	SI2	63.3	56	351
22	0.23	Very Good	E	VS2	63.8	55	352

Diamond Price Prediction Dataset

Linear Learner: Hyperparameters

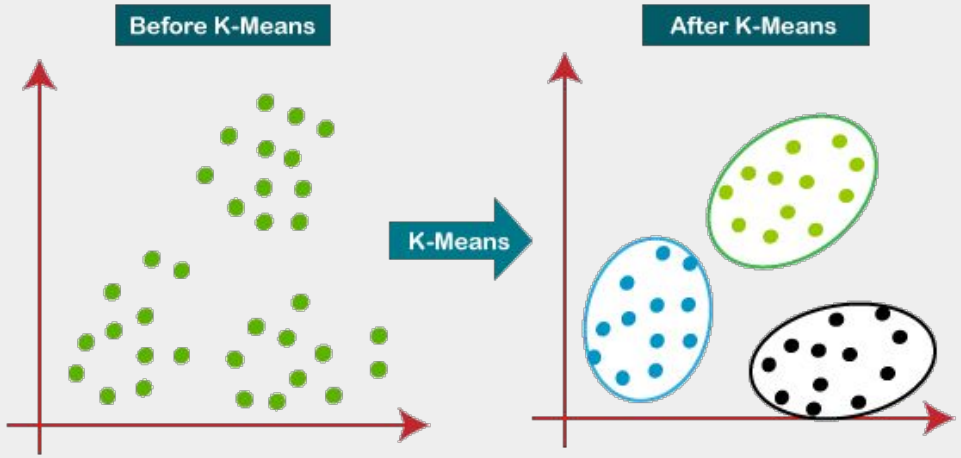
Parameter Name	Description
1. `predictor_type`	• Specifies the type of target variable. • Valid values: binary_classifier, multiclass_classifier, or regressor
2. `num_classes`	• The number of classes for the response variable. • Required when predictor_type is multiclass_classifier. • Valid Values: Integers from 3 to 1,000,000
3. `epochs`	• The maximum number of passes over the training data. • Valid Values: Positive integer • Default Value: 15
4. `feature_dim`	• The number of features in the input data. • Valid Values: auto or positive integer • Default Value: auto

Algorithm 2: K-Means

- unsupervised learning
- grouping similar data

APPLICATIONS:

- Customer segmentation
- Outlier/Anomaly detection
- Document classification
- Inventory categorization



K Means: Sample Datasets

Mall_Customers.csv (3.98 KiB)

Detail | Compact | Column

About this file

This file contains the basic information (ID, age, gender, income, spending score) about the customers

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (...
Unique ID assigned to the customer	Gender of the customer	Age of the customer	Annual Income of the customee	Score assigned by the mall based on customer behavior and spending nature
1 200	Female 56% Male 44%	18 70	15 137	1 99 (63.60 - 75.40, Count: 24)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99

Mall Customer Segmentation Data

emails.csv (1.43 GiB)

Detail | Compact | Column

file	message
allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.evans@thyme> Date: Mon, 14 May 2001 16:39:00 -0700 (PDT...
allen-p/_sent_mail/10.	Message-ID: <15464986.1075855378456.JavaMail.evans@thyme> Date: Fri, 4 May 2001 13:51:00 -0700 (PDT)...
allen-p/_sent_mail/100.	Message-ID: <24216240.1075855687451.JavaMail.evans@thyme> Date: Wed, 18 Oct 2000 03:00:00 -0700 (PDT...
allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.evans@thyme> Date: Mon, 23 Oct 2000 06:13:00 -0700 (PDT...
allen-p/_sent_mail/1001.	Message-ID: <30922949.1075863688243.JavaMail.evans@thyme> Date: Thu, 31 Aug 2000 05:07:00 -0700 (PDT...

Enron Email Dataset

apper.ph

K-Means: Hyperparameters

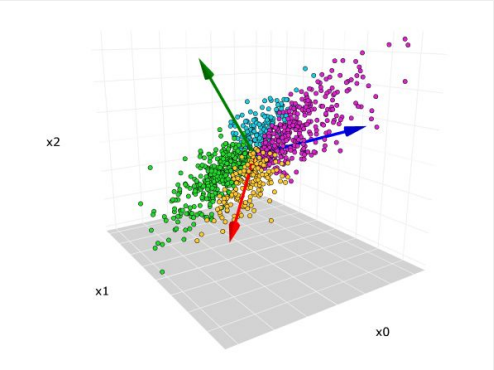
Parameter Name	Description
1. `k`	• The number of required clusters • Valid Values: Positive Integer
2. `extra_center_factor`	• The algorithm creates K centers = num_clusters * extra_center_factors as it runs and reduces the number of centers from K to k when finalizing model • Valid Values: Either a positive integer or auto • Default Value: auto
3. `init_method`	• Method by which the algorithm chooses the initial cluster centers • The standard k-means approach chooses them at random • K-means++ method chooses the first cluster center at random. Then it spreads out the position of the remaining initial clusters by weighting the selection of centers with a probability distribution that is proportional to the square of the distance of the remaining data points from existing centers • Valid Values: Either random or kmeans++ • Default Value: random

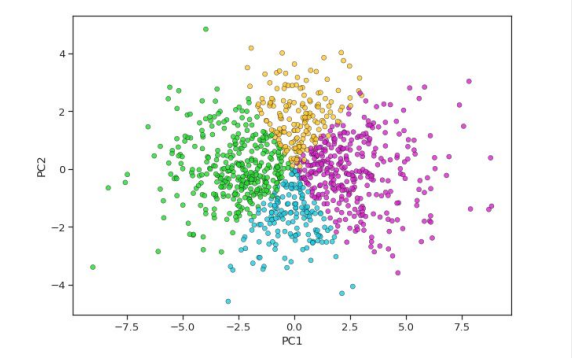
Algorithm 3: Principal Component Analysis (PCA)

- unsupervised machine learning
- reducing dimensionality of a dataset

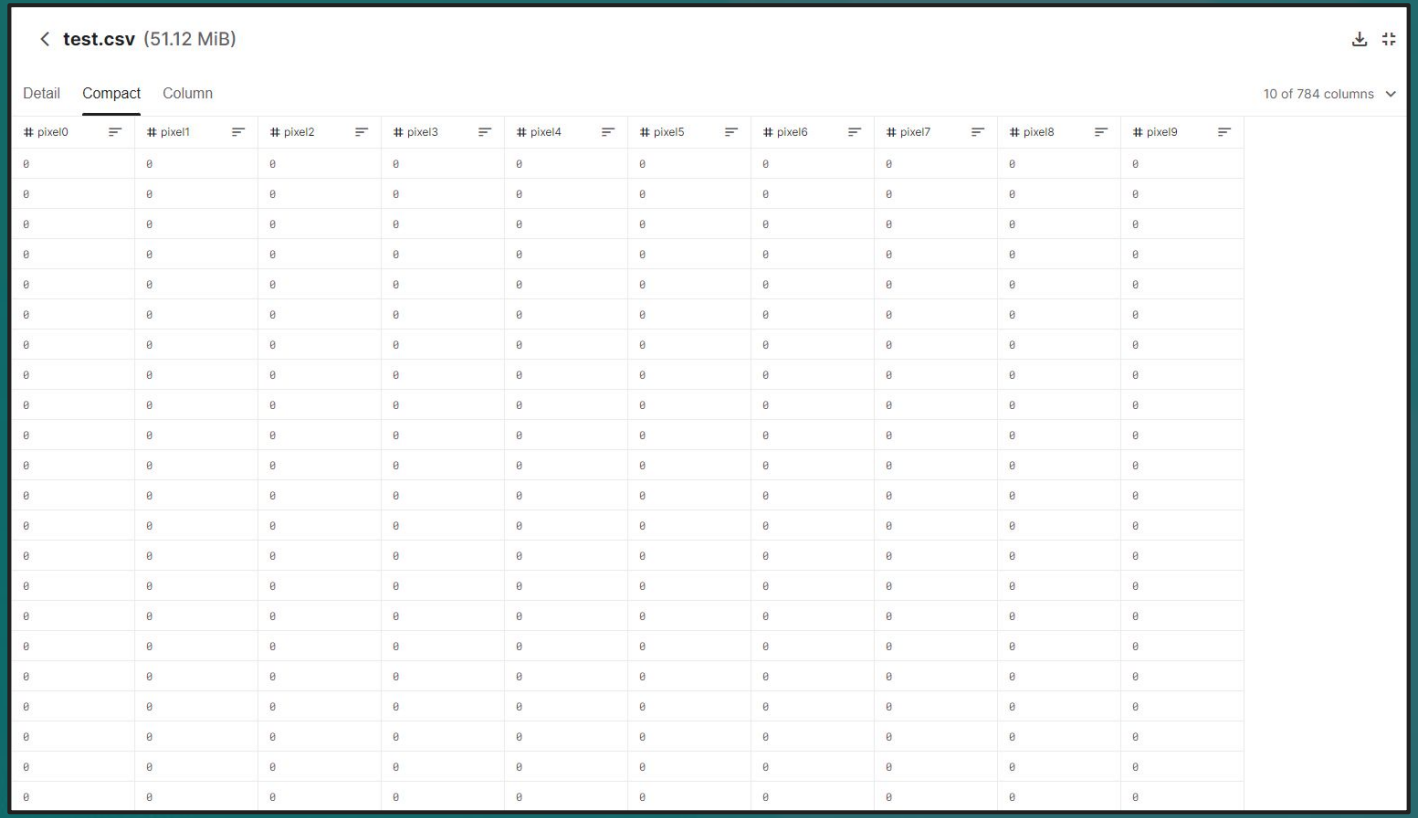
APPLICATIONS:

- Dimensionality reduction/de-noising of dataset for pre-processing or feature engineering
- Image Compression
- Multidimensional data visualization
- General data compression





PCA: Sample Datasets



MNIST Dataset

PCA: Hyperparameters

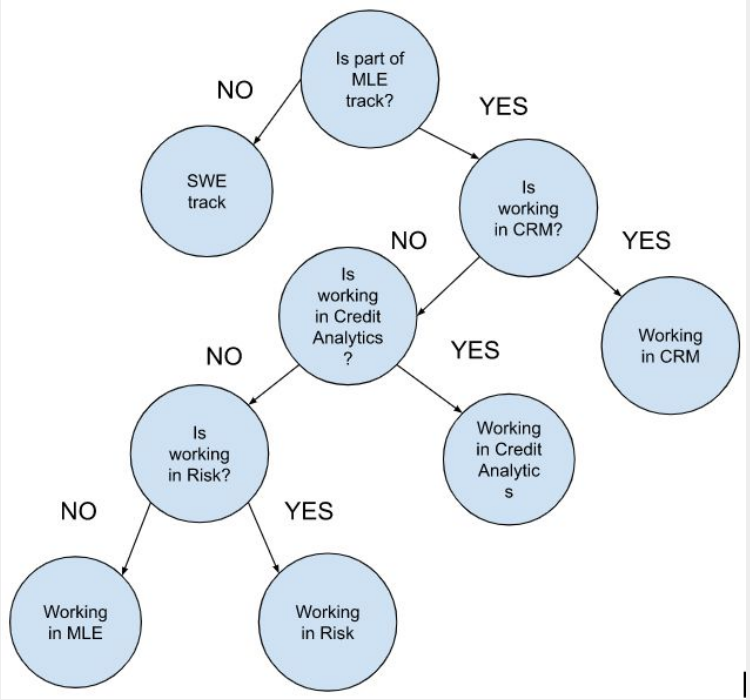
Parameter Name	Description
1. `num_components`	• The number of principal components to compute • Valid Values: Positive integer
2. `algorithm_mode`	• Mode for computing the principal components • Valid Values: regular or randomized • Default Value: regular • regular: For datasets with sparse data and a moderate number of observations and features • randomized: For datasets with both a large number of observations and features. This mode uses an approximation algorithm
3. `subtract_mean`	• Indicates whether the data should be unbiased both during training and at inference • Valid Values: One of true or false • Default Value: true

Algorithm 4: Random Cut Forest (RCF)

- unsupervised method
- anomaly detection

APPLICATIONS:

- Detection of fraudulent transactions on GCash
  - user regular patterns when using GCash
  - users may be grouped based on characteristics and activities
  - anomaly scores assigned to users



Example Decision Tree

RCF: Sample Datasets

Data Description

In this competition you are predicting the probability that an online transaction is fraudulent, as denoted by the binary target `isFraud`.

The data is broken into two files `identity` and `transaction`, which are joined by `TransactionID`. Not all transactions have corresponding identity information.

Categorical Features - Transaction

- `ProductCD`
- `card1` - `card6`
- `addr1`, `addr2`
- `P_emaildomain`
- `R_emaildomain`
- `M1` - `M9`

Categorical Features - Identity

- `DeviceType`
- `DeviceInfo`
- `id_12` - `id_38`

The `TransactionDT` feature is a timedelta from a given reference datetime (not an actual timestamp).

You can read more about the data from this post by the competition host.

Files

- train_{transaction, identity}.csv - the training set
- test_{transaction, identity}.csv - the test set (you must predict the `isFraud` value for these observations)
- sample_submission.csv - a sample submission file in the correct format

IEEE-CIS Fraud Detection on Vesta Card Payment Transactions

Input

order_brush_order.csv: It contains orders information.
Columns: [orderid, shopid, userid, event_time]

Each orderid represents a distinct transaction on Shopee.

Each unique shopid is a distinct seller on Shopee.

Each unique userid is a distinct buyer on Shopee.

Event Time refers to the exact time that an order was placed on Shopee.

Submission Format

Check each shop and determine whether it is deemed to have conducted order brushing. If a shop conducted order brushing, list the userid(s) that are identified as suspicious for the corresponding shopid.

Two columns required:

- shopid
- userid

- If a shop is not deemed to have conducted order brushing, assign the value 0
- Else, list the userid(s) that are identified as suspicious for the corresponding shopid
- If there is more than 1 userid identified as suspicious, list all the userids separated by "&", with the smaller numerical userid first.

shopid	userid
162014252	183926374
321014322	1923237&23421231
22754767	0

Your submission should have 18770 rows (excluding the headers), each with 2 columns.

Teams which do not make a successful submission will be considered to have a score of zero for this challenge

Order Brushing on Shopee

RCF: Hyperparameters

Parameter Name	Description
1. num_samples_per_tree	• Number of random samples given to each tree from the training data set. • Valid values: Positive integer (min: 1, max: 2048) • Default value: 256
2. num_trees	• Number of trees in the forest. • Valid values: Positive integer (min: 50, max: 1000) • Default value: 100
3. feature_dim	• The number of features in the data set (calculated by RCF estimator already, no need to specify) • Valid values: Positive integer (min: 1, max: 10000)
4. eval_metrics	• A list of metrics used to score a labeled test data set. The following metrics can be selected for output: accuracy precision_recall_fscore • Valid values: a list with possible values taken from accuracy or precision_recall_fscore • Default value: Both accuracy, precision_recall_fscore are calculated.

apper.ph