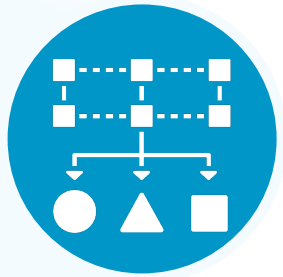
A supervised learning technique: classification

AI | BUSINESS SCHOOL



What is classification?

- Classification is the process of categorizing a given set of data into classes. The pre-defined classes act as our labels, or ground truth.
- The model uses the features of an object to predict its labels. E.g., filtering spam from non-spam emails or classifying types of fruits based on their color, weight and size.

What types of problems does classification solve?

AI BUSINESS SCHOOL

There are two types of classification problems

Binary

The output is restricted to two classes

Multi-class

The output has more than two classes

To solve classification problems: logistic regression

AI BUSINESS SCHOOL

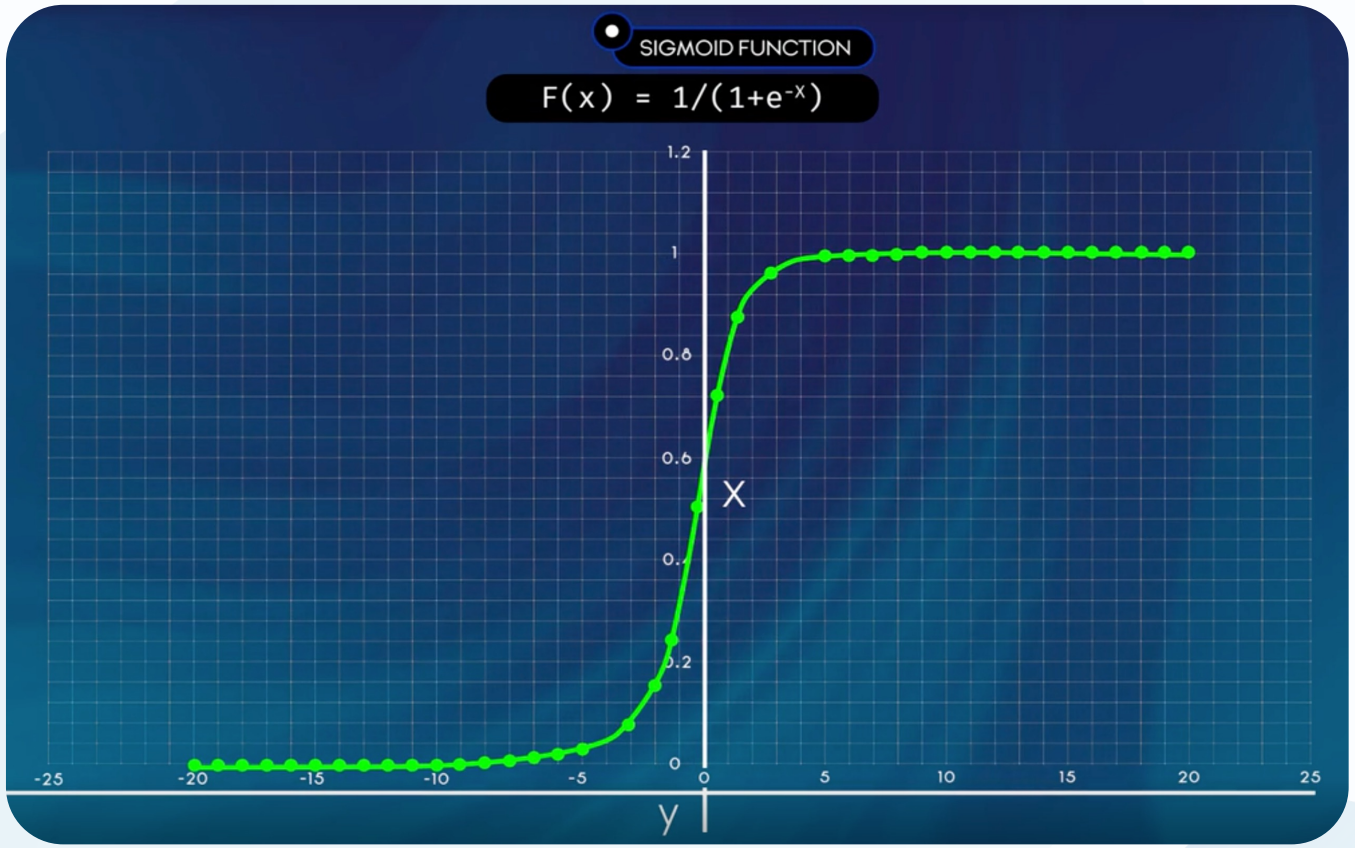
What is logistic regression?

Logistic regression is a linear regression but for classification problems. Unlike linear regression, logistic regression **doesn’t need a linear relationship** between input and output variables.

Logistic regression uses a logistic function: sigmoid function

AI BUSINESS SCHOOL

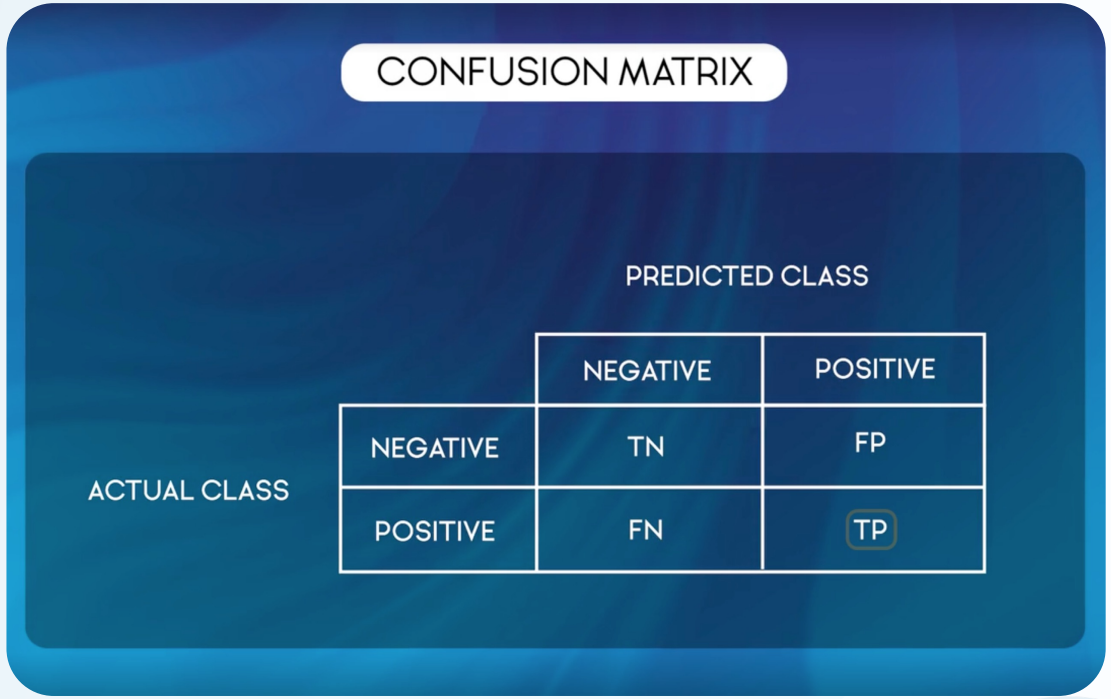
The sigmoid function takes any real input, and outputs a value between zero and one.



How can we measure the performance of a logistic regression classifier?

AI BUSINESS SCHOOL

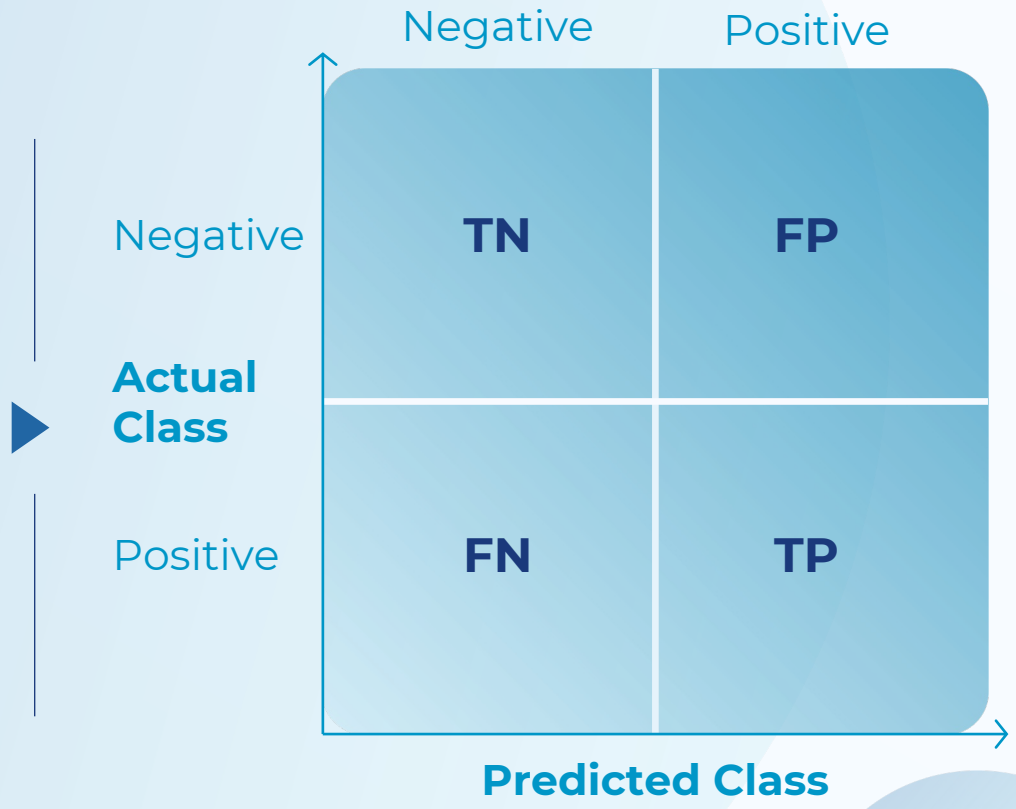
- Once we have the predicted results from our classification model (classifier), the results are compared with the actual label (ground truth)
- Then the performance of the model is being evaluated using the **confusion matrix**



Applying the confusion matrix to measure the model performance

AI BUSINESS SCHOOL

- **True positives (TP)** - results which were predicted as positive & ground truth were also positive.
- **False positives (FP)** - instances predicted as positives but actually were negative.
- **True negatives (TN)** - instances predicted as negatives & their ground truth was also negative.
- **False negatives (FN)** - instances predicted as negative but their ground truth was positive.



Possible Collaboration areas

AI BUSINESS SCHOOL

Accuracy

Indicates how accurately a result can be correctly predicted from the total amount of samples

Precision

Indicates how accurately positive instances were predicted and how many of them are positive

Recall (Sensitivity)

Indicates how many positive samples the classifier has falsely predicted

F1 score (F measure)

Indicates the equi-librium between the precision and the recall

The aim is to maximize true positives & true negatives; minimize false positives & negatives

The evaluation metrics

AI BUSINESS SCHOOL

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

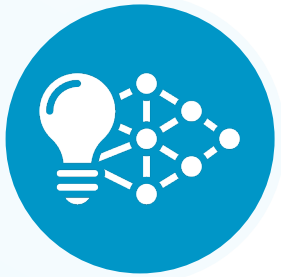
$$Precision = \frac{(TP)}{(TP+FP)}$$

$$Recall = \frac{(TP)}{(TP+FN)}$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Support vector machine (SVM)

AI | BUSINESS SCHOOL



What is support vector machine (SVM)?

- **Support vector machine** (SVM), is a supervised ML technique that can be used to solve classification and regression problems. It is, however, mostly used for classification.
- In this algorithm, each feature & data points are plotted in the space. Then, the SVM model finds boundaries to separates different data samples into specific classes.

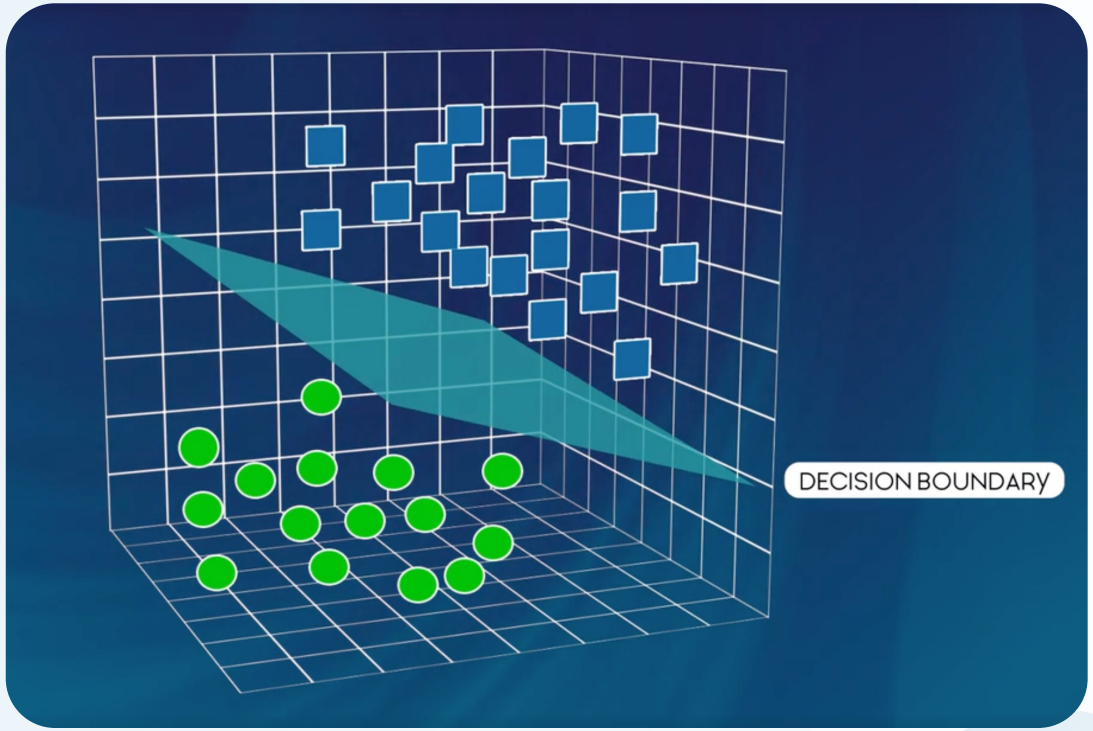
A practical example: finding a 2D plane that differentiates two classes

AI BUSINESS SCHOOL

Let's say we have a dataset of different animals of two classes: birds & fish

- **There are only three features:** body weight, body length, and daily food consumption
- We draw a **3D grid** and plot all these points

- A SVM model will try to find a 2D plane that differentiates the 2 classes



If there are more than three features, we would have a hyper-space

AI BUSINESS SCHOOL

A hyper-space is a space with higher than 3 dimensions like 4D, 5D etc., and a separating line in a dimension higher than 3, is called a **hyper-plane**.

- If the hyper-planes are linear, the SVM is called **Linear Kernel SVM**
- For nonlinear hyper-planes, a **Polynomial Kernel** or other advanced SVMs are used