AI Bootcamp

# Linear Classification

Module 13 Day 1

# Class Objectives

By the end of class, you will be able to:

1. Understand and explain the principles behind linear classification models.
2. Explain how the logistic regression model works as a binary classifier.
3. Explain how the SVM model works as a binary classifier.
4. Implement logistic regression and SVM, and assess their performance on sample datasets.
5. Compare and contrast different data scaling methods.

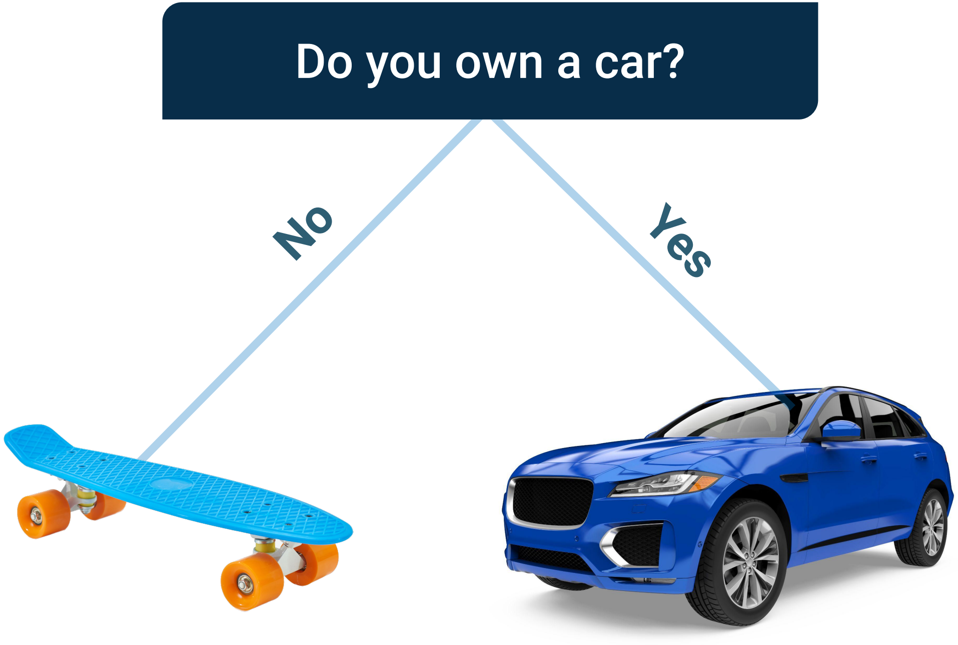
# Instructor **Demonstration**

Classification Overview

**Classification** is a method to predict discrete valued variables. A discrete variable has no middle, and its values cannot be divided.

# Classification

Consider a loan application that asks:



Do you own a car?

No

Yes

1. The possible answers are yes or no.
2. You either own a car or you don’t.
3. There is no middle value, so this type of variable, such as `car_ownership`, would be discrete.

# Classification

We can use classification to draw categorical conclusions about data.

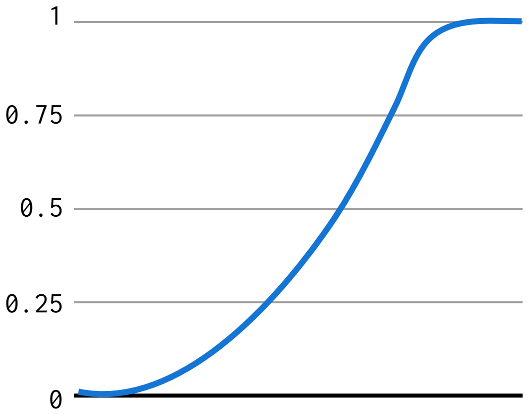
Instead of forecasting quantitative numbers, classification uses a binary (true-positive / true-negative) approach to predict membership in a category (i.e., will the outcome be of type A or type B).

| **Qualified** loan applicant | vs. | **Unqualified** loan applicant |
|---|---|---|
| Creditworthy | | Credit risk |
| 0 | | 1 |

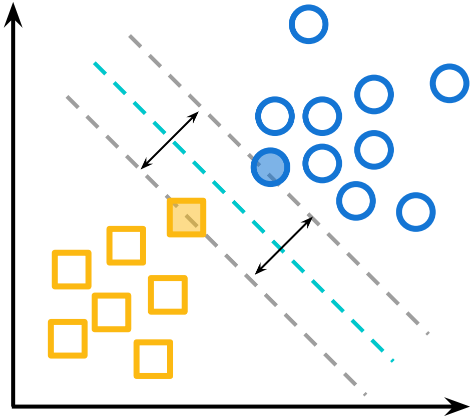
# Classification

Today you'll learn to perform classification using:

## Logistic regression



## Support vector machines (SVM)



**In the next class**, you'll cover k-nearest neighbors (KNN), decision trees, and random forest models.

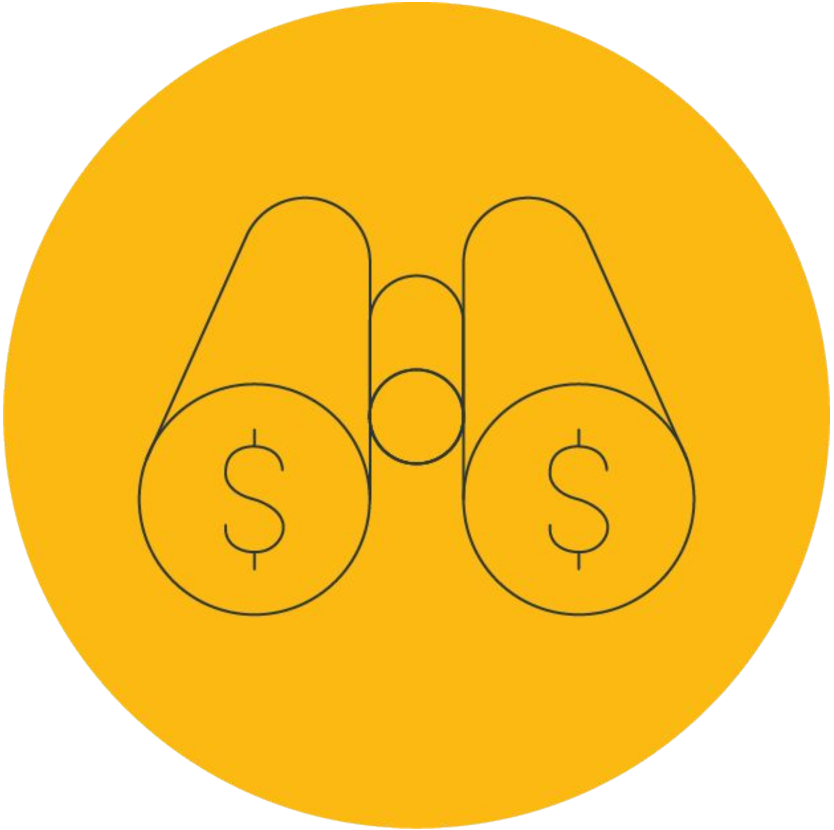
Next

# Classification

Classification models have greatly improved the ability for organizations to properly classify applicants, predict market decline, and classify fraudulent transactions or suspicious activity.

Imagine that we want to build a supervised learning model for a bank so it can determine whether to approve a loan to make a capital investment.

Suppose that we have data about two groups of startups: healthy and unhealthy firms.

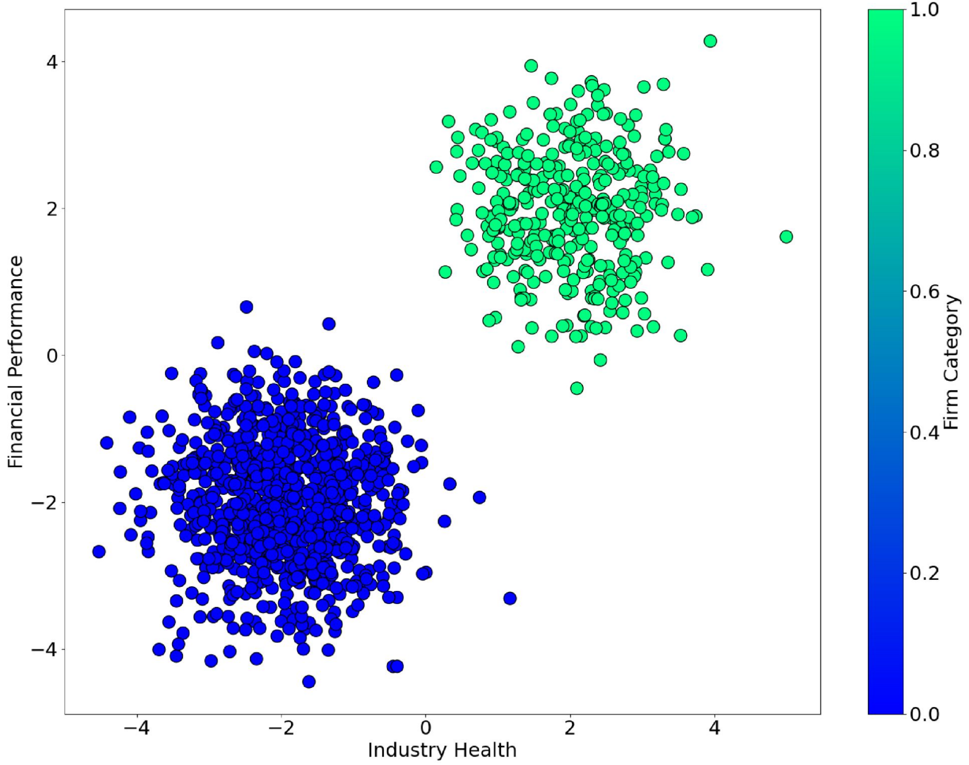


# Classification

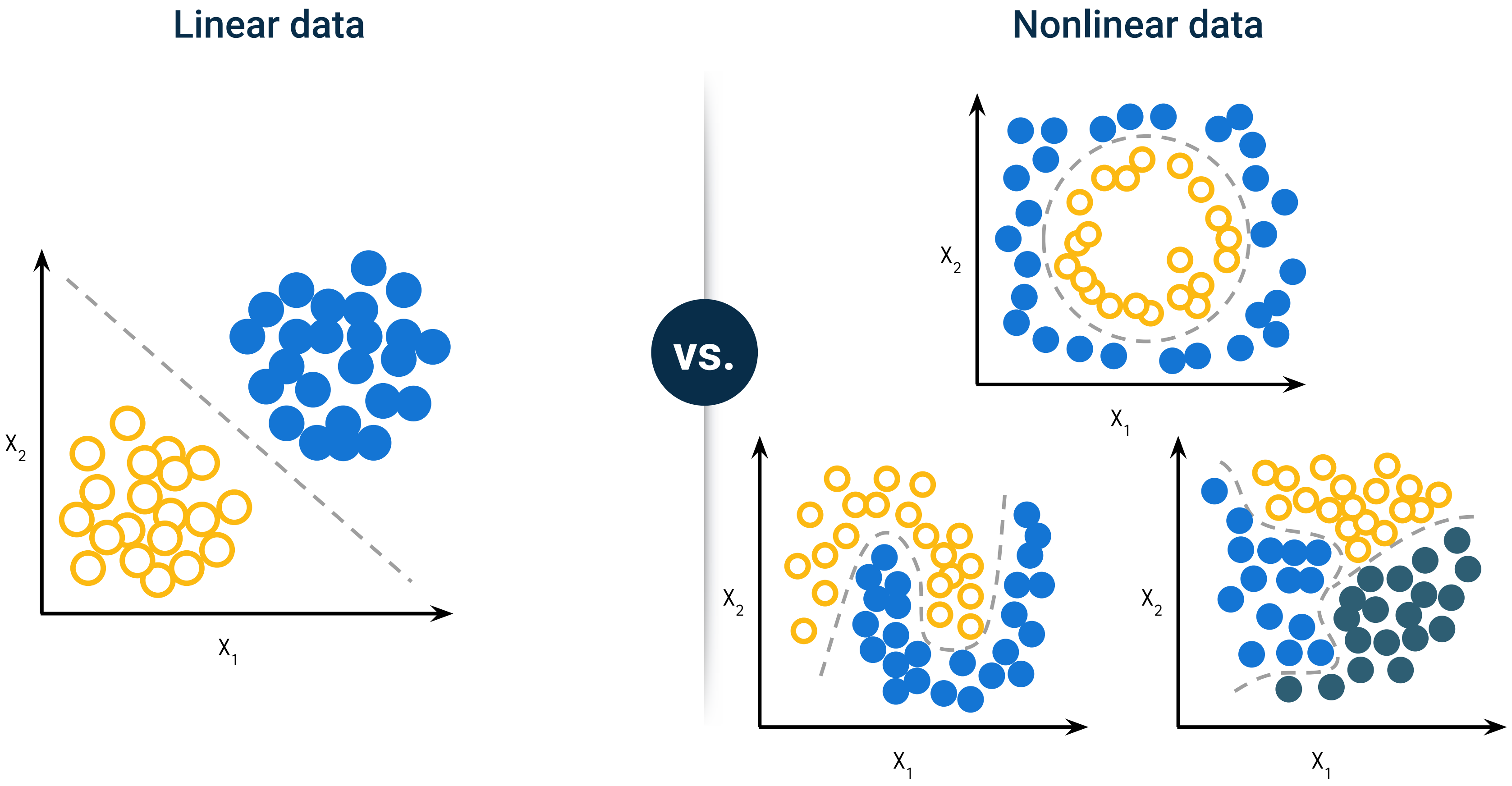
Classifying the health of startups to gauge the risk associated with a bank loan.

Classification models have allowed the financial industry to become more proactive.

Supervised learning algorithms can predict outcomes with a high degree of accuracy, which allows for more effective and efficient mitigation.



# Linear vs. Nonlinear Data



Linear data

Nonlinear data

# Methods for Using Binary Models With Multiclass Classification

## one-versus-rest

1. Bird vs. [cat, dog, fish]
2. Cat vs. [bird, dog, fish]
3. Dog vs. [bird, cat, fish]
4. Fish vs. [bird, cat, dog]

vs.

## one-versus-one

1. Bird vs. cat
2. Bird vs. dog
3. Bird vs. fish
4. Cat vs. dog
5. Cat vs. fish
6. Dog vs. fish

# Activity:

## Discuss Classification Examples

In this activity, you will discuss eight classification examples and determine if they’re binary or multiclass classification problems.

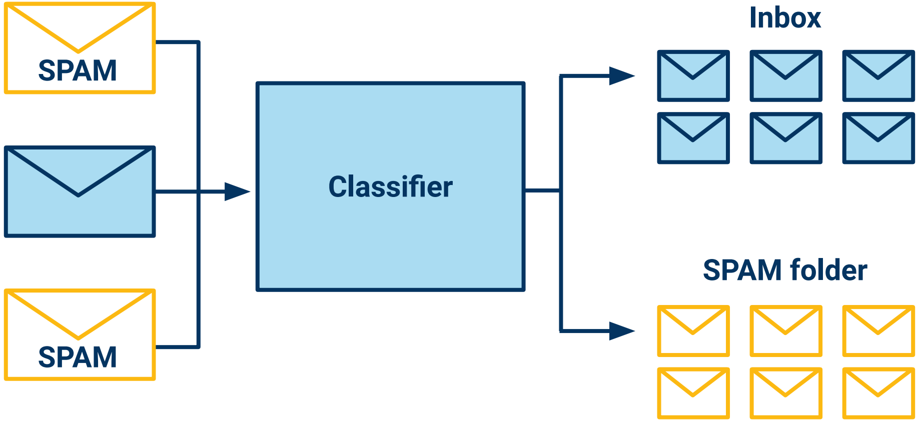
**Suggested Time:**

20 Minutes

# Activity:

Discuss Classification Examples

## 01. Filtering spam messages out of email inboxes



Is this a binary
or multiclass problem?

Click for answer

# Activity:

Discuss Classification Examples

## 02. Predicting whether an image is a cat or a dog

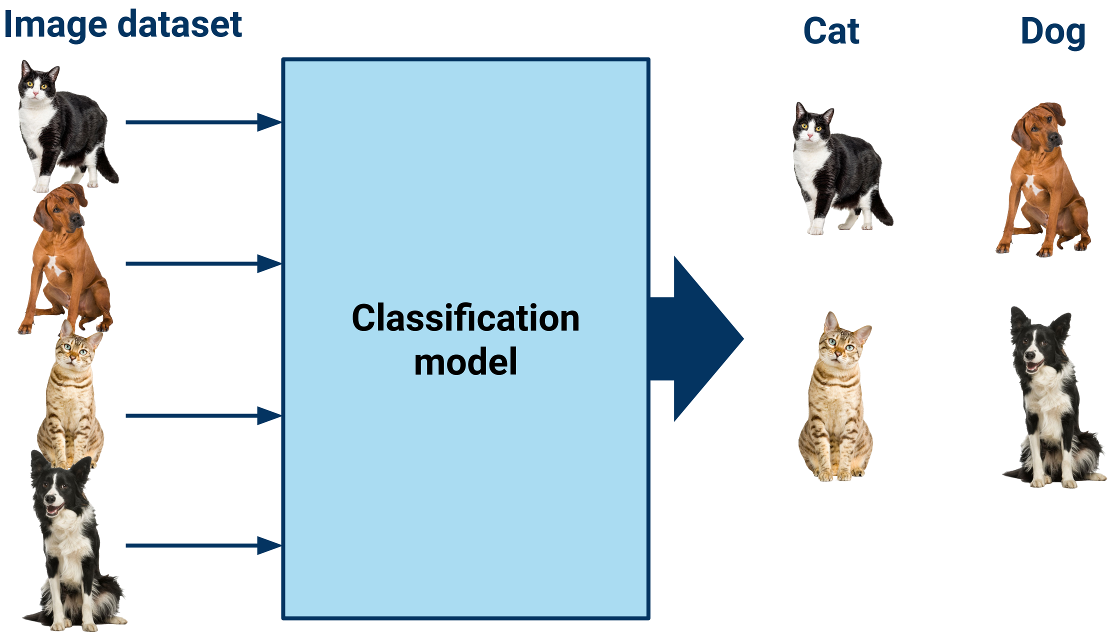


Image dataset

Classification model

Cat

Dog

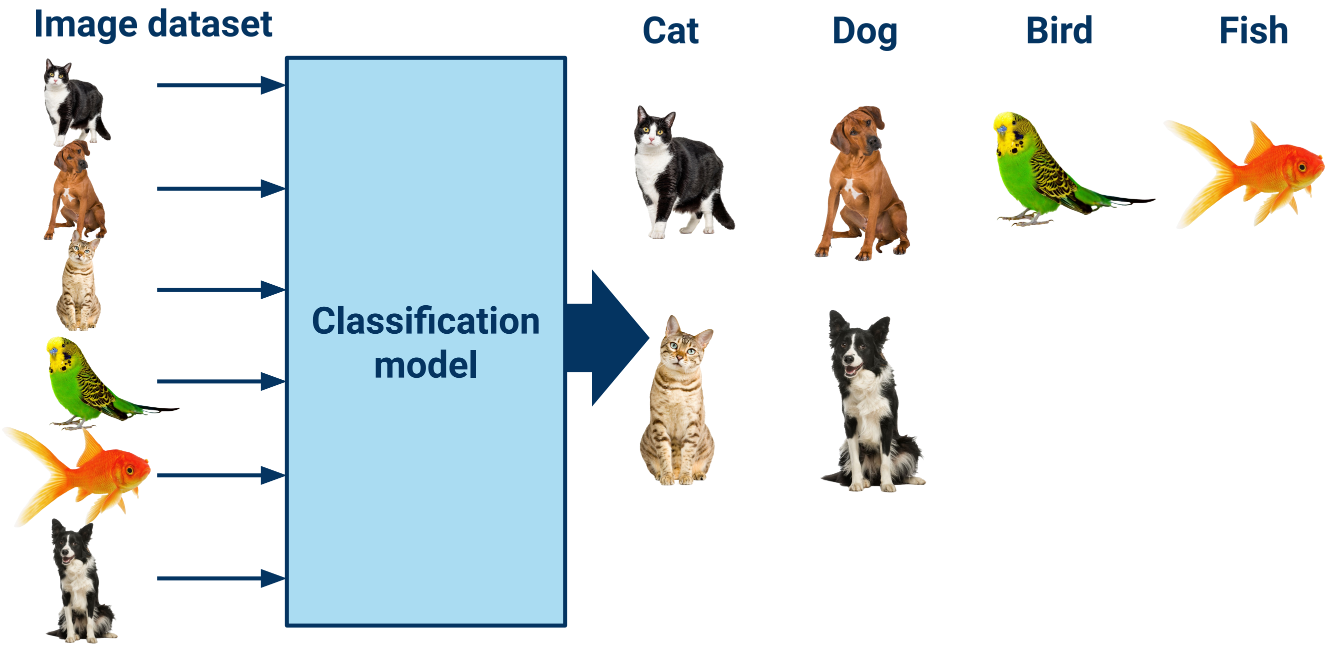
Is this a binary
or multiclass problem?

Click for answer

# Activity:

Discuss Classification Examples

## 03. Predicting whether an image is a cat, dog, bird, or fish



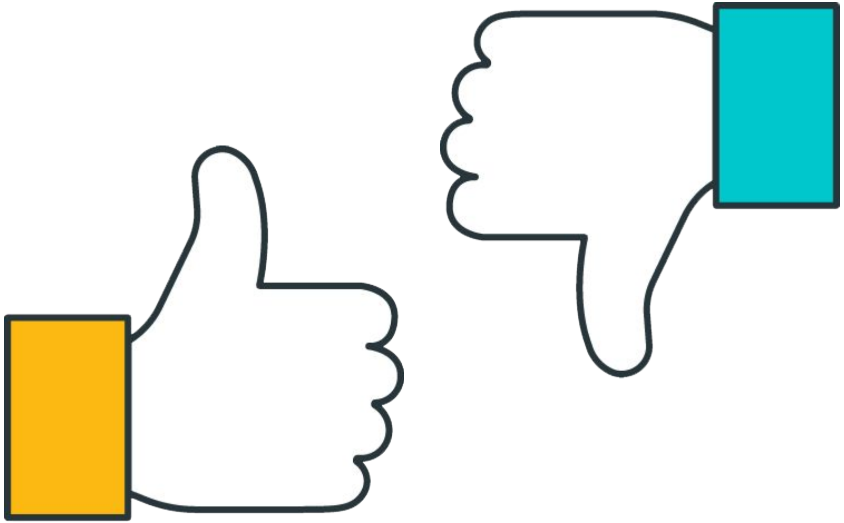
Is this a binary
or multiclass problem?

Click for answer

# Activity:

Discuss Classification Examples

## 04. Classifying customer reviews on social media as positive or negative



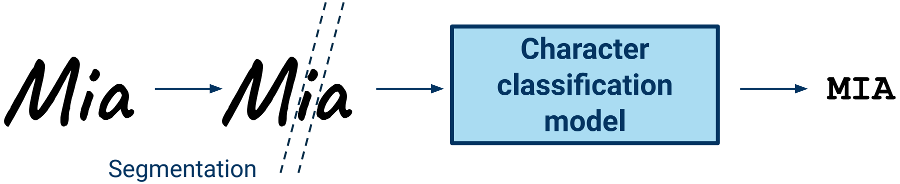
Is this a binary
or multiclass problem?

Click for answer

# Activity:

Discuss Classification Examples

## 05. Identifying individual letters from handwritten text



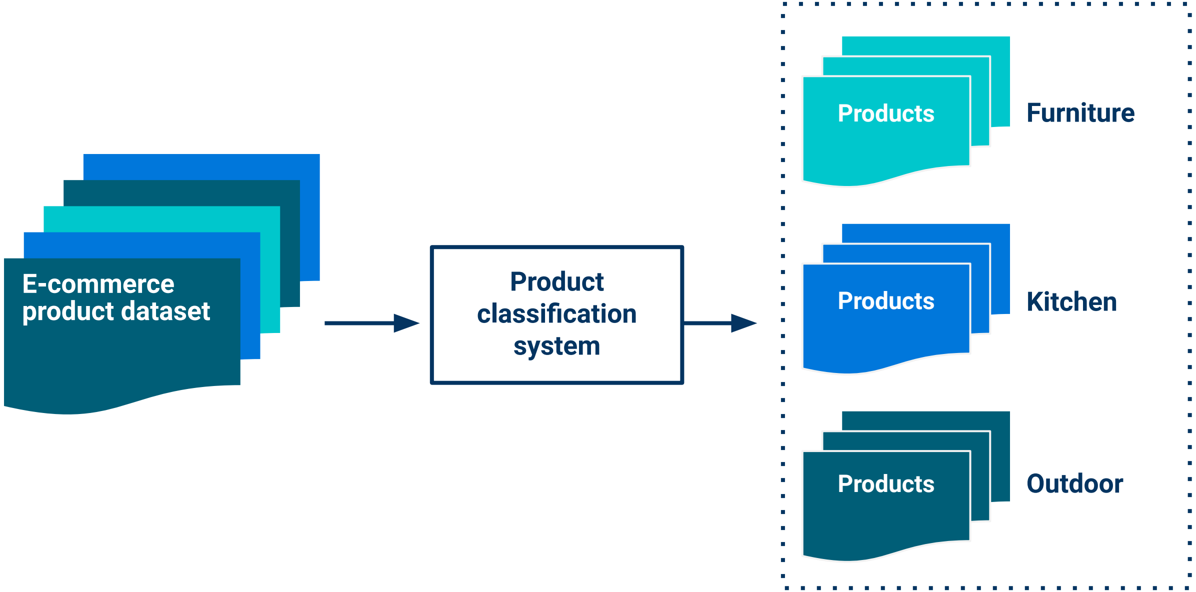
Is this a binary or multiclass problem?

Click for answer

# Activity:

Discuss Classification Examples

## 06. Categorizing e-commerce products as furniture, kitchen, or outdoor items



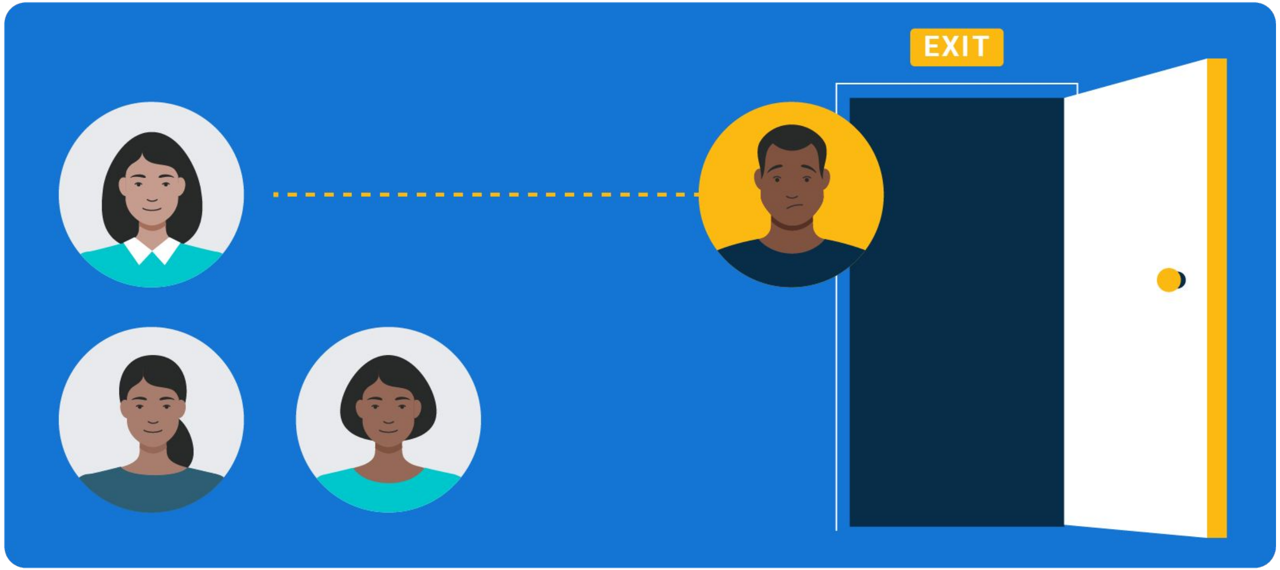
Is this a binary or multiclass problem?

Click for answer

# Activity:

## Discuss Classification Examples

### 07. Predicting whether a customer will churn or not



Is this a binary
or multiclass problem?

Click for answer

# Activity:

Discuss Classification Examples

## 08. Predicting whether credit card transactions are fraudulent or not

Is this a binary
or multiclass problem?

Click for answer

# Questions?

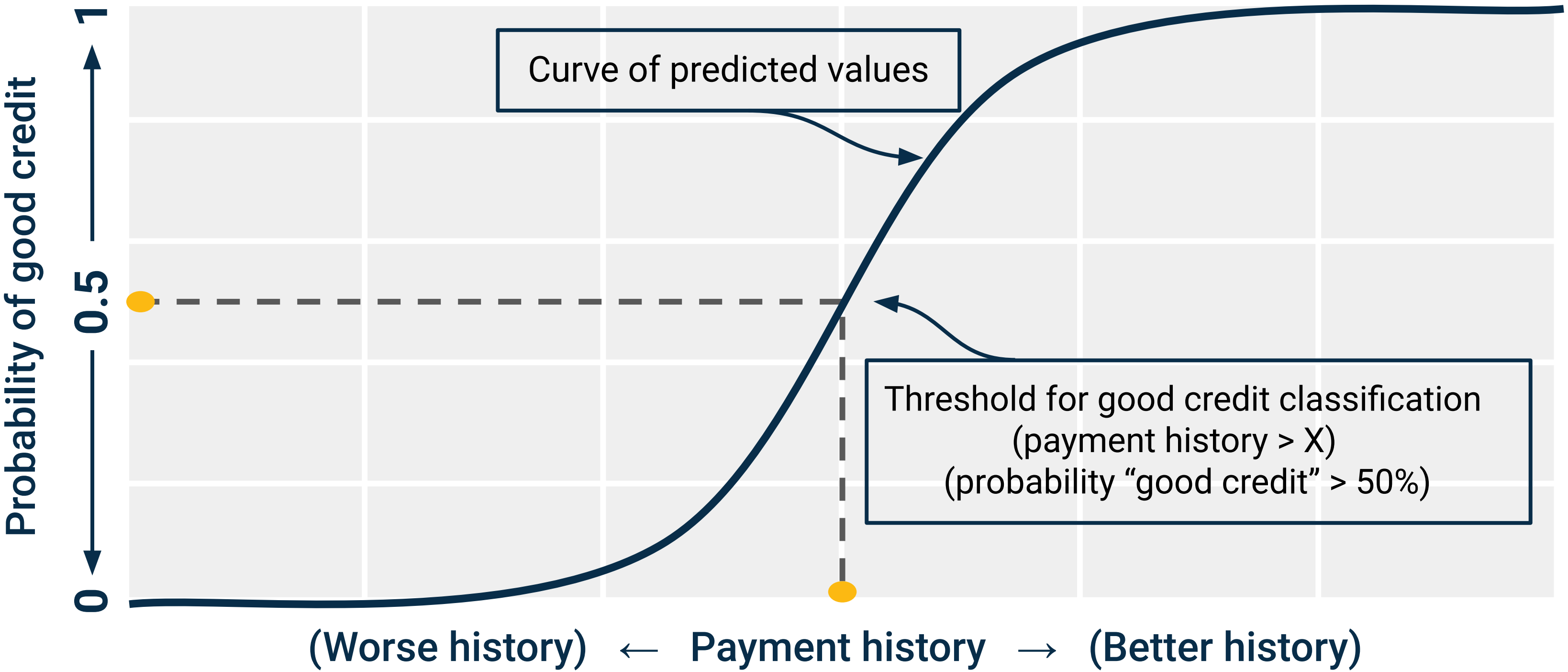
# Instructor **Demonstration**

Logistic Regression

**Logistic regression** is a statistical method for predicting binary outcomes from data.

# Making Predictions With Logistic Regression

Each data point receives a probability of being in the **1** category (e.g., “good credit”).

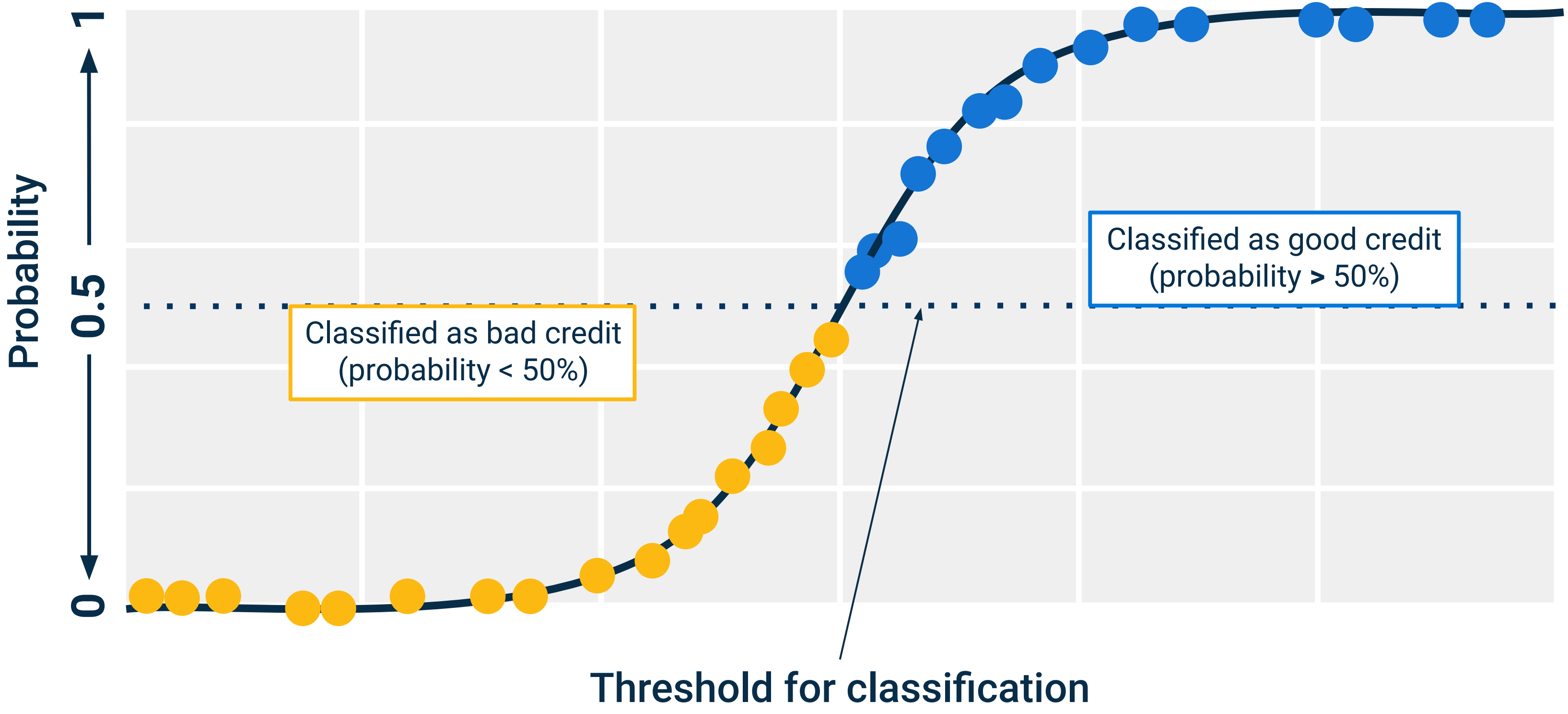


Curve of predicted values

Threshold for good credit classification
(payment history > X)
(probability “good credit” > 50%)

# Making Predictions With Logistic Regression

If the probability is above a certain threshold, that data point is estimated to be a **1** (“good credit”). Below that threshold, the data point is a (**0**).



Classified as good credit (probability > 50%)

Classified as bad credit (probability < 50%)

Threshold for classification

# The Sigmoid Function

Logistic regression converts using a **sigmoid** (or **squashing**) function:

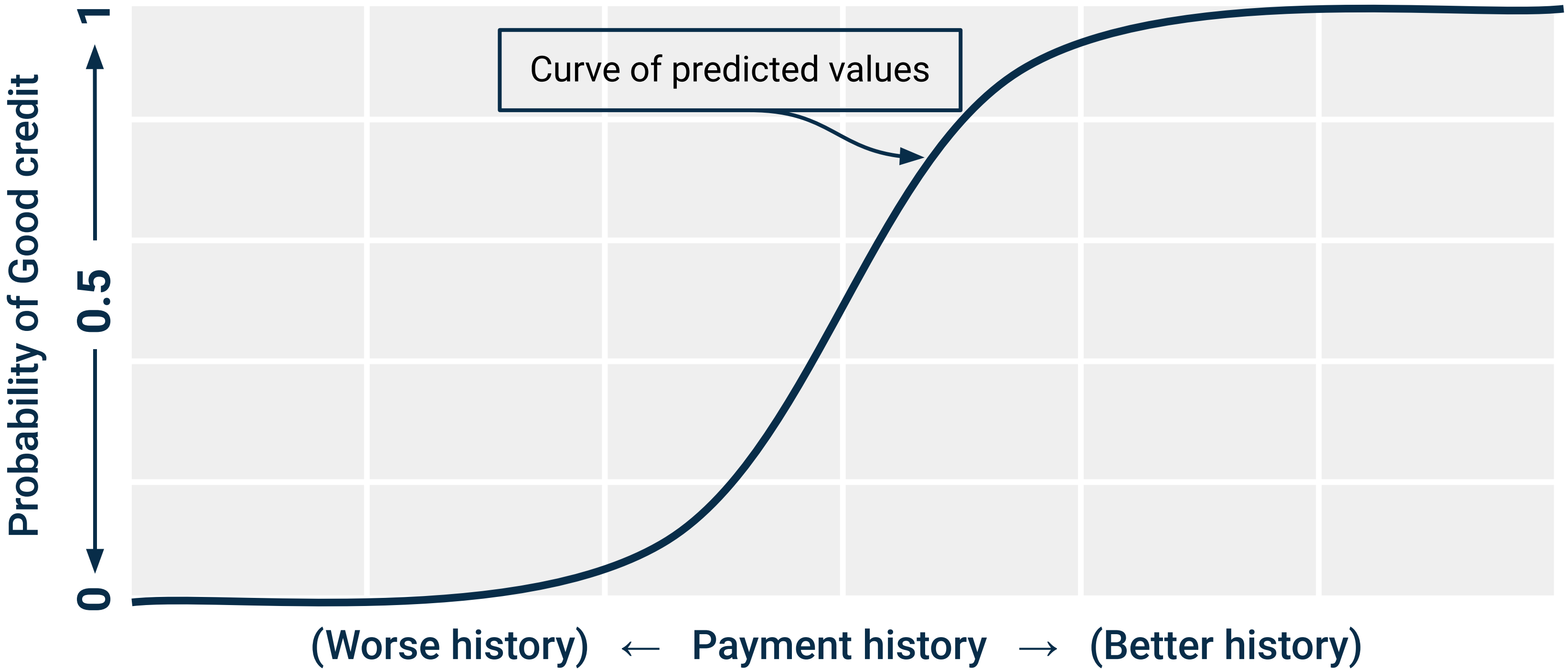
$$Probability\ of\ good\ credit = \frac{1}{1 + e^{-payment\ history}}$$

Performing behind the scenes, this function converts continuous data on the borrower (e.g., number of months without a delinquent payment) to a percentage probability of being a “good credit” borrower.

A good logistic regression model will use more information than just “payment history,” but the **sigmoid** function can still convert all this information into a probability.

# The Sigmoid Function

How do we translate continuous data like **payment history** into a probability of “good credit” ranging from 0 to 1?



# Steps in Logistic Regression Modeling

We can use logistic regression to predict which category or class a new data point should go into.

01

## Preprocess

This step consists of cleaning the data (such as removing rows with missing values) and splitting it into subsets for training and testing the model.

02

## Train

Training is when we use a large subset of our labeled data to teach the model to recognize classification patterns.

03

## Validate

In the validation step, we use a small subset of our labeled data to test how well the model is able to predict labels.

04

## Predict

Finally, we use our model to predict labels for unclassified data.

# Activity:

## Predicting Malware with Logistic Regression

In this activity, you will use logistic regression to identify malware apps that steal private information.

**Suggested Time:**
15 Minutes

# Time’s up!

Let’s review

# Questions?

**Break**

15 mins

# Instructor **Demonstration**

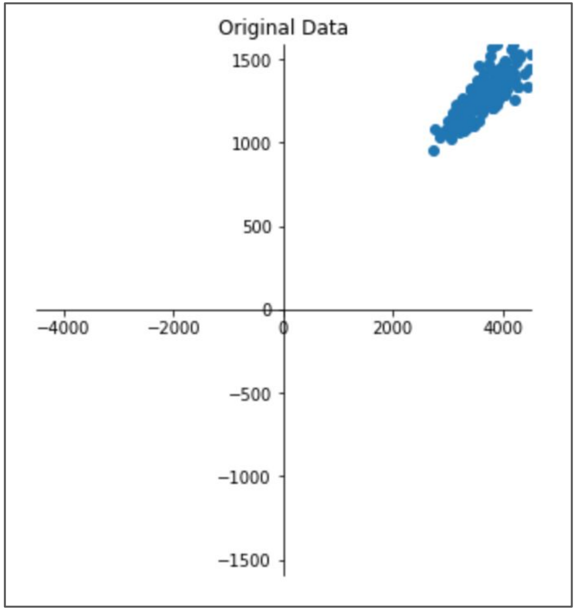
Data Leakage and Preprocessing Revisited

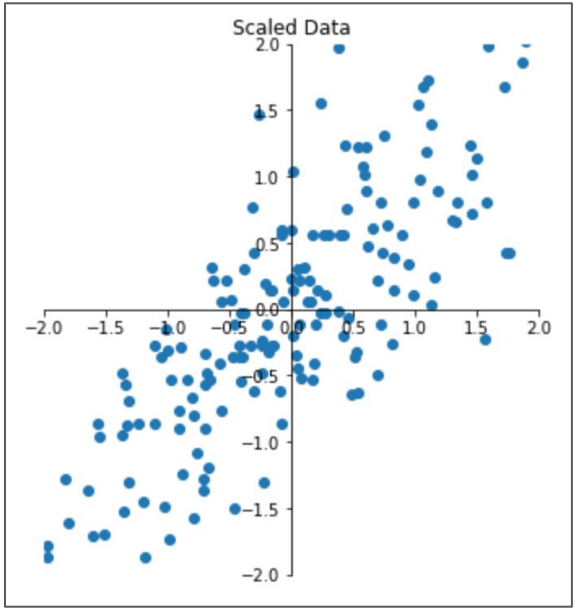
# Scaling/Standardization

We want all features to be shifted to similar numeric scales so that the magnitude of one feature doesn't bias the model during training.

`StandardScaler`

This scales data to have a mean of 0 and variance of 1. You should use `StandardScaler` when you do not have complete knowledge of your data.





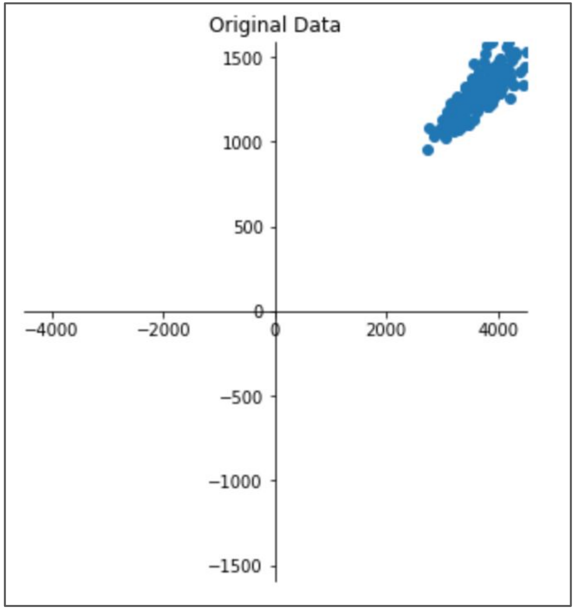
# Scaling/Normalization

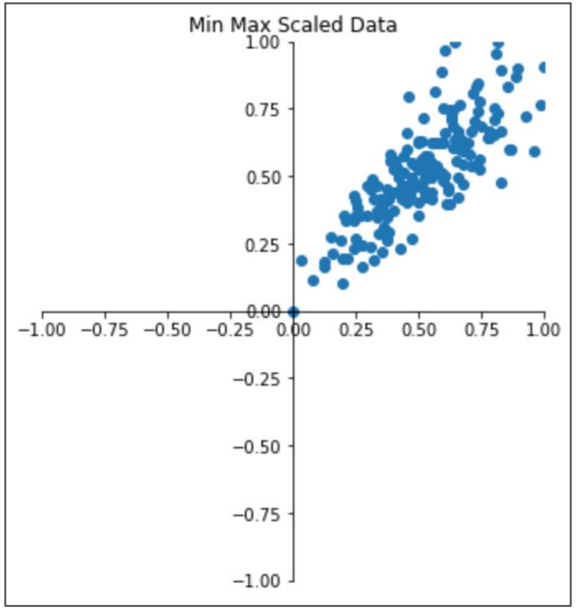
We want all features to be shifted to similar numeric scales so that the magnitude of one feature doesn't bias the model during training.

**`MinMaxScaler`**

`MinMaxScaler` is another scaler available in scikit-learn.

It scales feature data to a minimum of 0 and a maximum of 1.





# Scikit-learn's Preprocessing Paradigm

Preprocessors in scikit-learn follow the Fit -> Transform paradigm, similar to the Model -> Fit -> Predict paradigm for machine learning.

We fit our preprocessor (for example, **`StandardScaler`**) to training data. The preprocessor can be used to transform training data, testing data, or data to be predicted by a trained model.

## Make sure you fit preprocessors to training data!

If you fit your preprocessors before splitting your data, you are biasing the model with information from the testing set.

Remember, the test dataset is supposed to represent new data for the model to predict.

# Activity:

## Logistic Regression with Scaling Preprocessing

In this activity, you will practice scaling data prior to training a logistic regression model.

**Suggested Time:**

20 Minutes

# Time’s up!

Let’s review

# Questions?

# Instructor **Demonstration**

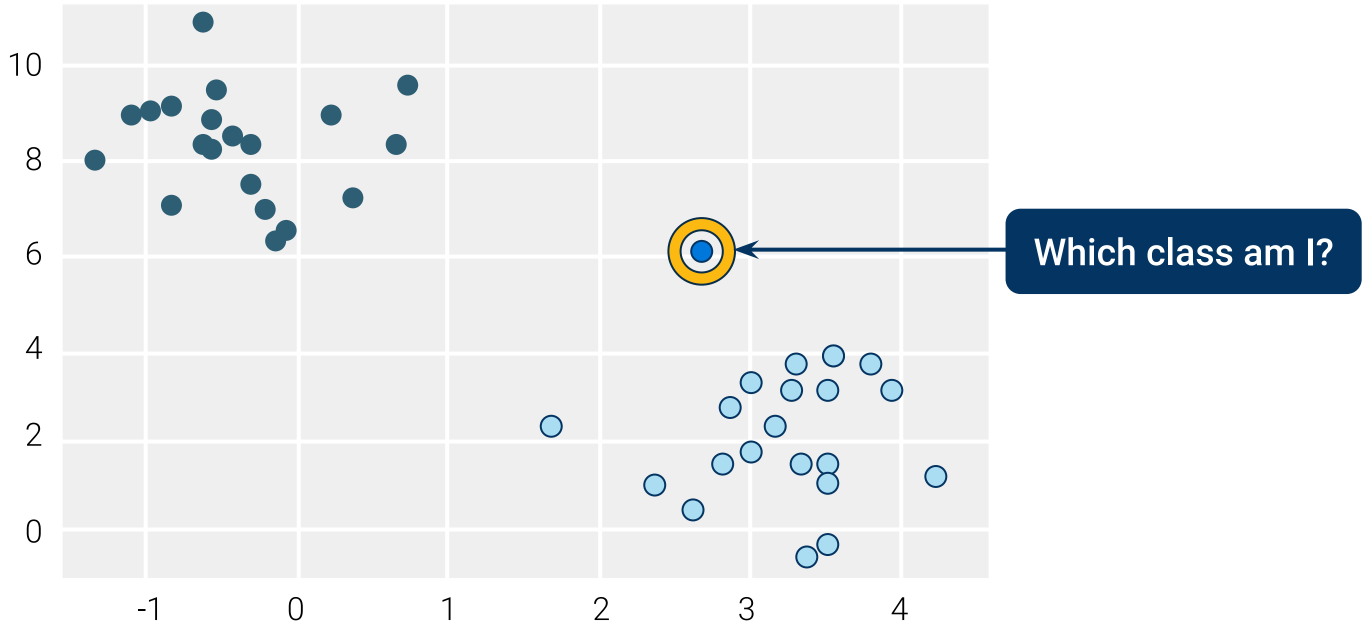
Introduction to SVM

The goal of a **linear classifier** is to find a line that separates two groups of data.

However, many possible lines might exist, with each creating a different boundary.

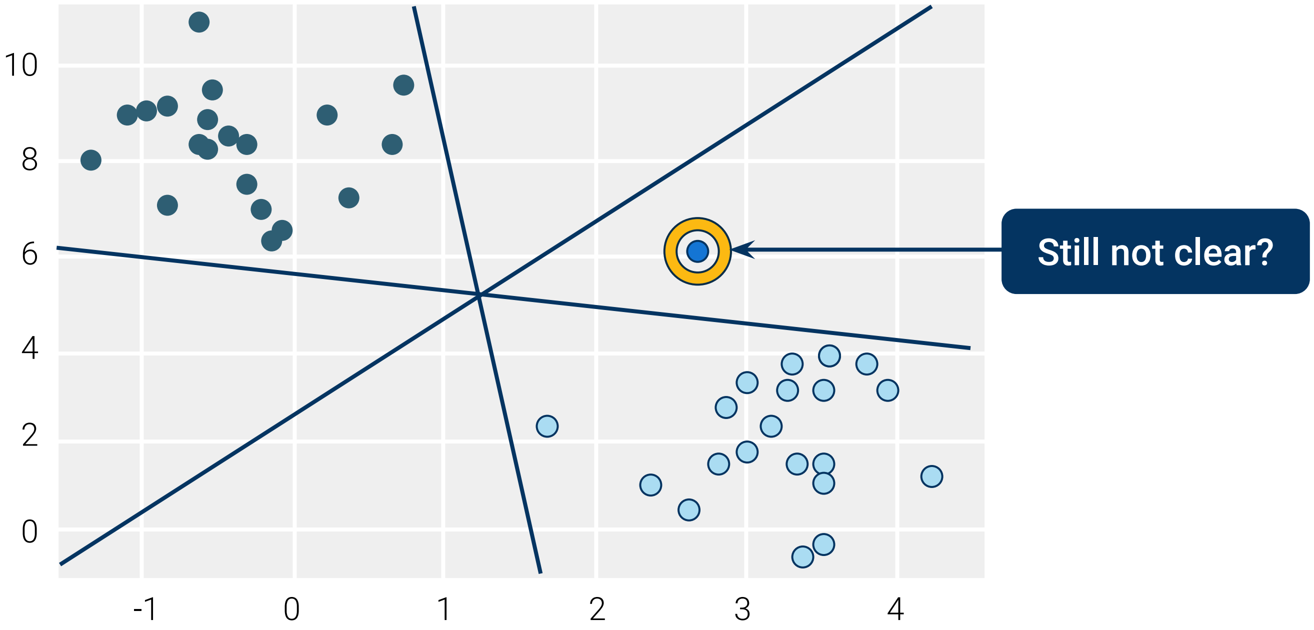
# Support Vector Machines

Choosing a less accurate line thus might result in the misclassification of new data.
The following image illustrates this problem:



# Support Vector Machines

The following image shows three possible lines that each separate the two existing groups of data.
A new data point can fall into either group, depending on the line:



The **support vector machines algorithm** tries to find a hyperplane that maximizes the boundaries between groups. Hyperplanes are decision boundaries in data spaces that are “flat.”

# Support Vector Machines

Sklearn follows a common pattern of model-fit-predict, which allows machine learning engineers to train, test, and evaluate a variety of machine learning models.

To illustrate this, we will look at a new model called an **SVM**.

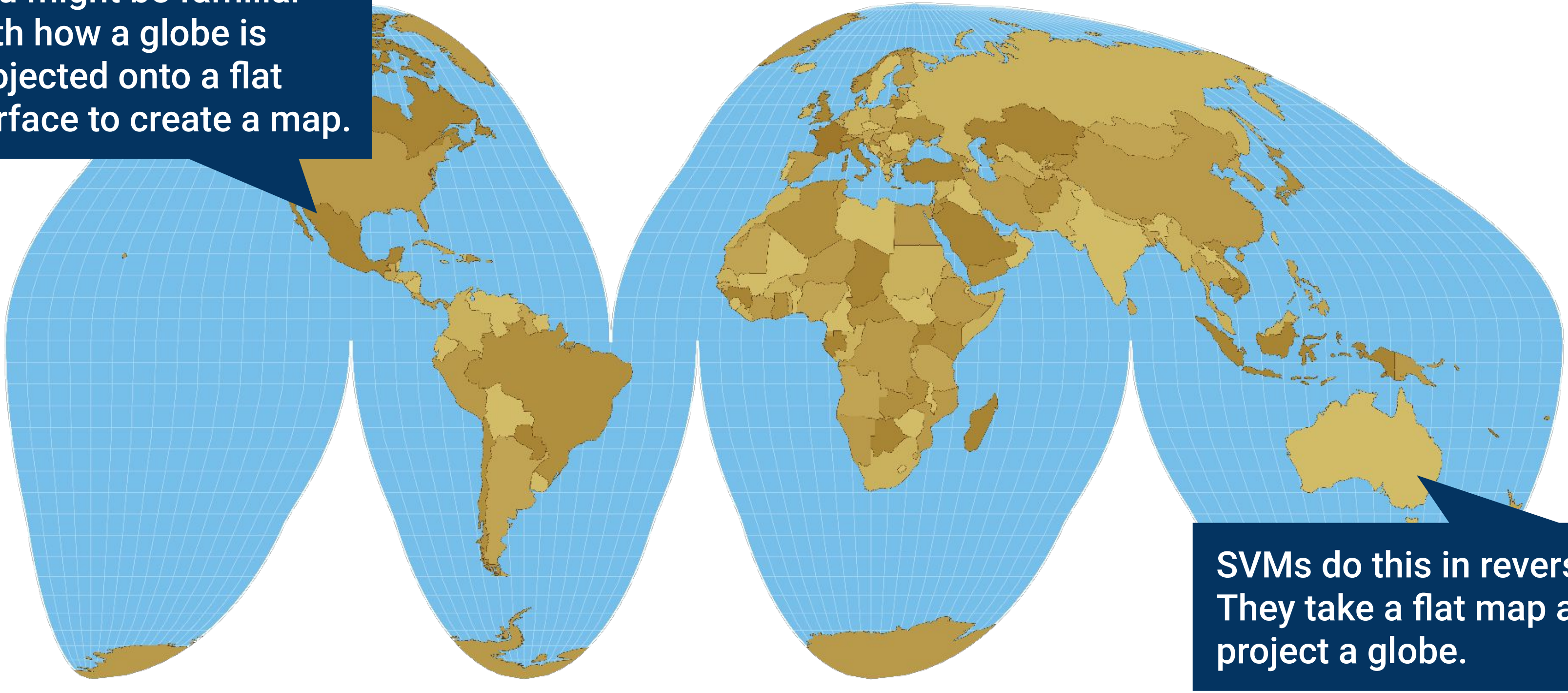
**SVM** is a supervised learning model that we can use for classification and regression analysis. SVM separates classes of data points into multidimensional space.

# Support Vector Machines

SVMs are a widely applied model in data science, especially for assessing credit risk and fraud detection.

# Support Vector Machines

The idea behind SVMs is that a dataset and its labels are projected into a higher dimensional space.

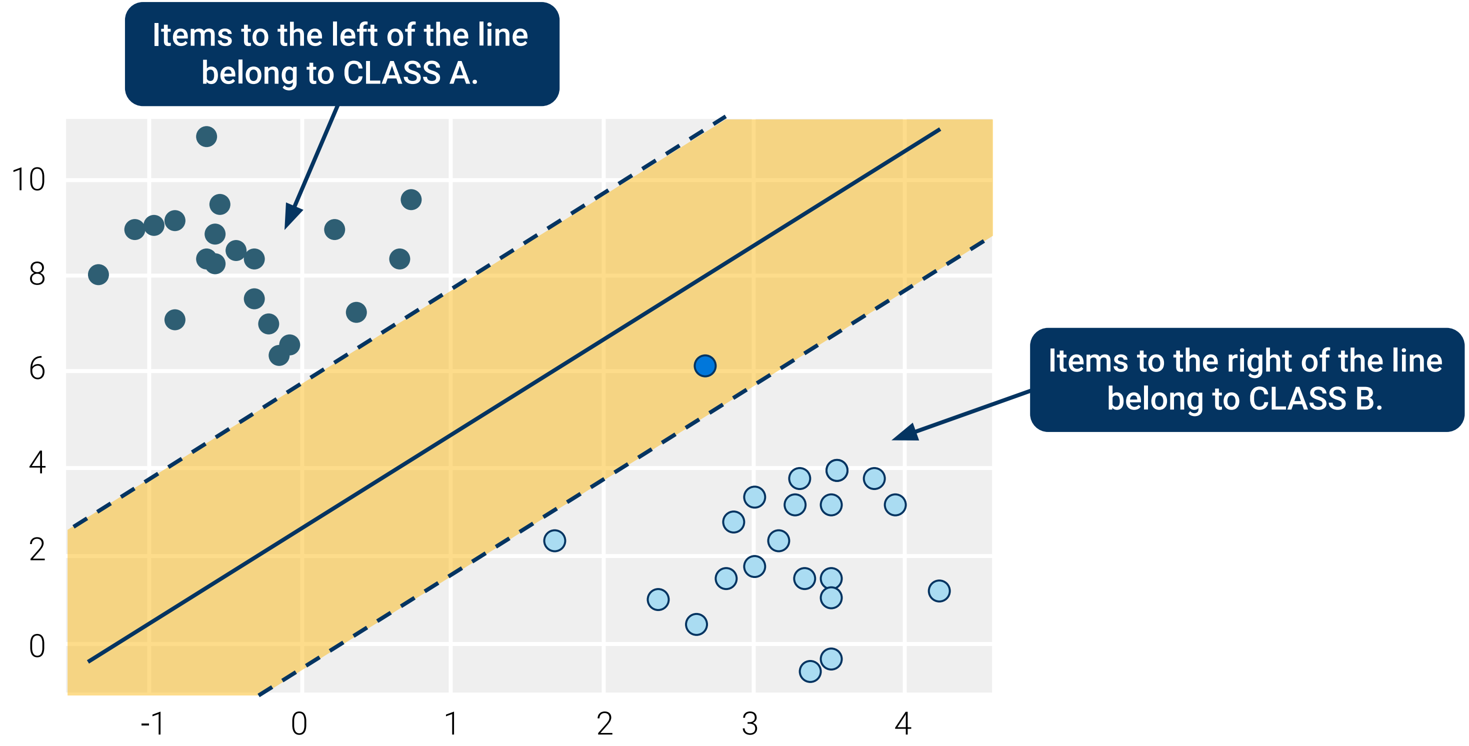


You might be familiar with how a globe is projected onto a flat surface to create a map.

SVMs do this in reverse. They take a flat map and project a globe.

# Support Vector Machines

This boundary's projection is called a **hyperplane**, and we can use it to classify the points.

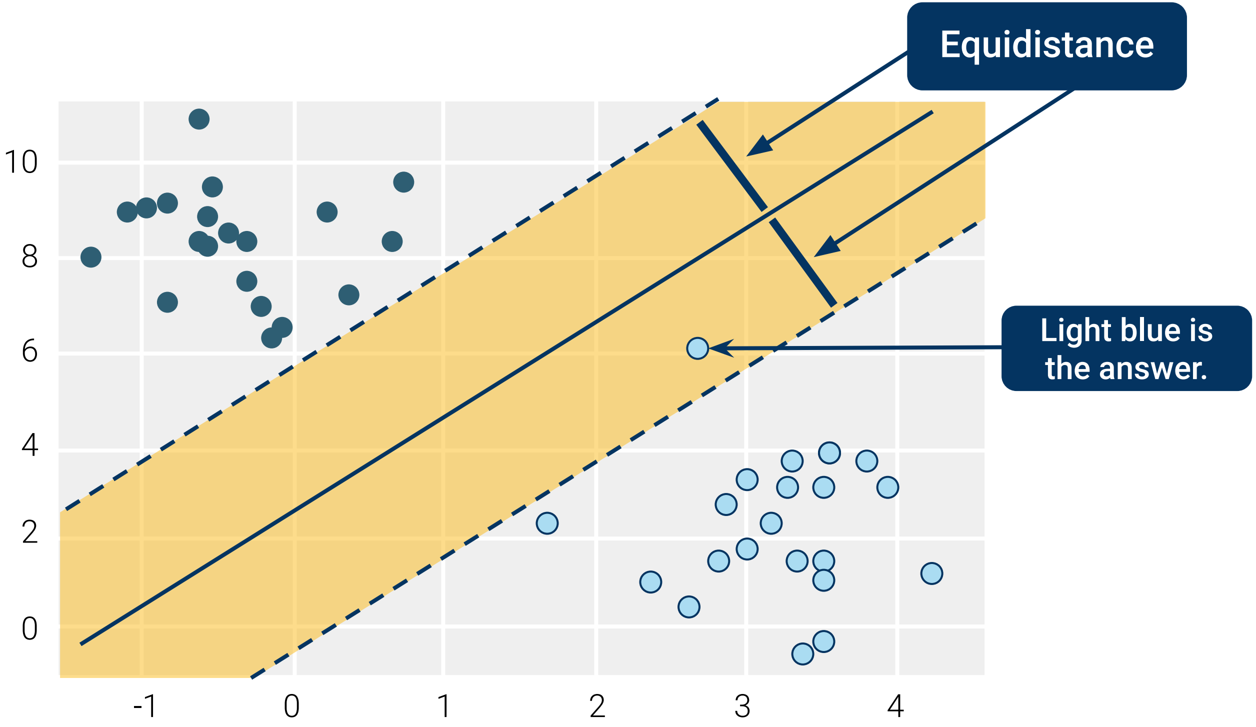


Items to the left of the line belong to CLASS A.

Items to the right of the line belong to CLASS B.

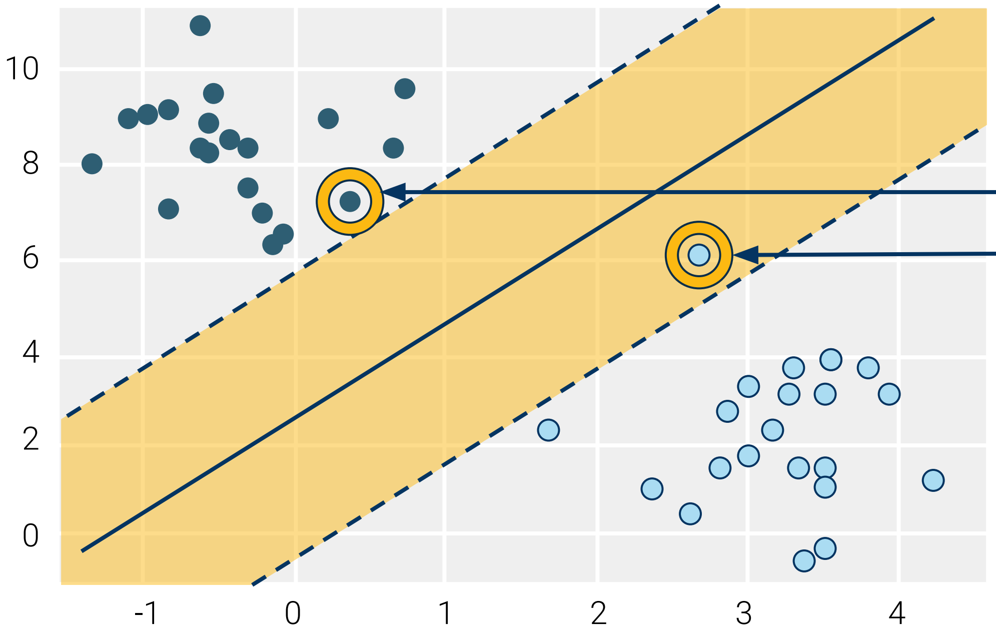
# Support Vector Machines

The goal with hyperplanes is to get the margin of the hyperplane equidistant to the data points for all classes.



# Support Vector Machines

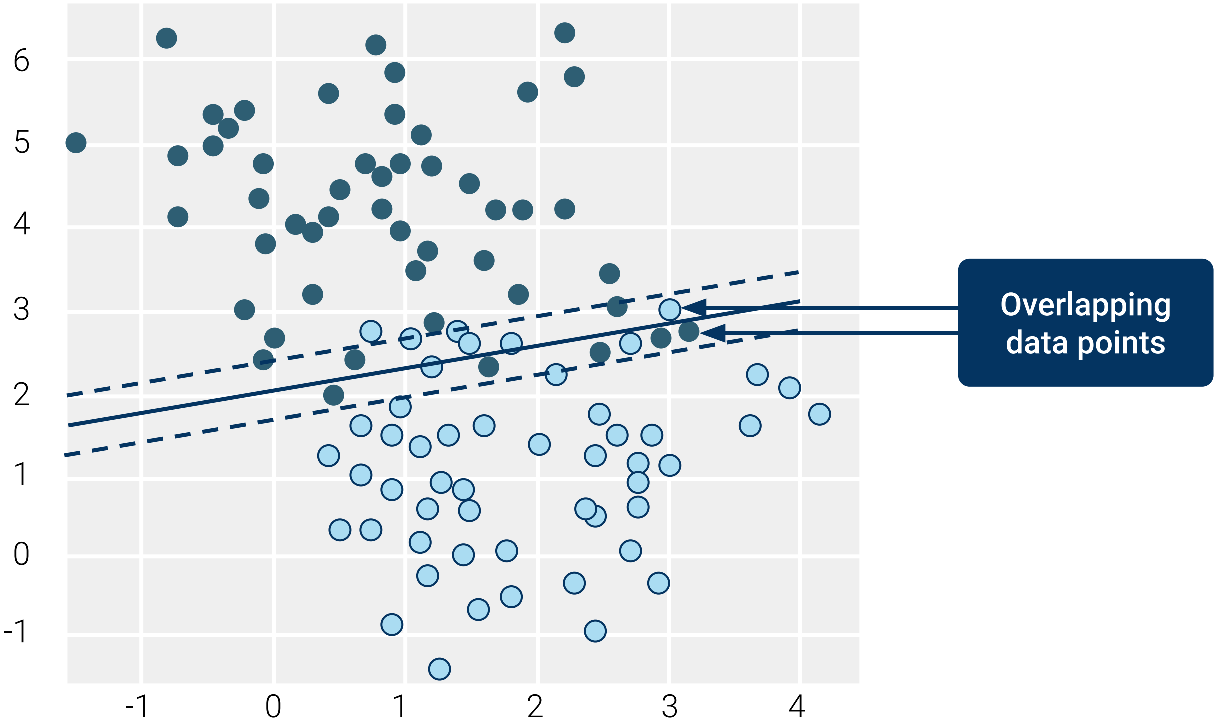
The data closest to/within the margin of the hyperplane are called support vectors. They define the boundaries of the hyperplane.



These points (support vectors) are sometimes the most difficult to classify because they live closest to the margin and could belong to either class.

# Support Vector Machine Models

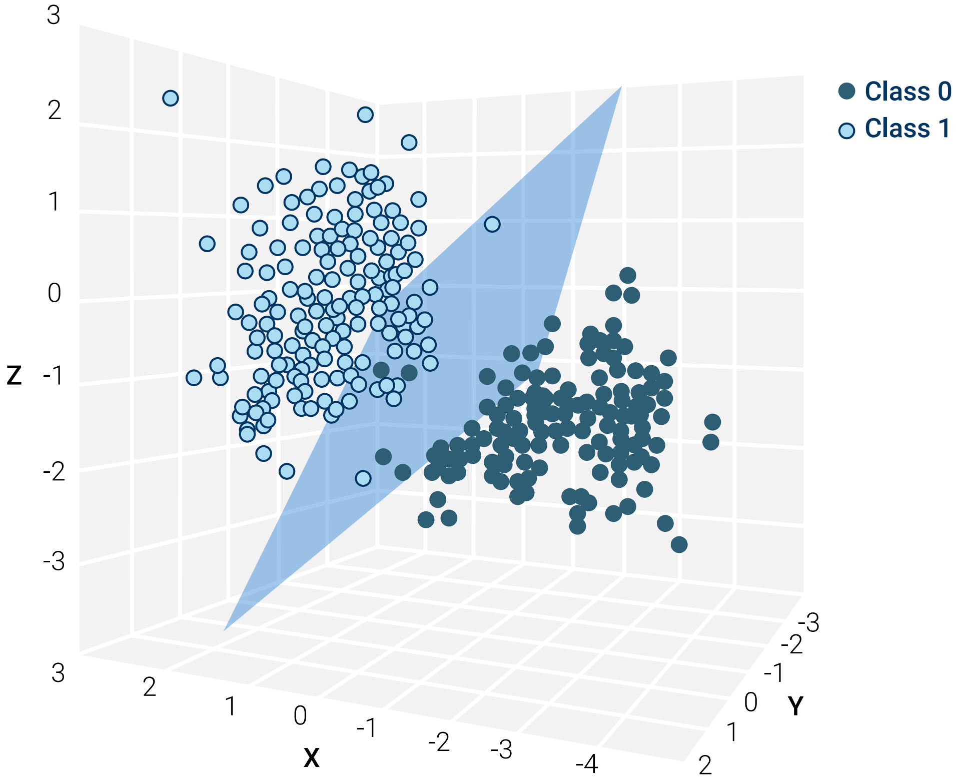
An SVM model with narrow margins and overlapping support vectors:



Overlapping data points

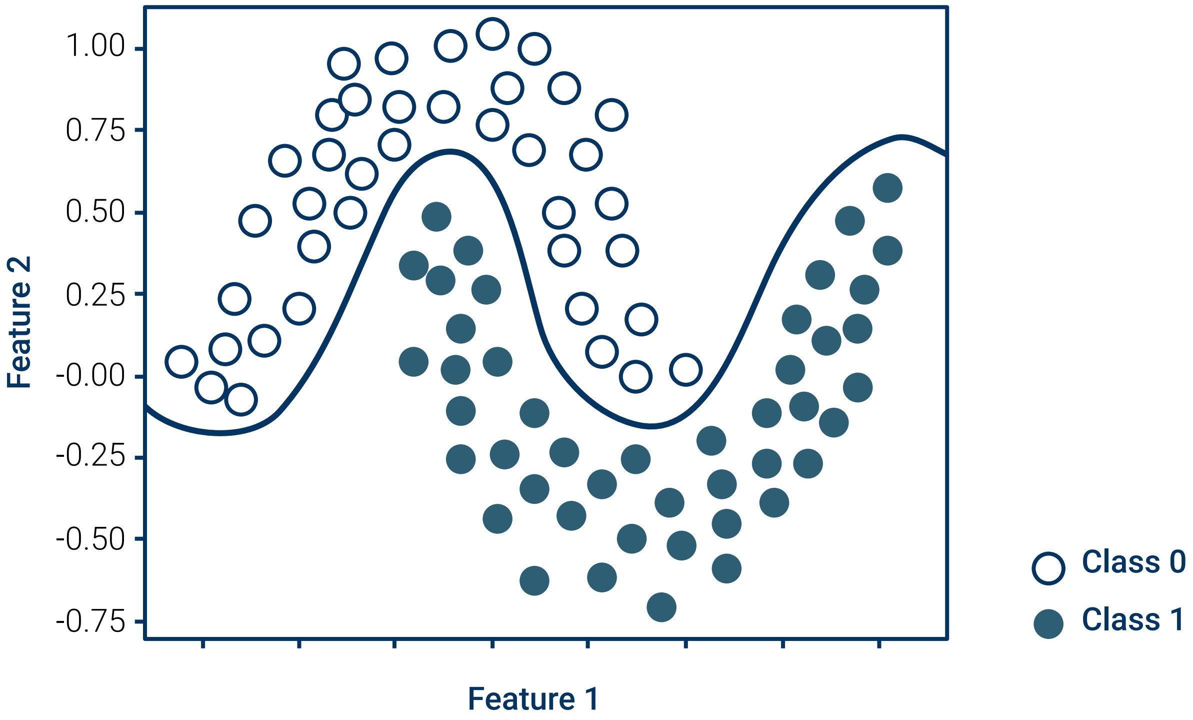
# Support Vector Machine Models

A hyperplane with an x, y, and z-axis:



# Support Vector Machine Models

A non-linear hyperplane viewed in two dimensions as a line:



# Instructor **Demonstration**

SVM Demo

# Activity:

## Predicting Malware with SVM

In this activity, you will use SVM to predict malware presence based on given data, and compare the accuracy score with the accuracy of the logistic regression model that used the same data.

**Suggested Time:**
15 Minutes

# Time’s up!

Let’s review

# Questions?

# Let’s **recap**

# Review the Class Objective

In this lesson you learned how to:

1 Understand and explain the principles behind linear classification models.

2 Explain how the logistic regression model works as a binary classifier.

3 Explain how the SVM model works as a binary classifier.

4 Implement logistic regression and SVM, and assess their performance on sample datasets.

5 Compare and contrast different data scaling methods.

# Next

In the next lesson, you will…

# Questions?

# The End