Non-linear Support Vector Machines Explained

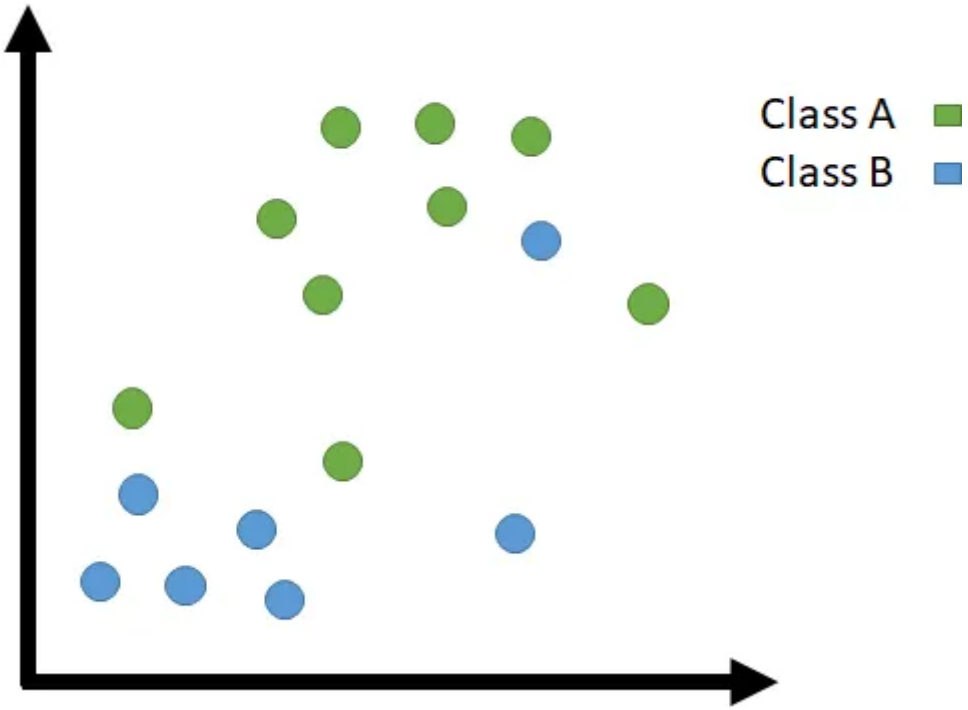
With video explanation | Data Series | Episode 9.3

Mazen Ahmed · Follow
5 min read · Sep 2, 2021

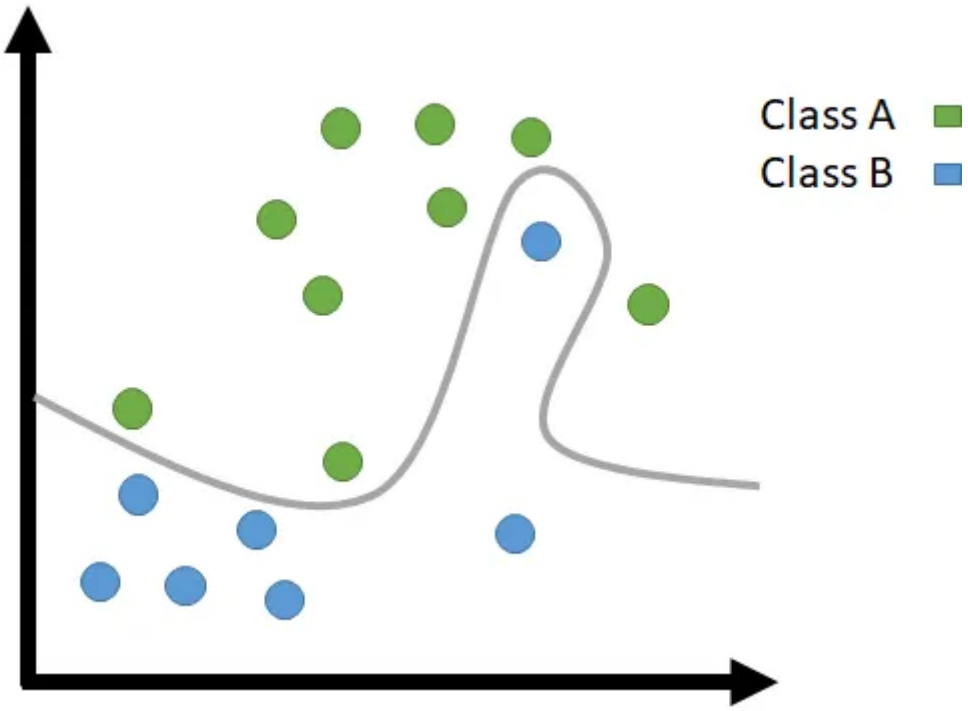
In the **previous episode** we explained what are support vector machines and the maths behind the algorithm. In this episode we discuss support vector machines for non-linearly separable data.

SVMs for non-linearly separable data

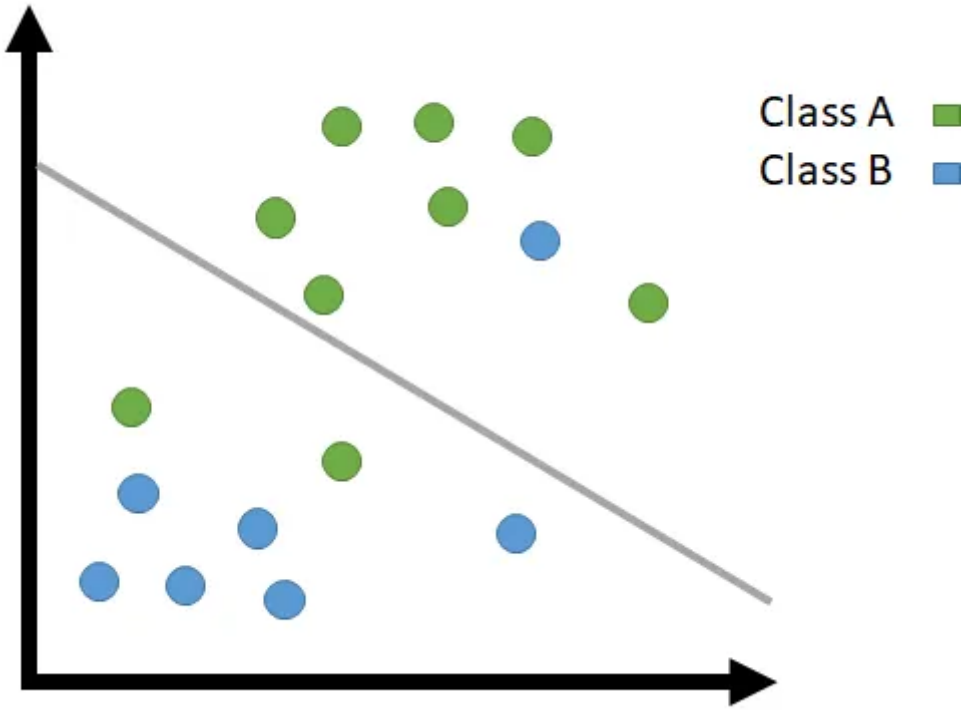
What if the data is **not linearly separable?** For example:



Calculating a **non-linear support vector machine** may overfit our data:



In this case we may still want a linear support vector machine but allow it to **make some mistakes.** This is known as a **soft-margin** support vector machine.



To do so we make some changes to the SVM **objective function** defined in the **previous episode** as :

$$\min \frac{1}{2}\|w\|^2$$

$$\text{Such that: } (w^T x_i + b) y_i \geq 1$$

$$\text{for } i = 1, \ldots, m$$

- Where **w** are the **coefficients** of the hyperplane or decision boundary, **b** the intercept and **m** in the number of training examples.
- x_i gives the position of data point i and y_i the label of data point i (class A or class B) 0 or 1.

To allow for misclassifications we make the following adjustment:

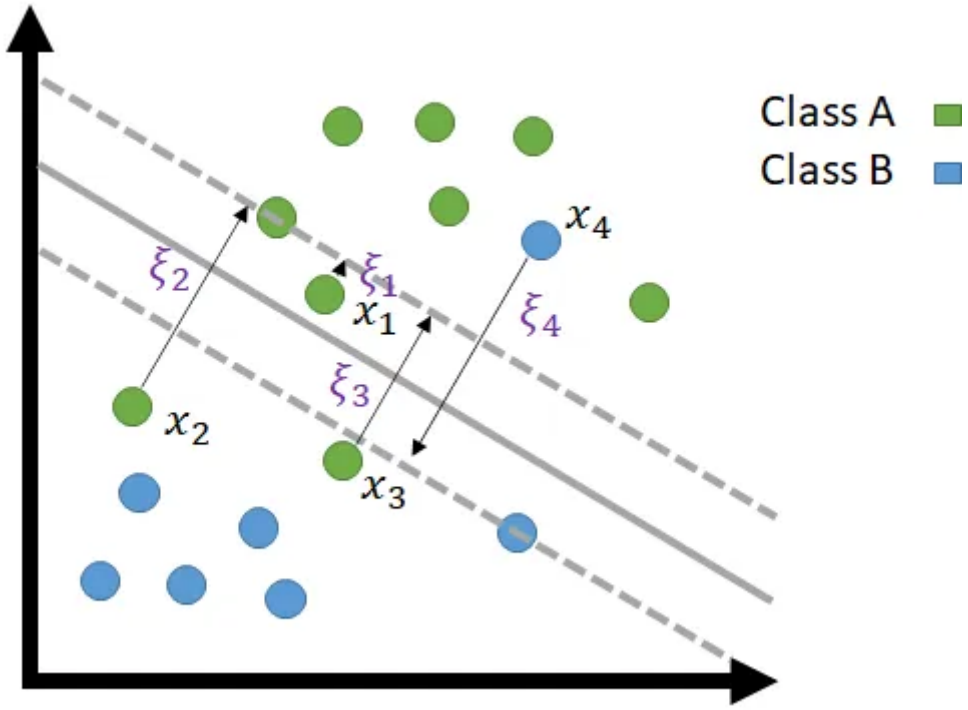
$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{Such that: } (w^T x_i + b)y_i \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$\text{for } i = 1, \ldots, m$$

- ξ_i (xi) is known as the slack variable or penalty. It gives the distance misclassified points are away from their classes margin:

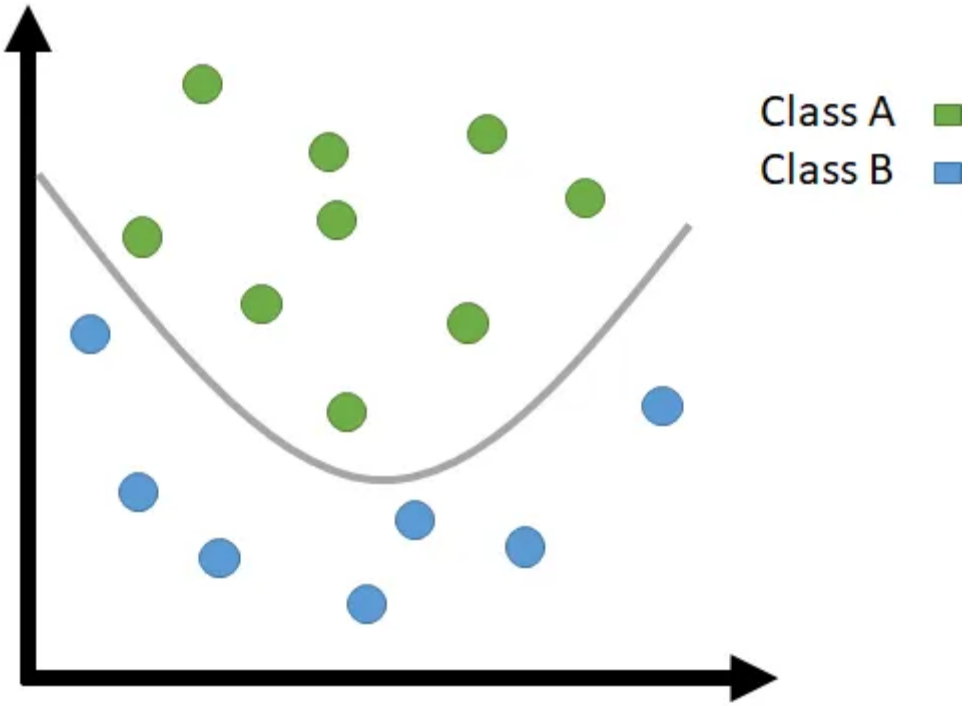


- C is a regularization parameter that controls the **strength** of the slack variable ξ.
- Since we are multiplying ξ with C, with lower values of C there are lower penalties to misclassified data points. The objective function would put less emphasis on minimising the penalties of misclassified points . The SVM won't try as hard to separate the data and therefore produce a more generalisable model which **reduces overfitting.**

- With larger values of C, there are bigger penalties to misclassified data points and therefore the SVM would try hard not to make mistakes and sperate the data more in order to minimise the objective function. This may lead to overfitting shown in the non-linear SVM above.

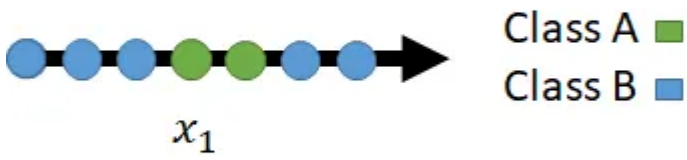
Non-linear Support Vector Machines

What if our data is better separated by a non-linear function. For example a polynomial function shown below:



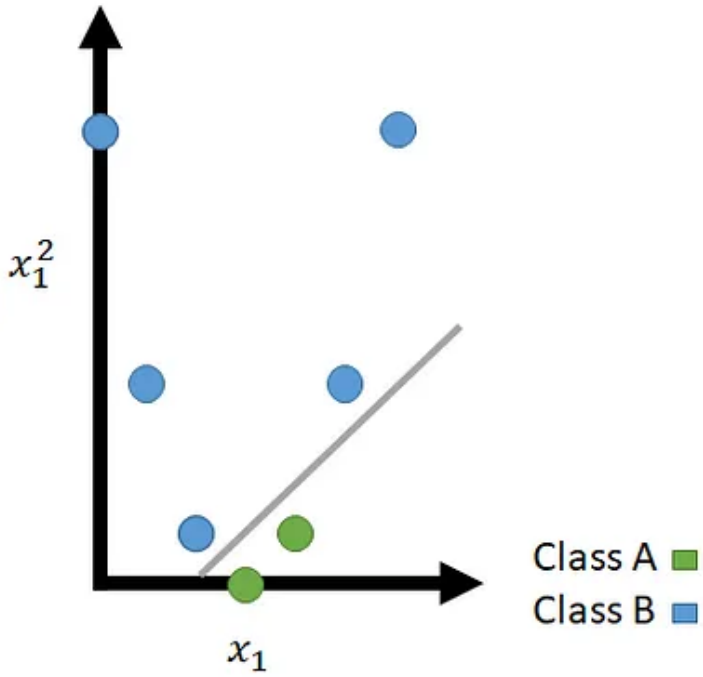
To produce a non-linear support vector machine we make use of what is called a **kernel function**. Depending on the kernel used, the kernel function transforms our data to a feature space where the data becomes more likely to be linearly separable. This is known as the 'kernel trick'.

To illustrate how this works take a look at the following 1 dimensional data the cannot be separated by a single line.



x_1	Class
-3	B
-2	B
-1	B
0	A
1	A
2	B
3	B

Using a polynomial kernel, we can project the data to a 2 dimensional space where we square each point.



x_1	x_1^2	Class
-3	9	B
-2	4	B
-1	1	B
0	0	A
1	1	A
2	4	B
3	9	B

We can see from above that the data has now become linearly separable.

For the above example we made use of the polynomial kernel, this is commonly used when the data can be separated by some polynomial function.

Kernels

The dot product.

$$x_1 = \begin{pmatrix} 1 \\ 2 \\ 7 \end{pmatrix} \quad x_2 = \begin{pmatrix} 3 \\ 2 \\ 4 \end{pmatrix}$$

$$x_1 \cdot x_2 = 1 \times 3 + 2 \times 2 + 7 \times 4 = 35$$

1. **Linear:** used for linearly separable data to speed up the calculation of the SVM classifier.

$$k(x_i, x_j) = x_i \cdot x_j$$

x_i and x_j are given as two single observations. In some cases we may just have one feature, for this case the kernel just becomes $k(x_i) = x_i$.

2. **Polynomial:** used for data that can be separated by a polynomial function with kernel coefficient γ, independent term r and degree d.

$$k(x_i, x_j) = (\gamma(x_i \cdot x_j) + r)^d$$

We applied this kernel to the above example with γ= 1, r= 0 and d=2. Since we had only one feature x_1 — this kernel essentially squared each observation.

3. **Gaussian radial basis function (RBF)** : this kernel works for many cases and is used when there is no prior knowledge about the data.

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

$$\text{For } \gamma > 0$$

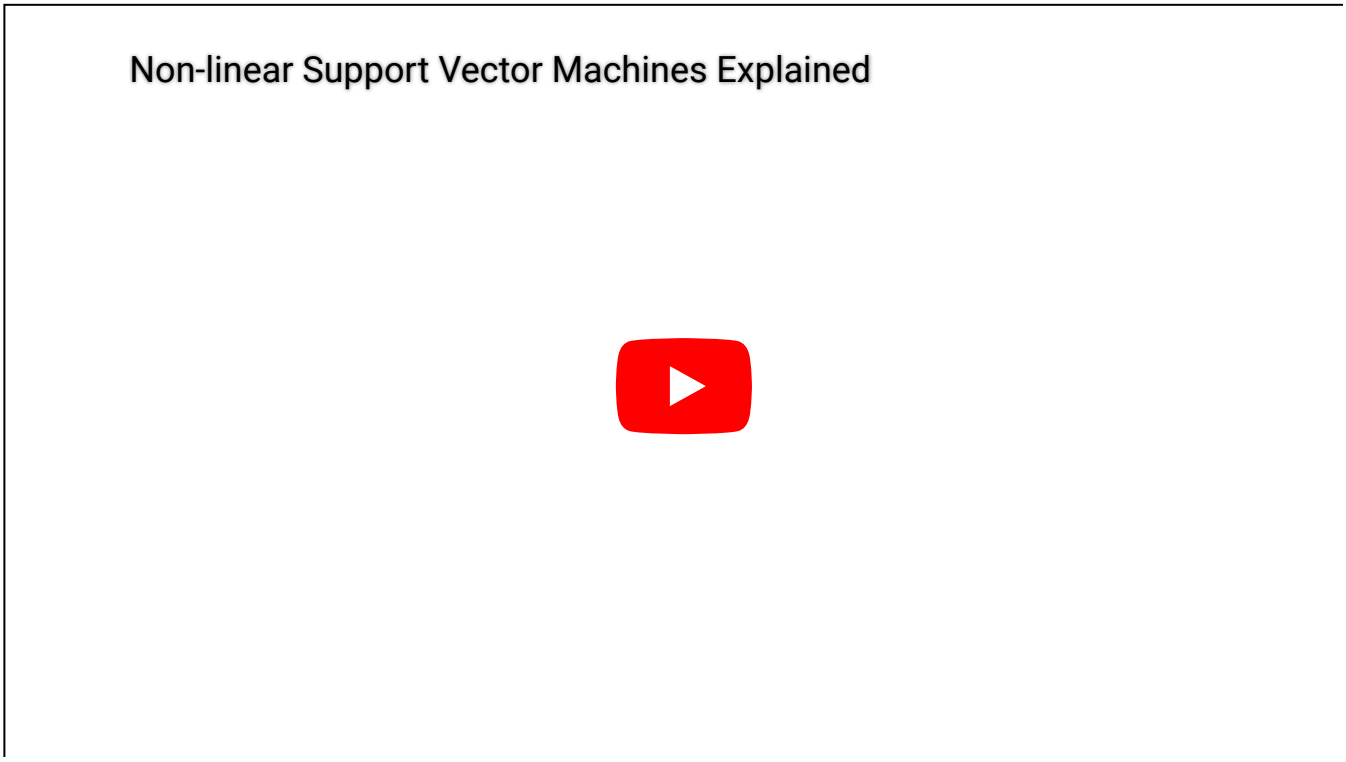
Summary

- We can separate non-linearly separable data without overfitting using a soft-margin support vector machine. For this we introduce the slack term ξ to the objective function.

- To produce non-linear support vector machines we make use of the kernel function which maps our data to a feature space where it becomes more likely to be linearly separable.
- **The Gaussian radial basis function** (RBF) kernel works for many cases and is used when there is no prior knowledge about the data.

Prev Episode | Next Episode

If you have any questions please leave them below!



Machine Learning AI Svm Data Science

Follow

Written by Mazen Ahmed

205 Followers

PhD Student in Bioinformatics that has produced an on-going series of data science tutorials with Python examples.

More from Mazen Ahmed

Mazen Ahmed

Ways to Evaluate your Regression Model

The different methods | Data Series | Episode 10.1

· 5 min read · Oct 6, 2021

9

Mazen Ahmed

Linear Support Vector Machines Explained

With video explanation | Data Series | Episode 9.1

· 5 min read · Feb 11, 2021

4

Mazen Ahmed

Evaluating your Regression Model in Python

Step-by-step follow along | Data Series | Episode 10.2

✦ · 6 min read · Oct 21, 2021

4

Mazen Ahmed

Data Science Project | Clustering Mixed Data

Start to Finish Clustering Analysis | Data Series | Project 3

✦ · 8 min read · Dec 29, 2021

64 1

See all from Mazen Ahmed

Recommended from Medium

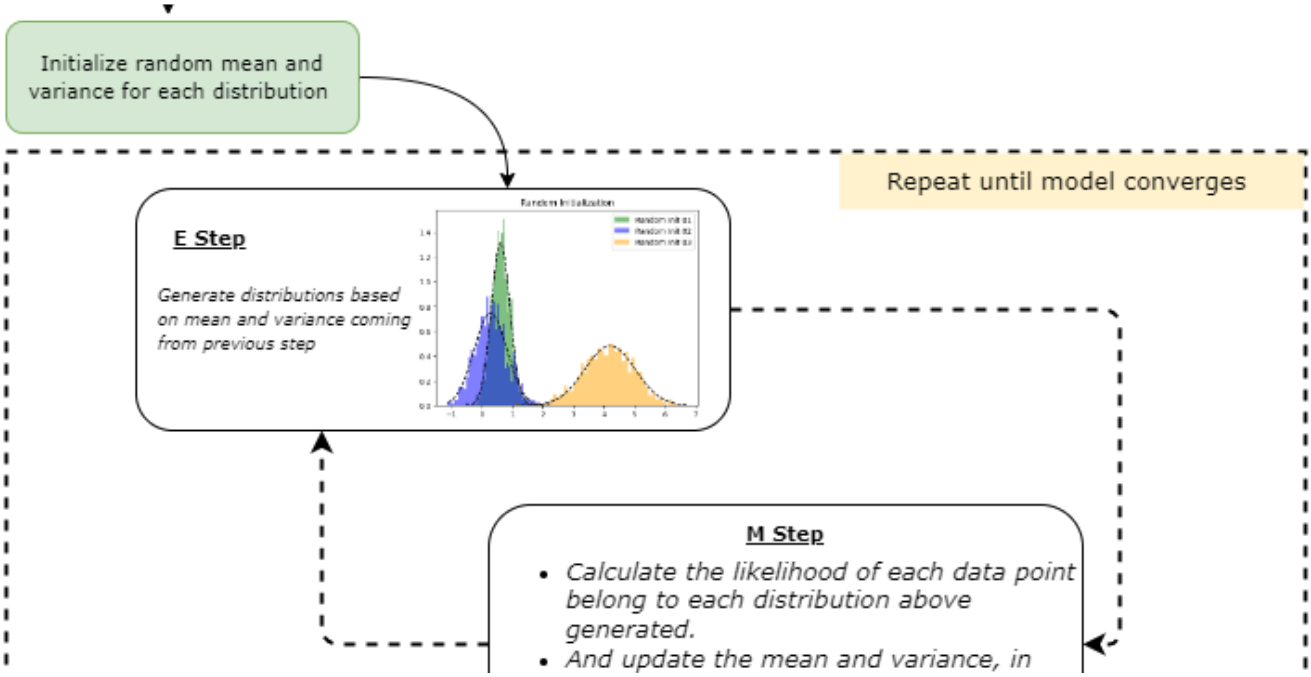
Aziz Budiman

Machine Learning in 4 Minutes: Support Vector Machines

In part 2 on this mini-series of #4MinutesML, we will uncover the mysteries behind the use of Support Vector Machines (SVM) and its...

4 min read · Jun 3

19



Ransaka Ravihara in Towards Data Science

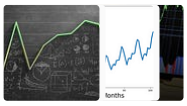
Gaussian Mixture Model Clearly Explained

The only guide you need to learn everything about GMM

9 min read · Jan 10

41 4

Lists



Predictive Modeling w/ Python

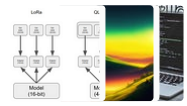
18 stories · 57 saves

Practical Guides to Machine Learning

10 stories · 71 saves

The New Chatbots: ChatGPT, Bard, and Beyond

13 stories · 29 saves



Natural Language Processing

369 stories · 21 saves

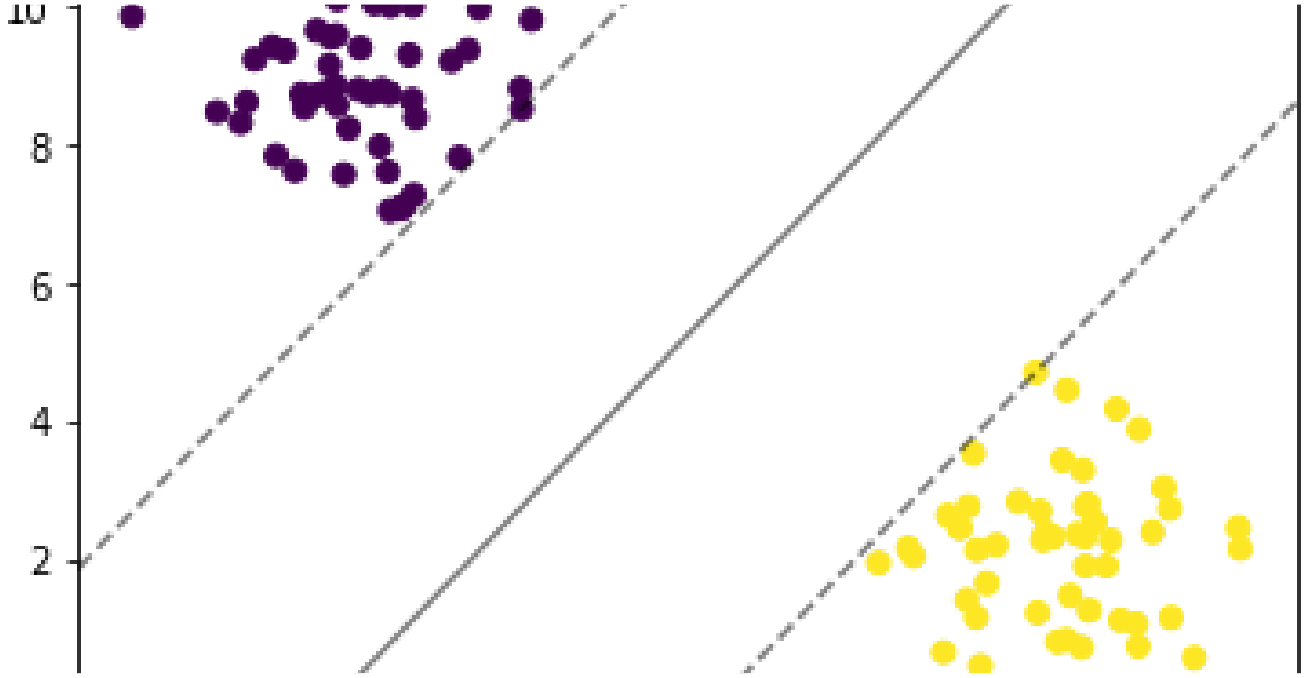
Zeki

Practical Example of Clustering and Radial Basis Functions (RBF)

Clustering is a technique used in machine learning and data analysis to group similar data points together. The goal of clustering is to...

9 min read · Feb 17

4



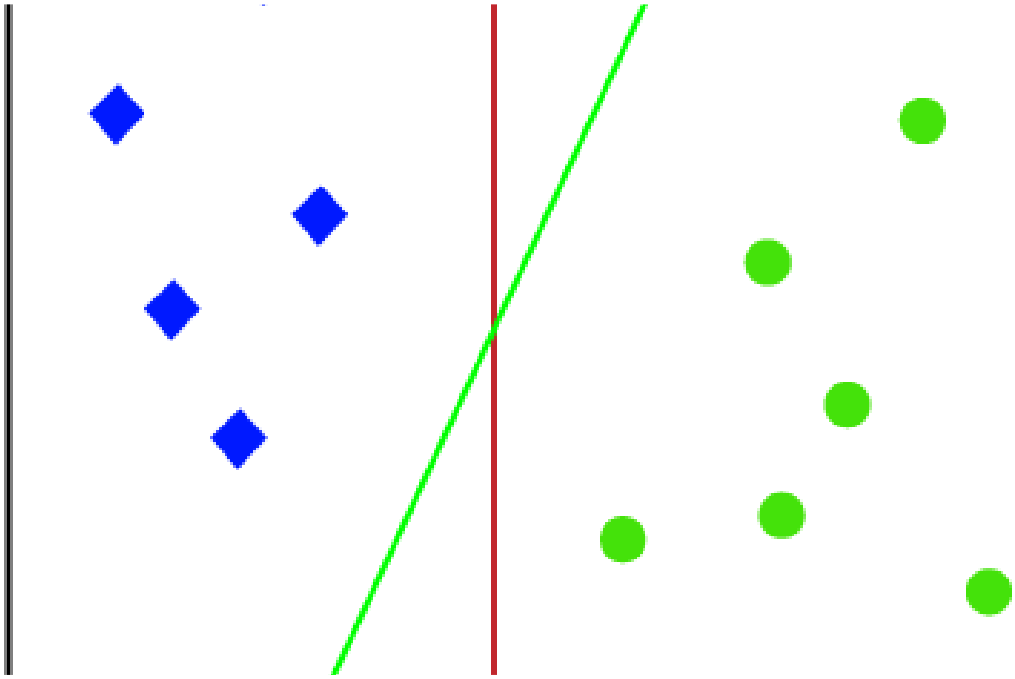
Sharmasaravanan

Introduction to Support Vector Machine (SVM) in Machine Learning

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for classification and regression analysis. SVM works...

3 min read · Mar 11

11

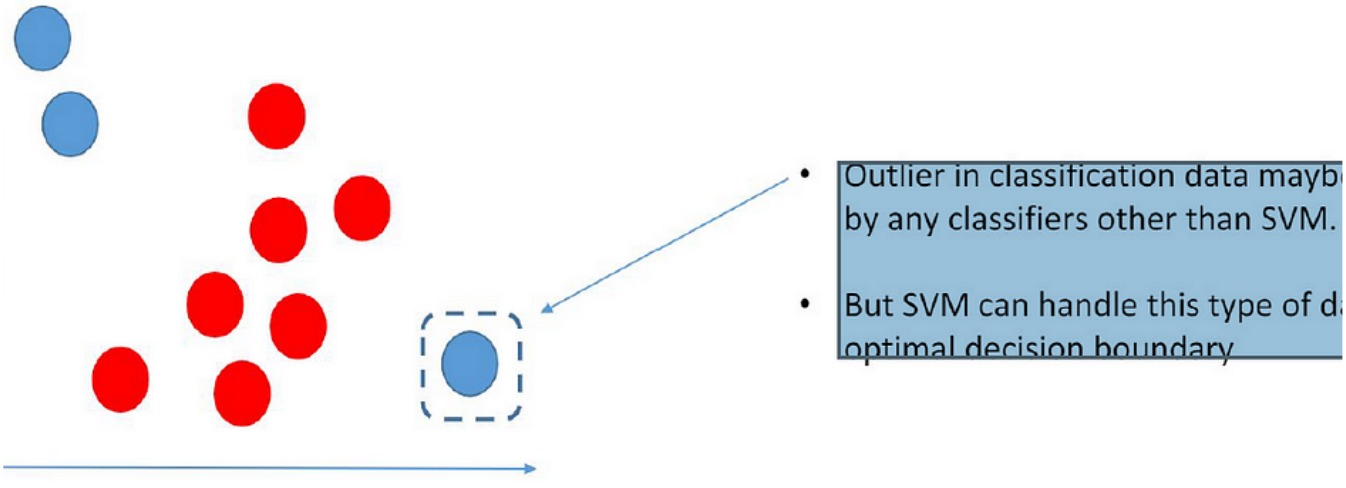


Detheharshada

Support vector machines(SVM)

Support Vector Machine (SVM) is a relatively simple Supervised Machine Learning Algorithm used for classification and/or regression. It is...

4 min read · May 3



YashwanthReddyGoduguchintha

Support Vector Machine Algorithm

SVM:-

5 min read · Apr 5

 1