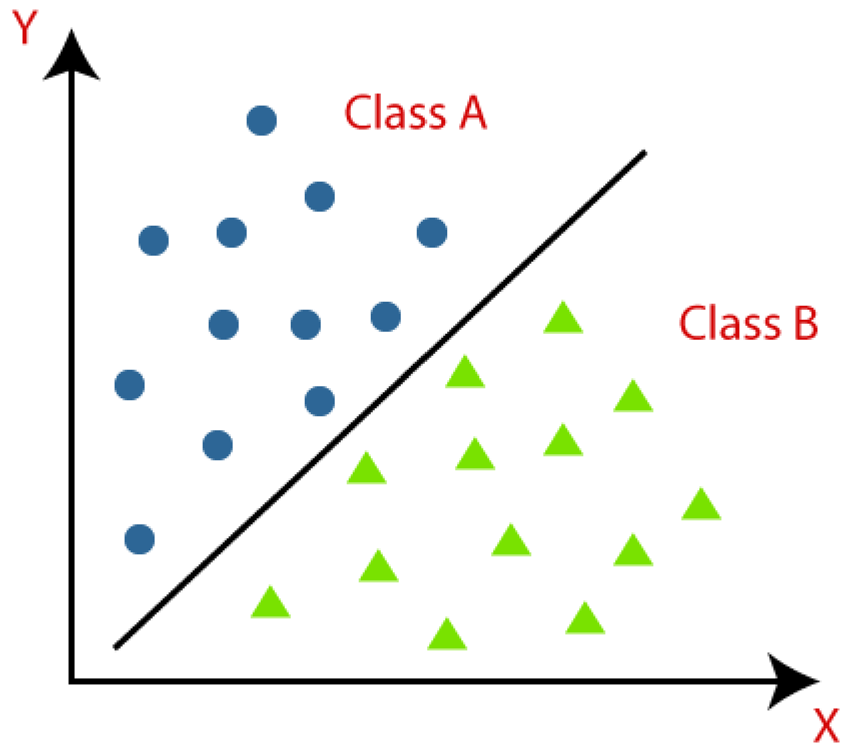
Supervised Learning - Classification

UNIT - II

Topics

- ✓ Basic concepts and applications of classification
- ✓ Naïve Bayes Classification
- ✓ Logistic Regression
- ✓ K-Nearest Neighbors
- ✓ Classification Trees
- ✓ Support Vector Machines
- ✓ Evaluation Measures for Classification Techniques

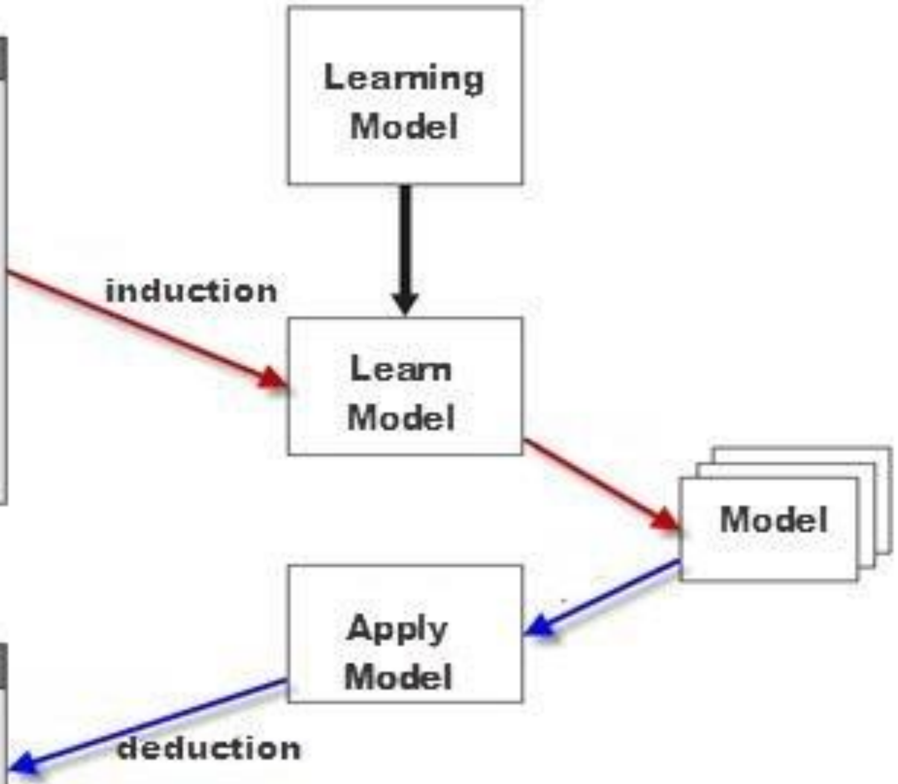


What is Classification?

- classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.
- Task of assigning objects to one of several predefined categories
- **Def** : Classification is the task of learning a target function f that maps each attribute set X to one of the predefined class labels y.
  - The target function is known as **classification model**.
- Examples
  - Classification of email as spam or ham
  - Classification of handwritten digits

Training Set

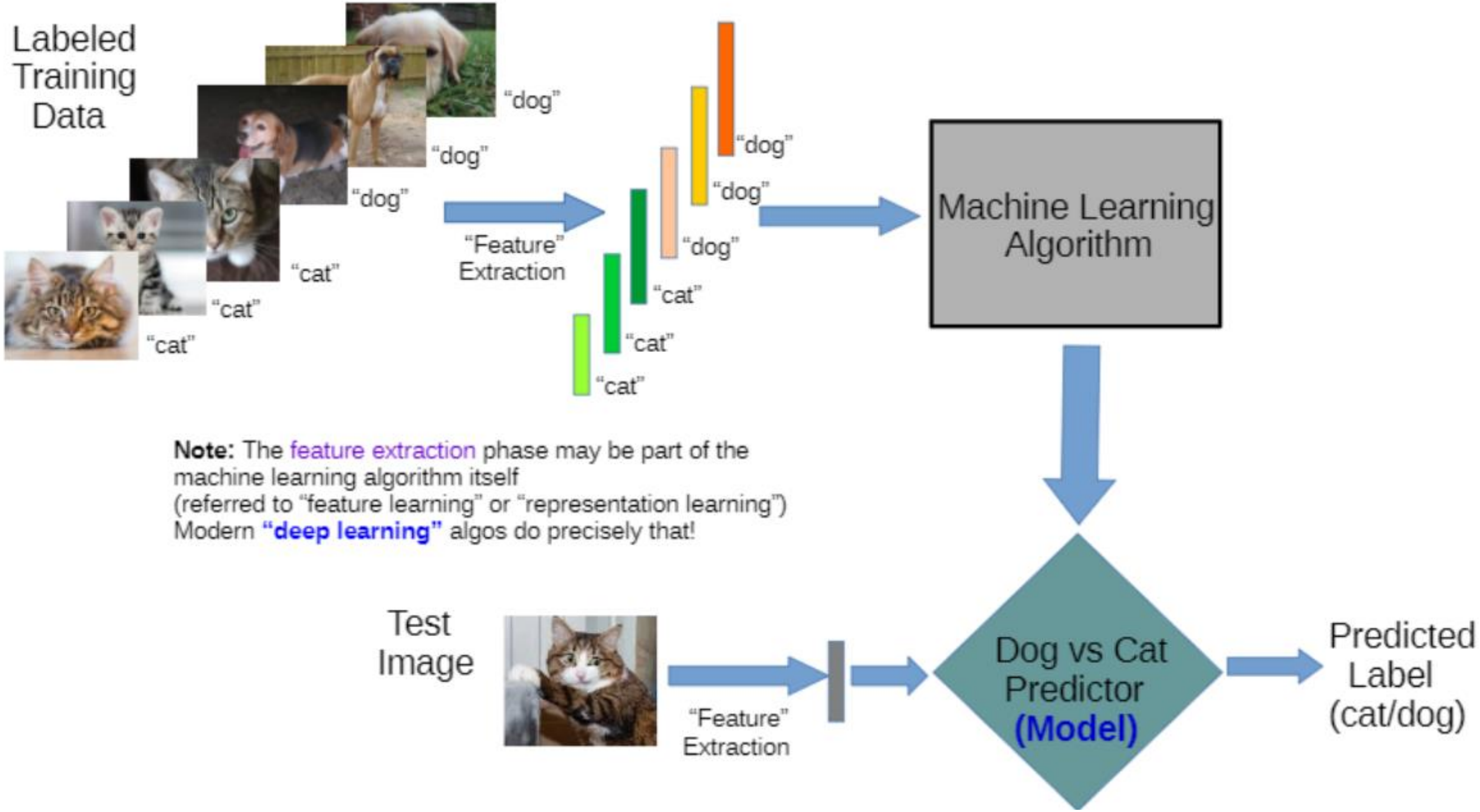
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes



Testing 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	?

Typical Classification Problem



More about classification

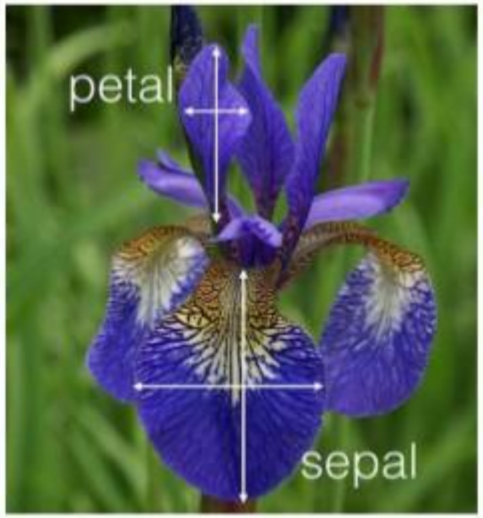
- The input data for classification task is a collection of records.
  - Each record is a tuple (X, y)
  - X is the attribute set, y is the label
  - Class label y must be **discrete**
  - The attribute set can contain both **discrete and continuous**.
- Classification algorithms are most suited for predicting or describing datasets with binary or nominal categories.

Classification Algorithms

- Various algorithms present to do classification are
  - Decision Tree Classifier
  - KNN
  - Neural Networks
  - SVM
  - Logistic Regression
  - Naïve Bayes
- Each technique employs a learning algorithm
- Key objective of these algorithms is to build models with good generalization capability.

Sample Datasets

Supervised learning *classification* problem
(using the Iris flower data set)



Training / test data

Features | Labels

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica

	Gender	Height	Weight	Index	Status
0	Male	174	96	4	Obesity
1	Male	189	87	2	Normal
2	Female	185	110	4	Obesity
3	Female	195	104	3	Overweight
4	Male	149	61	3	Overweight
5	Male	189	104	3	Overweight
6	Male	147	92	5	Extreme Obesity
7	Male	154	111	5	Extreme Obesity
8	Male	174	90	3	Overweight
9	Female	169	103	4	Obesity

Sample Datasets

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Sample Datasets

age	income	student	credit_rating	buys_computer
<=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

Applications of Classification

- Sentiment Analysis
- Email Spam Classification
- Categorizing cells as malignant or benign based on MRI scans
- Document classification
- Image classification
- Image and speech recognition
- Language Modelling
- Machine Translation

Linear Models

- ✓Makes predictions using a linear function of the input features..

- ✓What is the general prediction formula for linear regression?
  - ✓Linear regression finds the parameters which minimizes the mean squared error between y and $\hat{y}$.

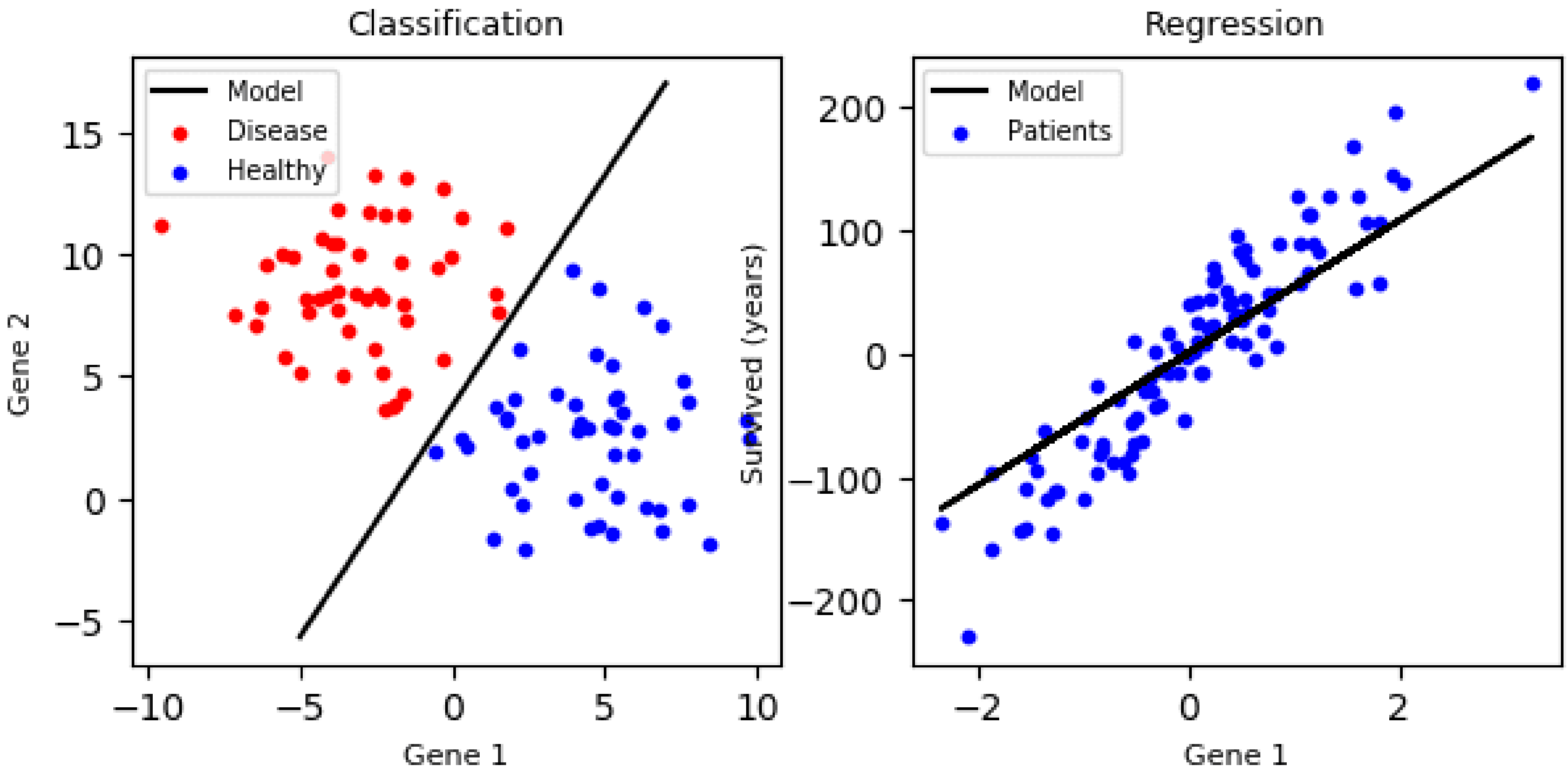
- ✓Can we use a linear model for classification?

- ✓ linear models can also be used for classification by following

$$\hat{y}=wx+b>0$$

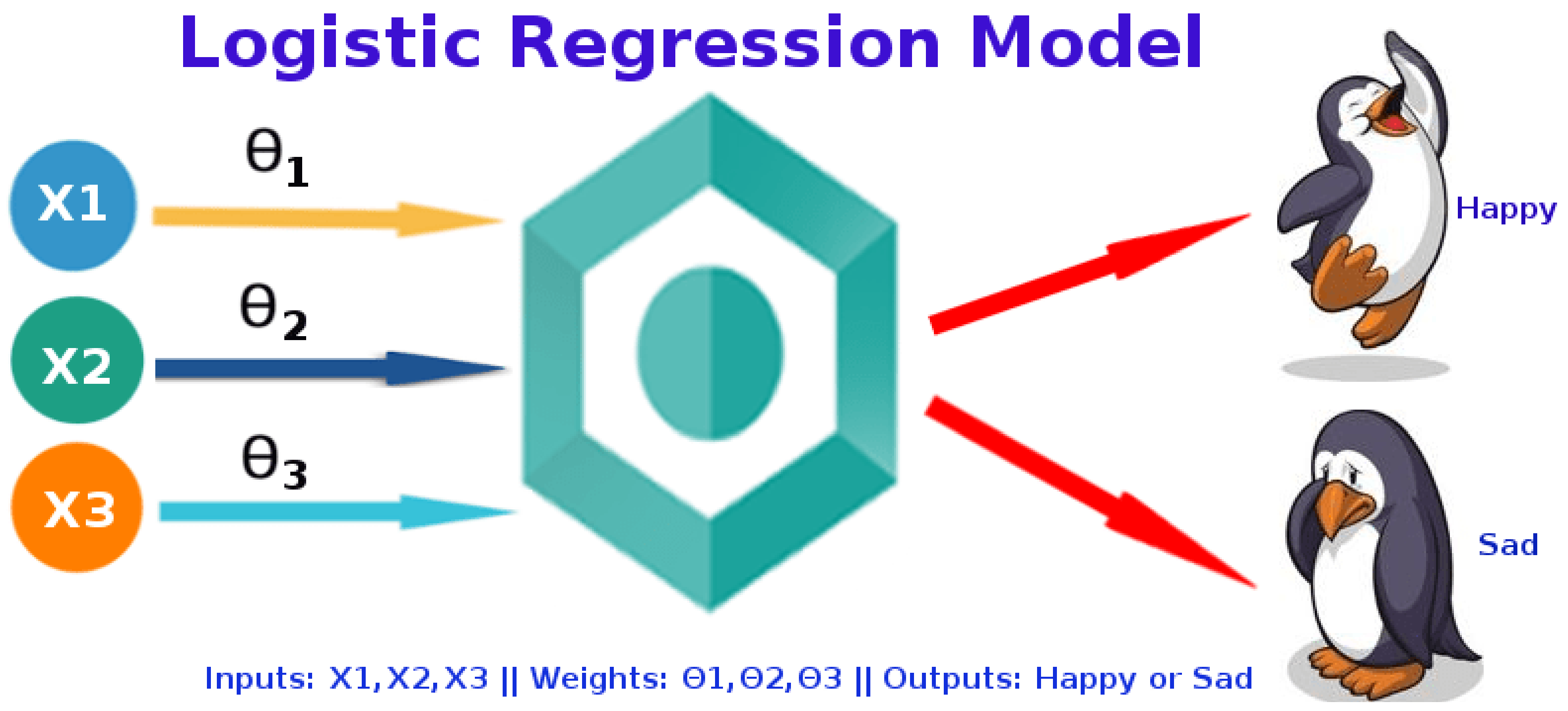
- ✓ Instead of just returning the linear sum, we threshold the predicted value at 0
  - ✓If **wx+b<0** then the predicted class is **-1**.
  - ✓If **wx+b>0** then the predicted class is **1**.

- ✓ The two most common linear classification algorithms are
  - ✓Logistic Regression
  - ✓Support Vector Machines.



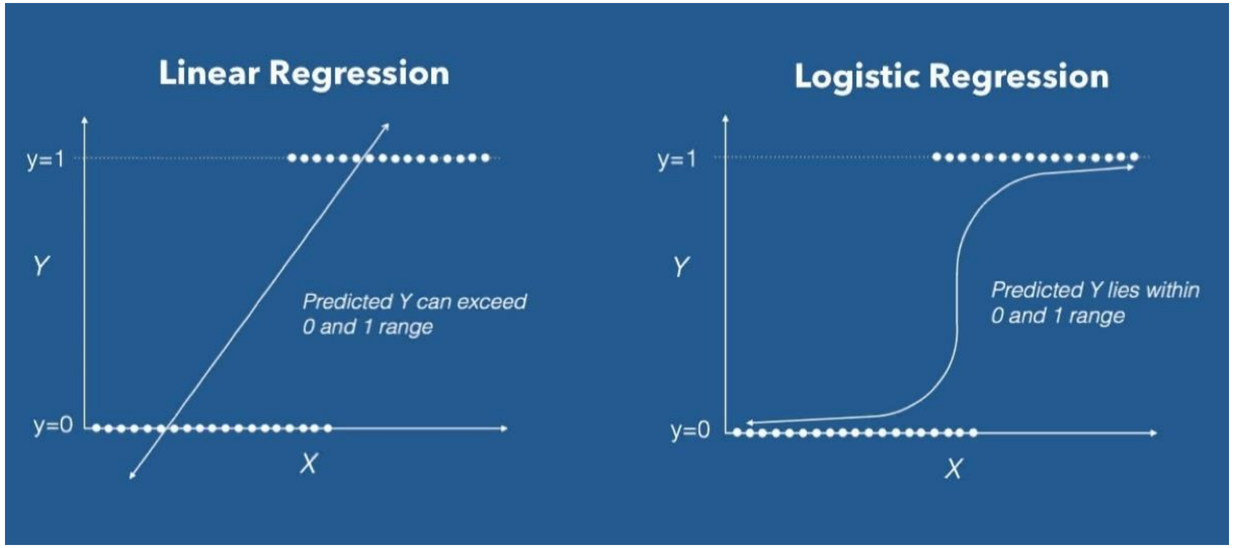
Classification
Model
Disease
Healthy
Gene 2
Gene 1
15
10
5
0
−5
−10
−5
0
5
10
Regression
Model
Patients
Survived (years)
Gene 1
200
100
0
−100
−200
−2
0
2

Logistic Regression



Logistic Regression

- A statistical method for predicting binary classes. predicts the probability of occurrence of a binary event utilizing a logit function.
- The target variable is dichotomous in nature.
  - Any examples ...?
- In general, a linear function with threshold creates a classifier. but it causes few problems
  - Outliers leads to misclassifications
  - This classifier announces completely a confident prediction of 1 or 0, even to the examples that are close to the boundary



Logistic Regression Sigmoid function

To resolve the issues - soften the threshold function

The threshold function is approximated using a continuous, differentiable logistic function (sigmoid)

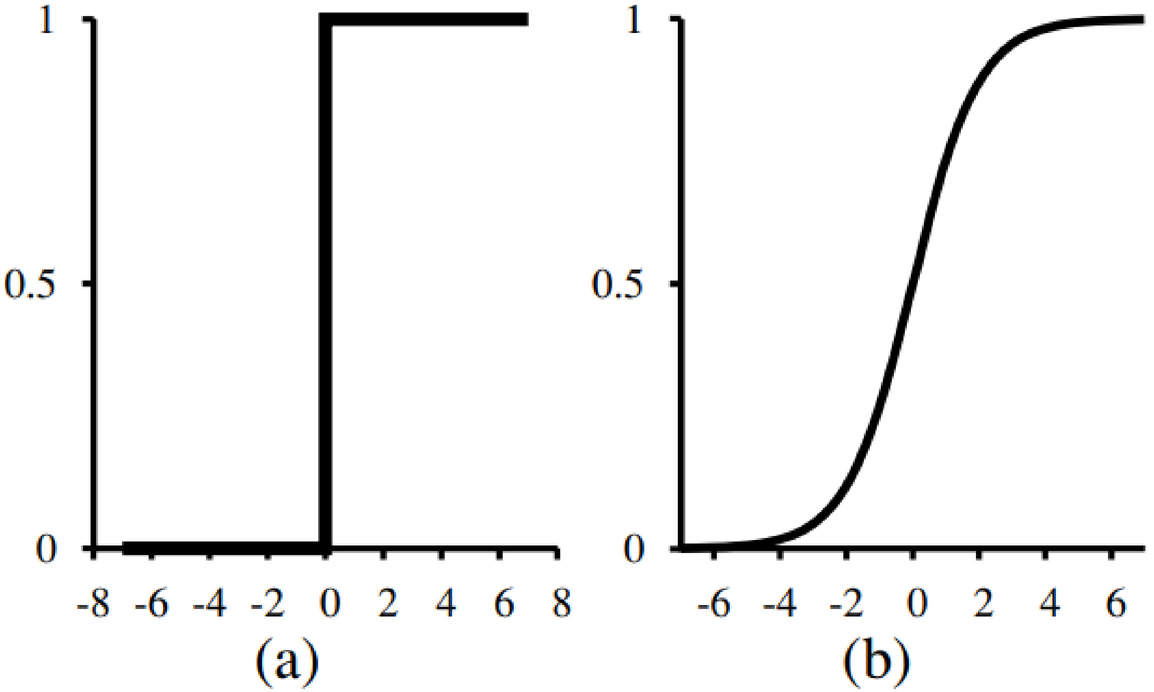


Figure (a) shows the threshold with 0/1 output. It is non-differentiable at z=0

Figure (b) shows the logistic function (also called sigmoid)

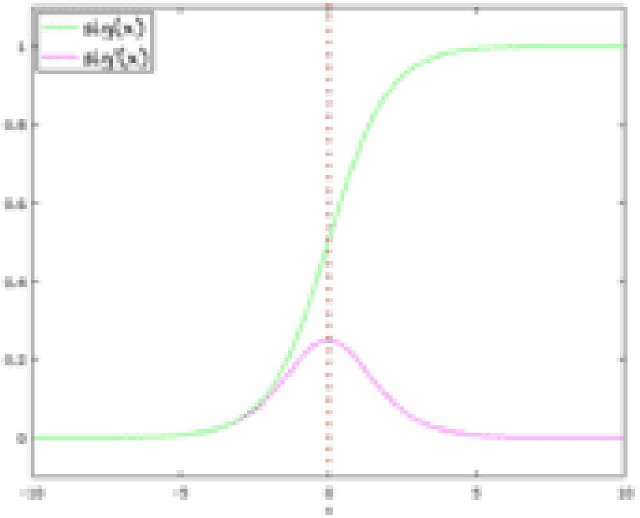
Logistic Regression Sigmoid

$$Logistic(wx) = \frac{1}{1 + e^{-wx}}$$

Sigmoid has most convenient mathematical properties

The output is a number between 0 and 1.

It can be interpreted as a probability of a belonging to a class labelled 1.



Plot of σ(x) and its derivate σ'(x)

Domain: $(-\infty, +\infty)$
Range: $(0, +1)$
$\sigma(0) = 0.5$

Other properties

$$\sigma(x) = 1 - \sigma(-x)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Logistic Regression

The process of fitting the weights of the model to minimize loss on a dataset is called Logistic Regression

Why the name Logistic Regression?

Data is fit into linear regression model, which then be acted upon by a logistic function predicting the target categorical dependent variable.

Logistic Regression
Finding optimal values of W…

For a Single Example (X, y)

$$\frac{\partial\, g\big(f(x)\big)}{\partial x} = g^1\big(f(x)\big) \,.\, \frac{\partial\, f(x)}{\partial x}$$

$$\frac{\partial}{\partial\, w_i}\, Loss(W) = \frac{\partial}{\partial\, w_i}\, (y - \hat{y})^{\,2}$$

$$= 2\,(y - \hat{y})\, \frac{\partial}{\partial\, w_i} (y - \hat{y})$$

$$= -2\,(y - \hat{y})\, g^1(w.x)\, \frac{\partial}{\partial\, w_i}\, w.x$$

$$= -\,2\,(y - \hat{y})\, g^1(w.x)\, x_i$$

Derivative of the logistic function is

$$g^1(z) = g(z).\,(1 - g(z))$$

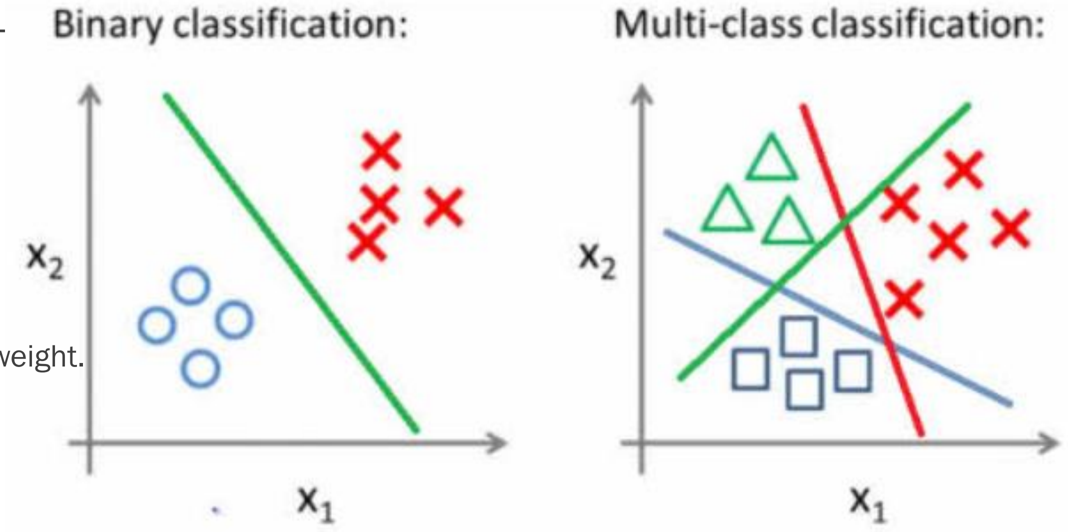
$$g^1(w.x) = g(w.x).\big(1 - g(w.x)\big) = \hat{y}.\,(1 - \hat{y})$$

Weight updates for minimizing the loss

$$w_i \leftarrow w_i - \alpha\, \frac{\partial}{\partial\, w_i}\, Loss(W)$$

$$w_i \leftarrow w_i - \alpha\, (y - \hat{y})\, \hat{y}\, (1 - \hat{y})\, x_i$$

Binary vs. Multi-class classification



- **Binary Classification**
  - Only two class instances are present in the dataset.
  - It requires only one classifier model.
  - Confusion Matrix is easy to derive and understand.
    - Example:- Check email is spam or not, predicting gender based on height and weight.
- **Multi-class Classification**
  - Multiple class labels are present in the dataset.
  - The number of classifier models depends on the classification technique we are applying to.
  - **One vs. All:-** N-class instances then N binary classifier models
  - **One vs. One**:- N-class instances then N* (N-1)/2 binary classifier models
  - The Confusion matrix is easy to derive but complex to understand.
    - Example:- Check whether the fruit is apple, banana, or orange.

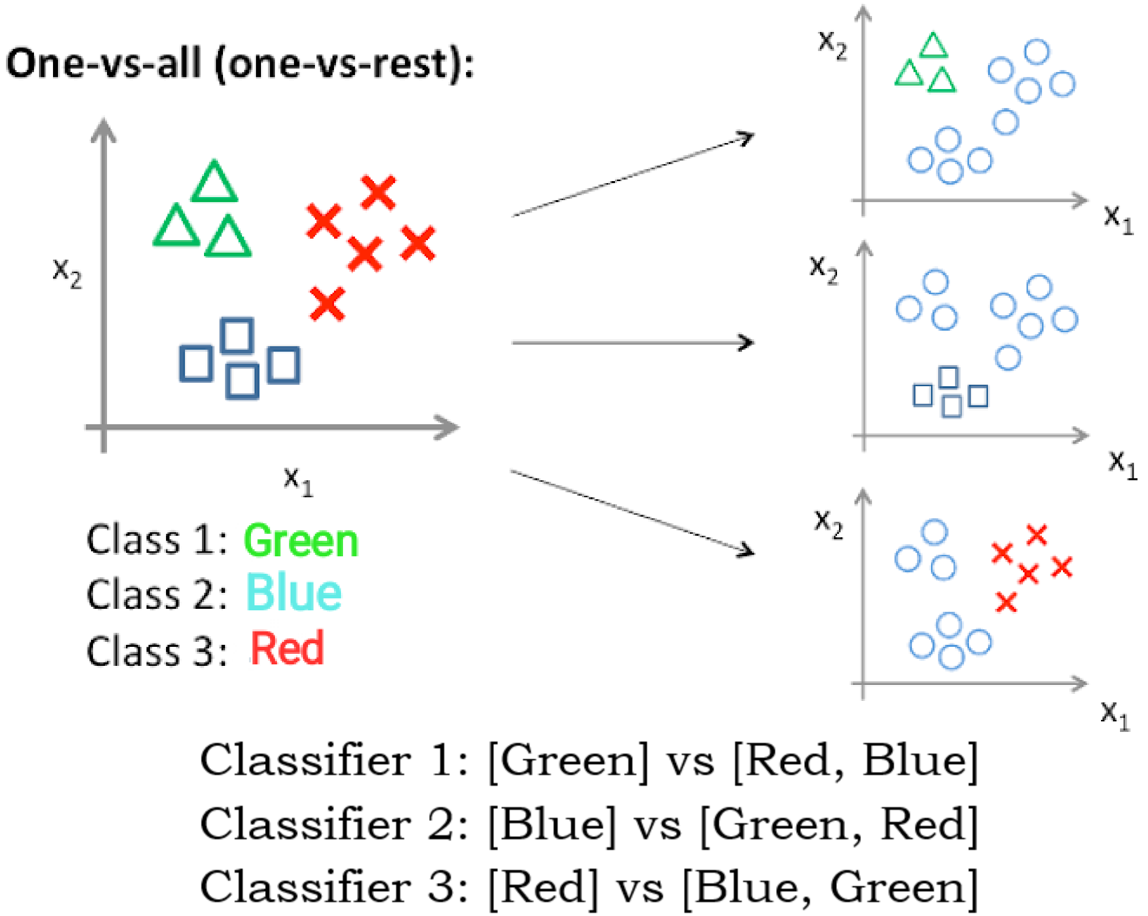
Logistic Regression
Multi-class Classification - Training

one vs. rest (one vs. all) approach –

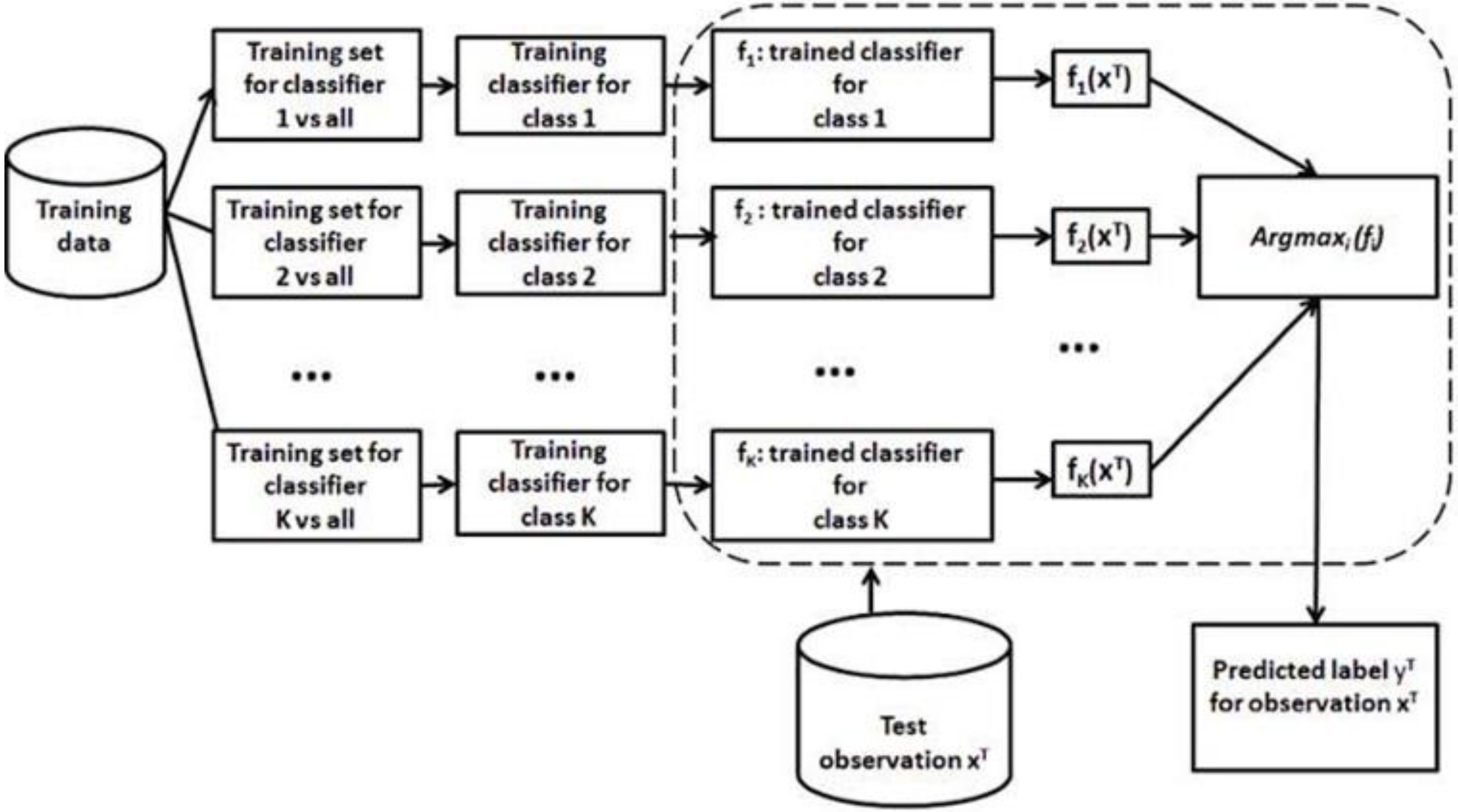
a binary model is learned for each class that tries to separate that class from all the other classes.

for the N-class instances dataset, it generates the N-binary classifier models

The number of class labels present in the dataset and the number of generated binary classifiers must be the ___________



Logistic Regression
Multi-class Classification - Predictions



Logistic Regression Advantages

- Logistic regression is easier to implement, interpret, and very efficient to train.
- It makes no assumptions about distributions of classes in feature space.
- It can easily extend to multiple classes
- It is very fast at classifying unknown records.
- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable
- It can interpret model coefficients as indicators of feature importance.
- Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets. One may consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.

Logistic Regression Disadvantages

- If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
- It constructs linear boundaries.
- The major limitation is the assumption of linearity between the dependent variable and the independent variables.
- Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.

Confusion Matrix

and Classification Evaluation Metrics

TN	FP
FN	TP

Evaluation measures for classification

What is the need of evaluating a classifier?

How you will evaluate a email spam classifier?

What are positive and negative classes in the above task?

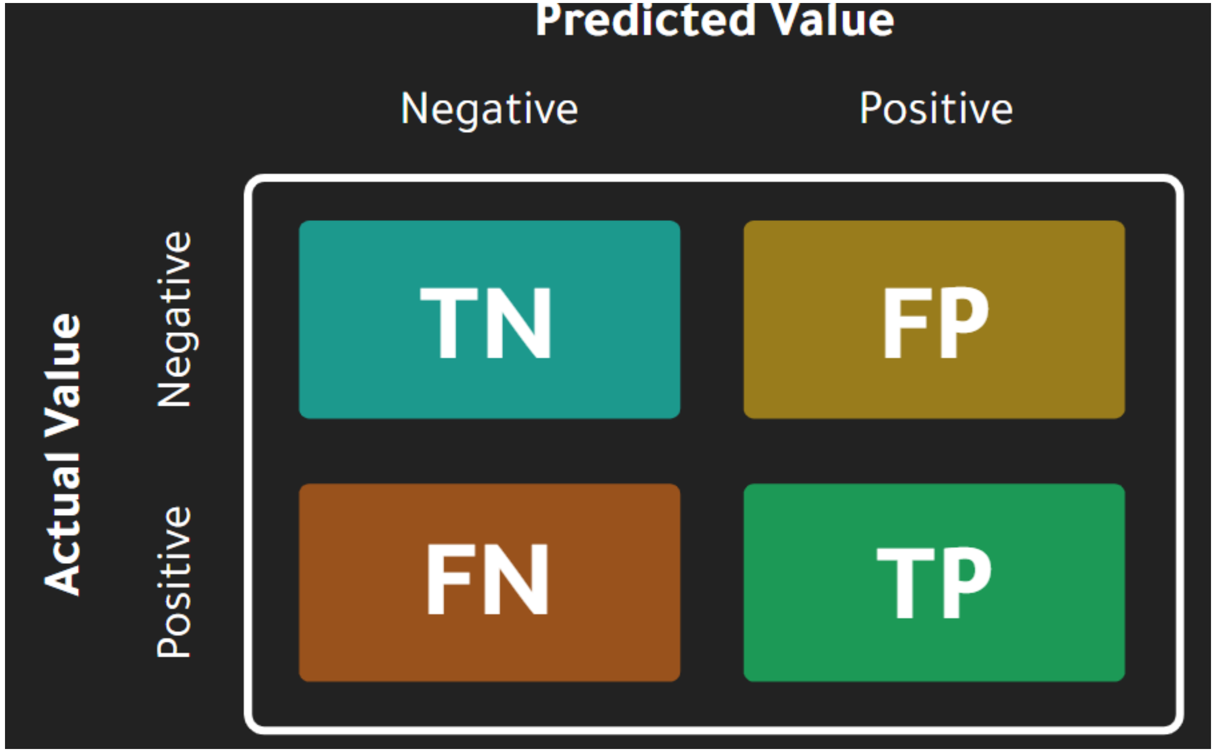
What is a confusion matrix?

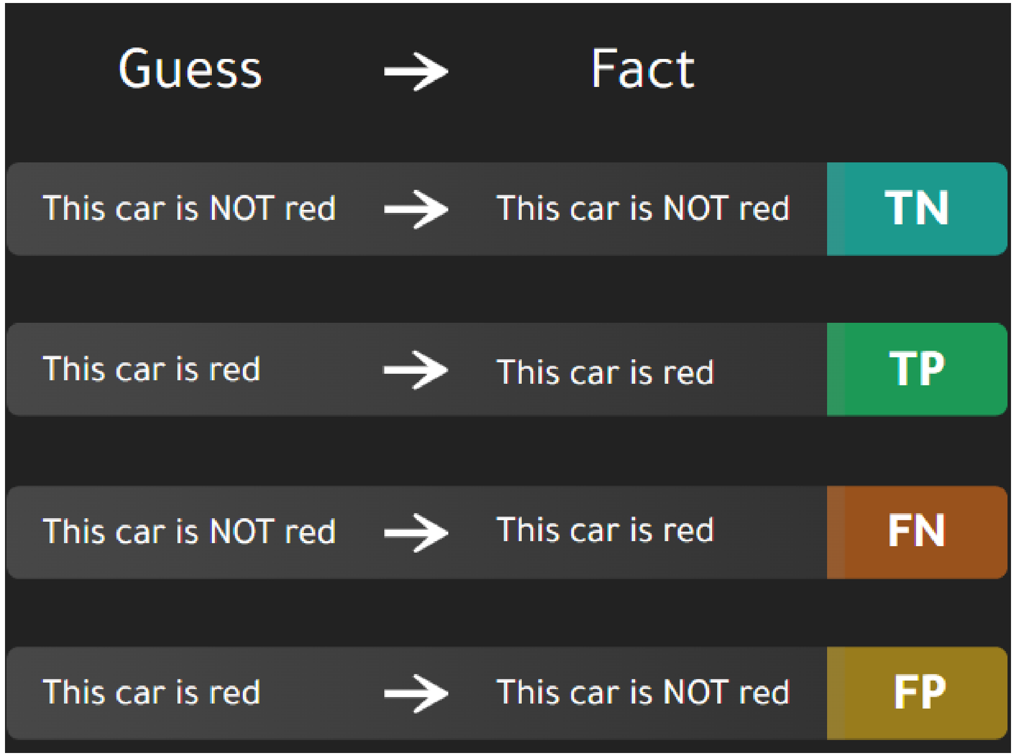
- A confusion matrix is a table for visualizing how an algorithm performs with respect to the human gold labels, using two dimensions (system output and gold labels), and each cell labelling a set of possible outcomes.

Ready for few more new words….

- True Positive, True Negative, False Positive, False Negative

Confusion matrix





TP, TN, FP, FN

In simple words

- A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class.
- A false positive is an outcome where the model incorrectly predicts the positive class. And a false negative is an outcome where the model incorrectly predicts the negative class.

For example, in the spam detection case example,

- true positives are documents that are indeed spam that our system correctly said were spam.
- False negatives are documents that are indeed spam but our system incorrectly labelled as non-spam

gold standard labels

		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** $= \frac{\text{tp}}{\text{tp+fp}}$
	system negative	**false negative**	**true negative**	
		recall $= \frac{\text{tp}}{\text{tp+fn}}$		**accuracy** $= \frac{\text{tp+tn}}{\text{tp+fp+tn+fn}}$

Accuracy

$$\frac{TP + TN}{FP + TP + TN + FN}$$

Positive Predictive Value

Precision

$$\frac{TP}{TP + FP}$$

Sensitivity or True Positive Rate

Recall

$$\frac{TP}{TP + FN}$$

F1-Score

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy

- tells what percentage of all the observations our system labelled correctly
- Not suitable for unbalanced datasets

Precision

- measures the percentage of the items that the system detected (i.e., the system labeled as posit

Recall

- measures the percentage of items actually present in the input that were correctly identified by the systemive) that are in fact positive (i.e.,are positive according to the human gold labels)

F-measure

- a single metric that incorporates aspects of both Precision and Recall.

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

The β parameter differentially weights the importance of recall and precision, based on the needs of an application.

Exercise

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

For the above confusion matrix, calculate and fill the following.

True positives	
False positives	
False negatives	
True negatives	
Precision	
Recall	
F1-Score	
accuracy	

Exercise

TN 9	FP 1
FN 1000	TP 9000

TN 9000	FP 1000
FN 1	TP 9

Evaluating with more than two classes

	gold labels			
system output	urgent	normal	spam	
urgent	8	10	1	**precision**$_u = \frac{8}{8+10+1}$
normal	5	60	50	**precision**$_n = \frac{60}{5+60+50}$
spam	3	30	200	**precision**$_s = \frac{200}{3+30+200}$
	recall$_u = \frac{8}{8+5+3}$	**recall**$_n = \frac{60}{10+60+30}$	**recall**$_s = \frac{200}{1+50+200}$	