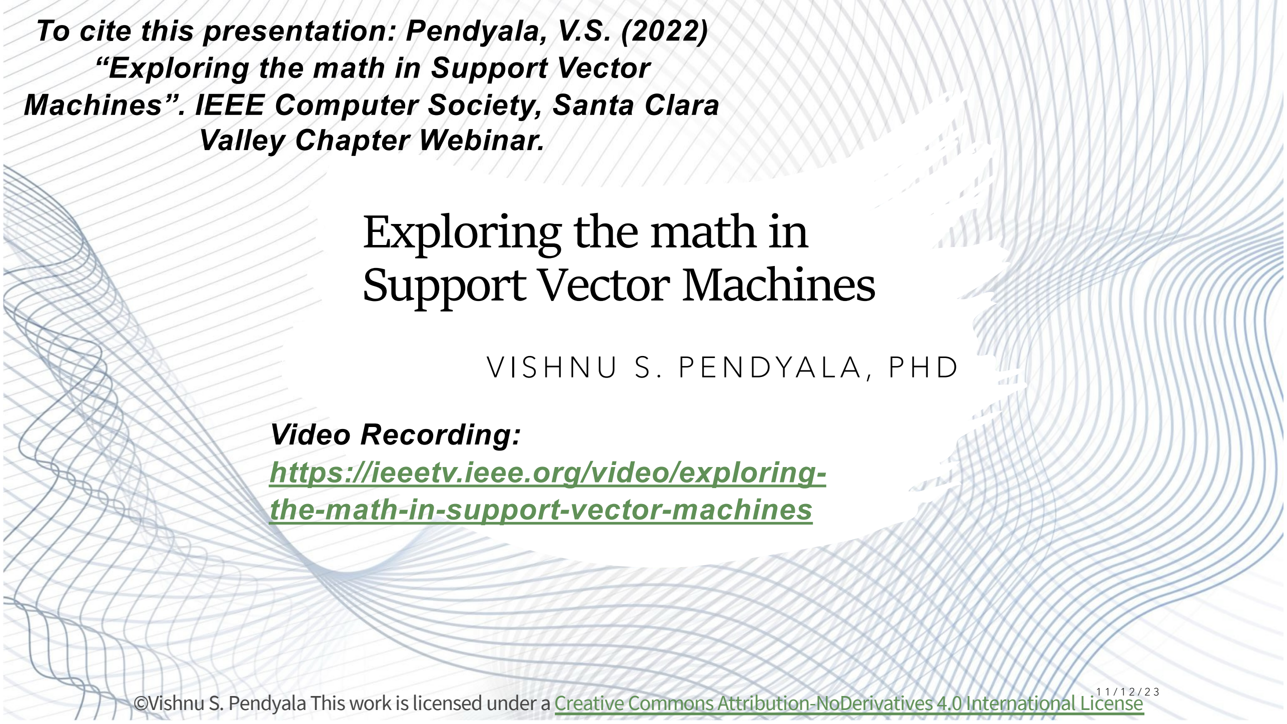
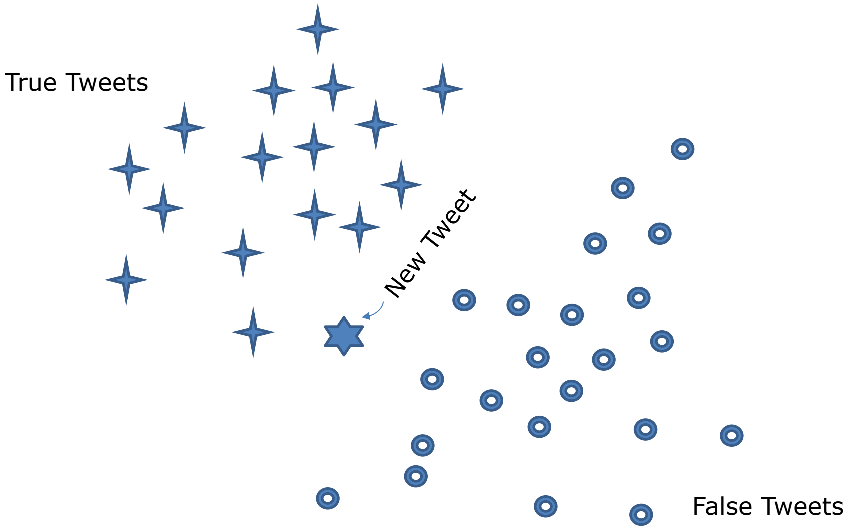


To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

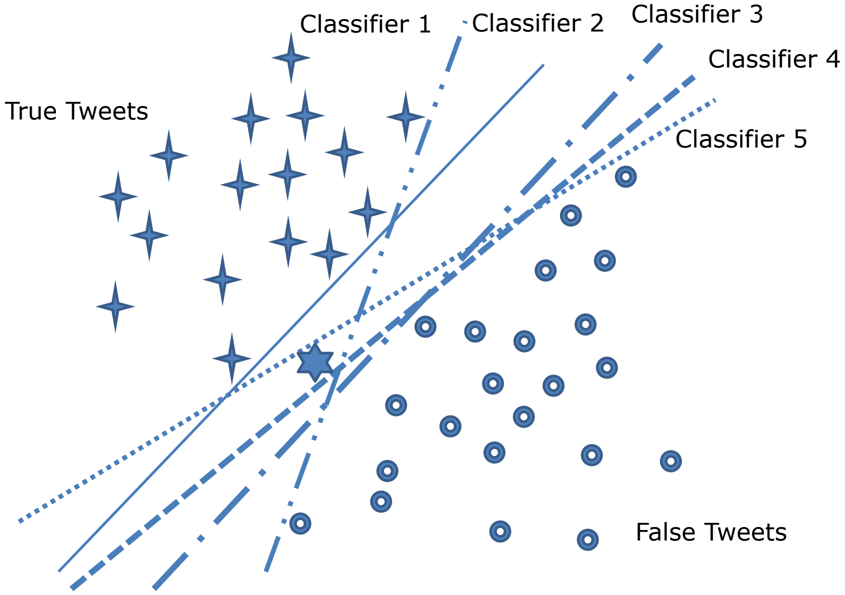
Exploring the math in Support Vector Machines

VISHNU S. PENDYALA, PHD

Video Recording:
https://ieeetv.ieee.org/video/exploring-the-math-in-support-vector-machines

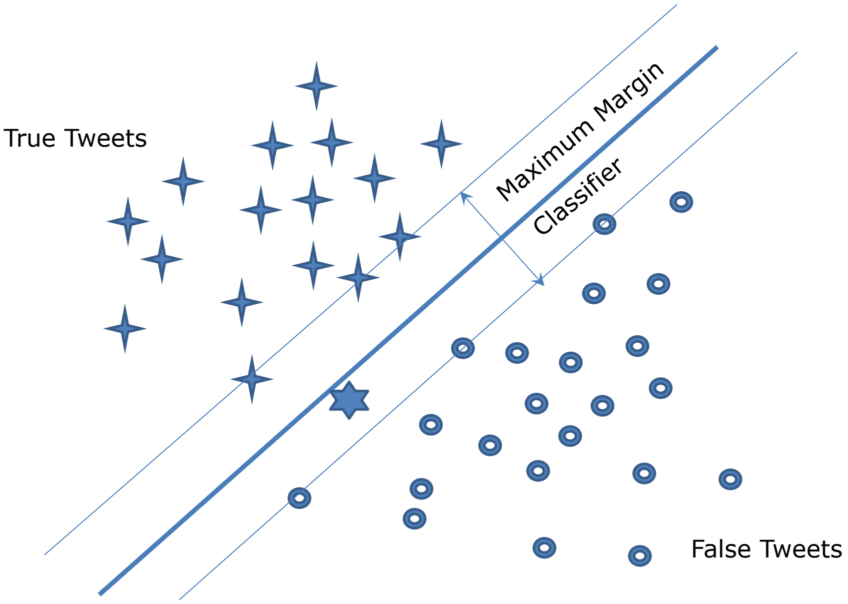


Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.



11/12/23

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

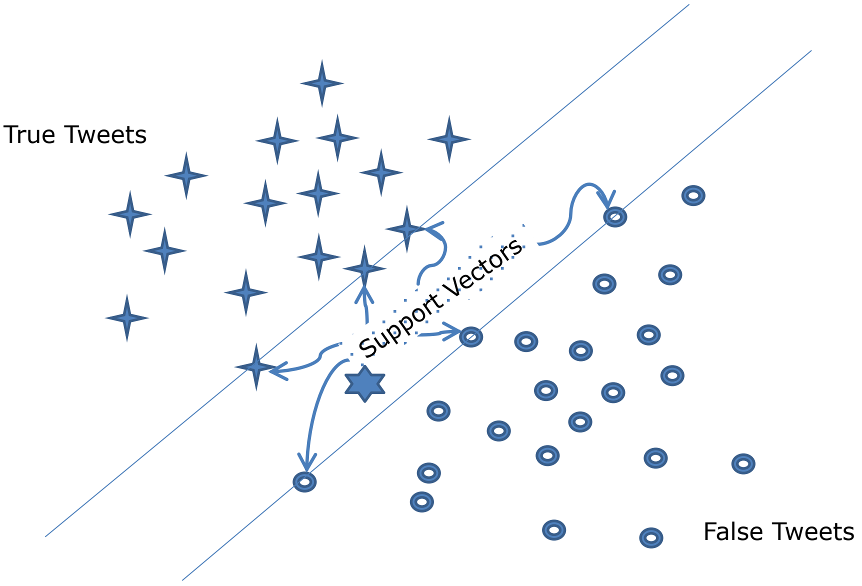


11/12/23

Maximum Margin Classifier

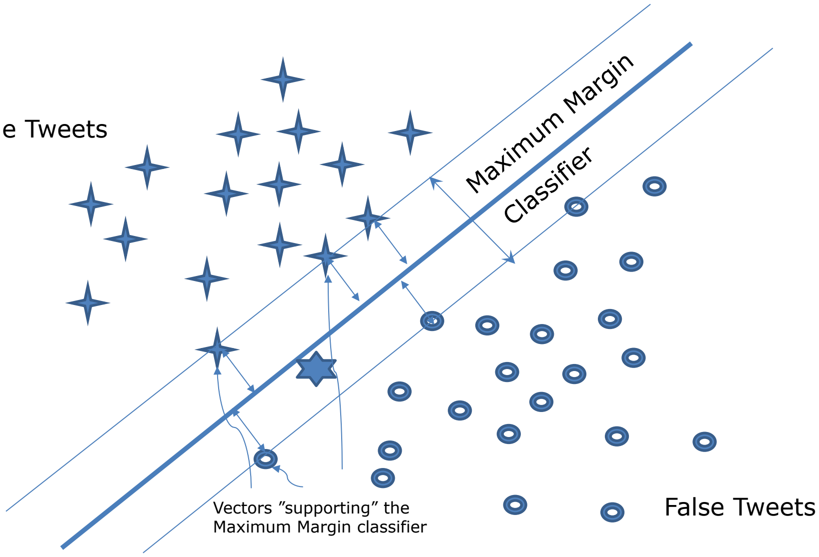
Arbiter

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.



11/12/23

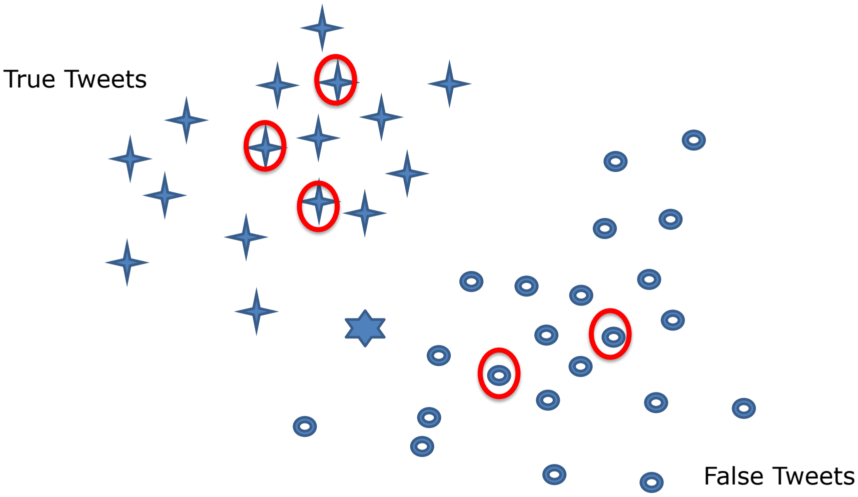
Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.



11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

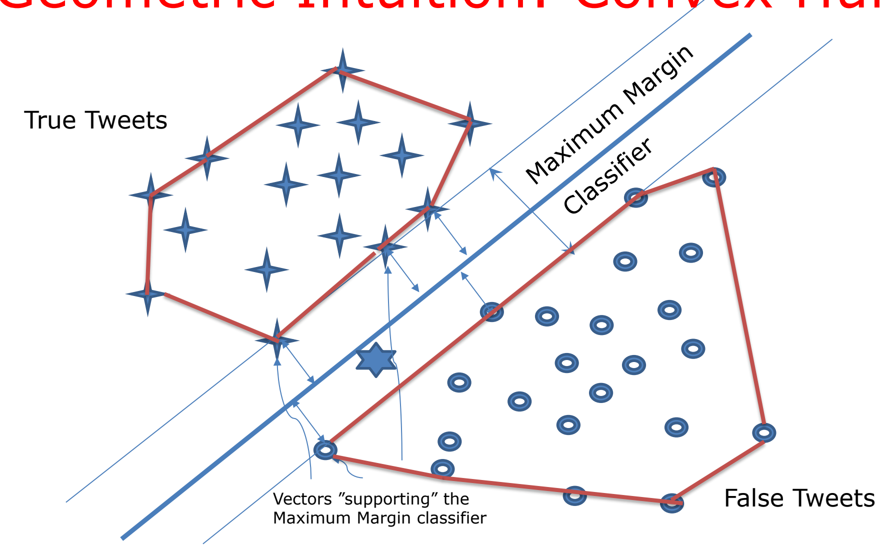
To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.



11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Geometric Intuition: Convex Hulls

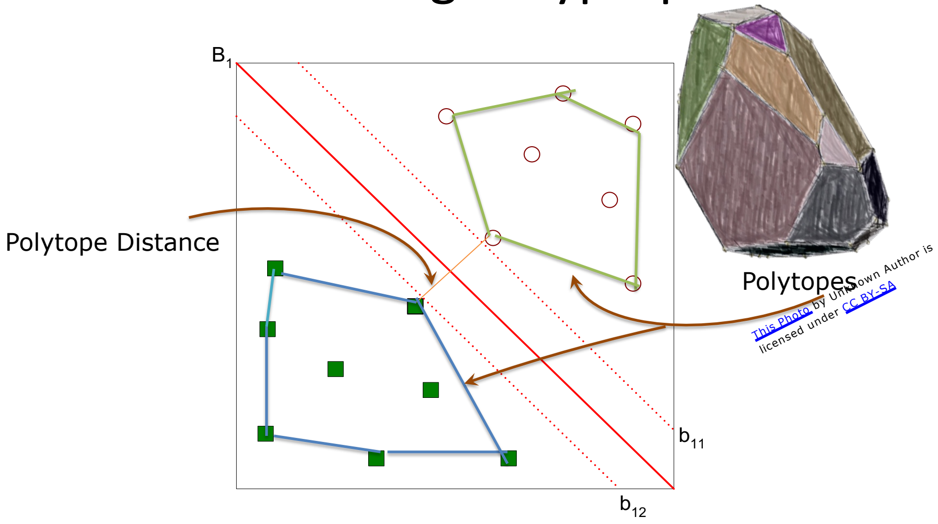


11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Maximum Margin Hyperplane



This Photo by Unknown Author is licensed under CC BY-SA

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

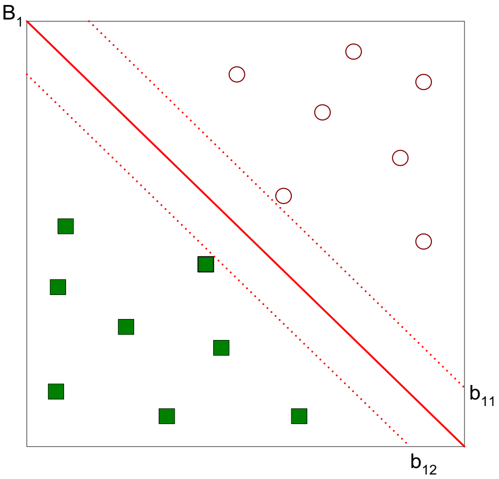
What is the length of the Maximum Margin?

Remember how to compute the distance between two parallel lines?

ax + by + c1 = 0

ax + by + c2 = 0

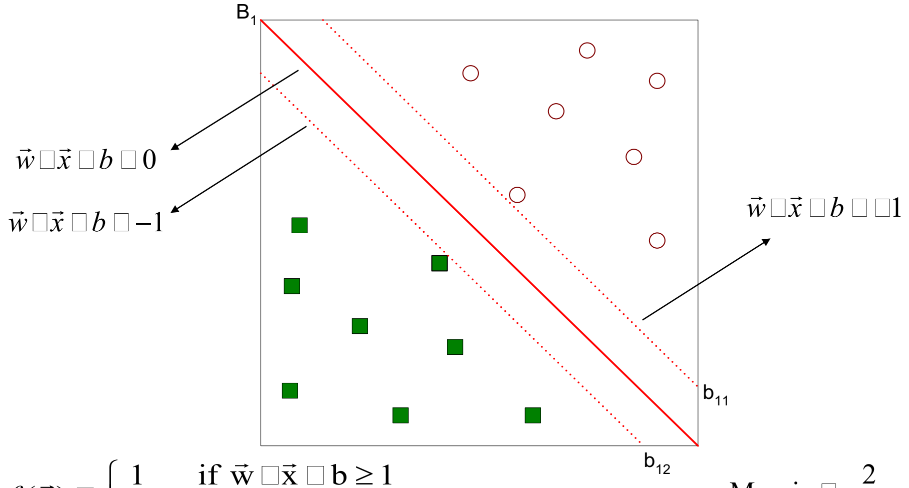
$$d = \frac{|c_2 - c_1|}{\sqrt{a^2 + b^2}}$$



Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Support Vector Classifier


B_1
$\vec{w} \square \vec{x} \square b \square 0$
$\vec{w} \square \vec{x} \square b \square -1$
$\vec{w} \square \vec{x} \square b \square \square 1$
b_{11}
b_{12}

$$f(\vec{x}) \square \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \square \vec{\mathrm{x}} \square \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \square \vec{\mathrm{x}} \square \mathrm{b} \leq -1 \end{cases}$$

$$\text{Margin} \square \frac{2}{\| \vec{w} \|}$$

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Linear SVM Problem Formulation

- Objective is to maximize: $\text{Margin} \, □ \, \frac{2}{\|\vec{w}\|}$

a) Which is equivalent to minimizing: $L(\vec{w}) \, □ \, \frac{\|\vec{w}\|^2}{2}$

b) Subject to the following constraints:

$$y_i \, □ \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \, □ \, \vec{\mathrm{x}}_{\mathrm{i}} \, □ \, \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \, □ \, \vec{\mathrm{x}}_{\mathrm{i}} \, □ \, \mathrm{b} \leq -1 \end{cases}$$

or

$$y_i(\mathbf{w} \, □ \, \mathbf{x}_i \, □ \, b) \geq 1, \quad i \, □ \, 1,2,\ldots,N$$

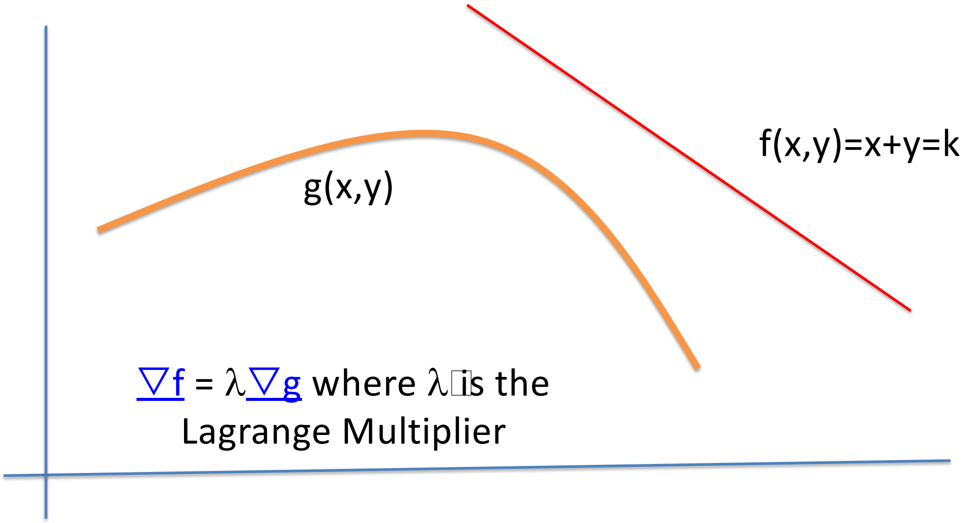
- This is a constrained optimization problem

=> Solve it using Lagrange multiplier method

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

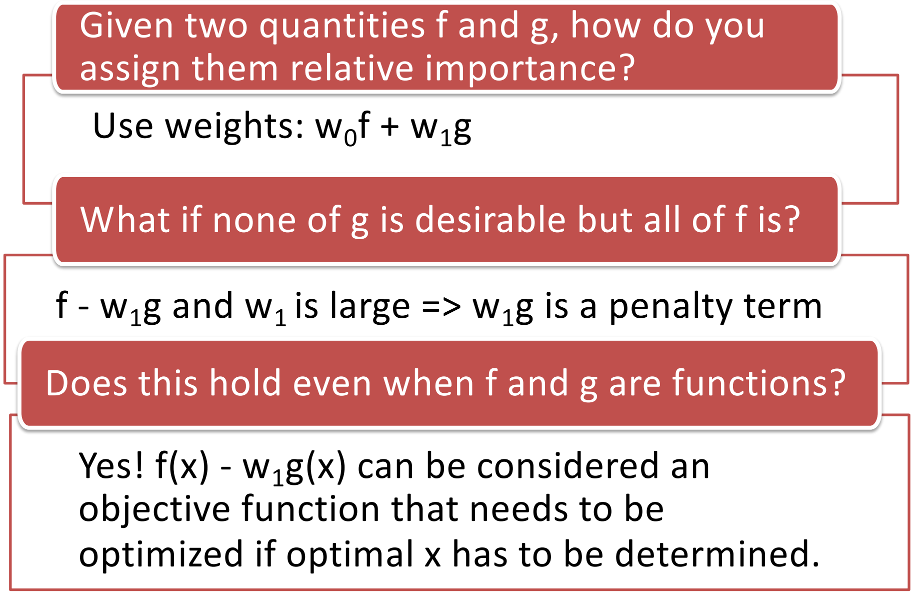
Lagrange Multipliers: Geometric Intuition

Find $\text{argmax}_{x,y} x+y$ such that $g(x,y) = 0$



$\nabla f = \lambda \nabla g$ where λ□is the Lagrange Multiplier

Leading questions



Given two quantities f and g, how do you assign them relative importance?

Use weights: $w_0f + w_1g$

What if none of g is desirable but all of f is?

f - w_1g and w_1 is large => w_1g is a penalty term

Does this hold even when f and g are functions?

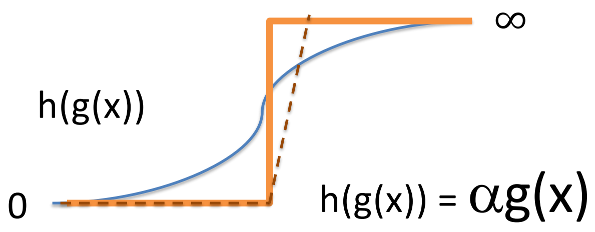
Yes! f(x) - $w_1g(x)$ can be considered an objective function that needs to be optimized if optimal x has to be determined.

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Constrained Optimization

1. In general, if we want to optimize f(x) under a given constraint g(x), we consider
   a) f(x) + h(g(x)) where h(g(x)) = 0 if the constraint is met and infinity if it is not met.
   b) What is a good (smoother) approximation to a step function?



Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Lagrange Multipliers Rephrased

- Find argmax$_{x,y}$ f(x,y) such that g(x,y)=c

  is equivalent to finding (x,y,α) such that

  ∇L = 0 where

  L(x,y,α) = f(x,y) – α(g(x,y)–c)

⇒ $\delta_\alpha L(x,y,\alpha) = -g(x,y) + c = 0$……….(1)

$\delta_x L(x,y,\alpha) = \delta_x f(x,y) - \alpha\delta_x g(x,y) = 0$….(2)

$\delta_y L(x,y,\alpha) = \delta_y f(x,y) - \alpha\delta_y g(x,y) = 0$…..(3)

- From (2) and (3) ∇f = α∇g

  Can also be written as

  $\delta_{x,y} f(x,y) = \alpha\delta_{x,y} g(x,y)$

Source: Abhisek Jana

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

What are f and g in the case of SVM?

$$f: \quad L(\vec{w}) \square \frac{\|\vec{w}\|^2}{2}$$

$$g: \quad y_i(\mathbf{w} \square \mathbf{x}_i \square b) \geq 1, \quad i \square 1,2,\ldots,N$$

How many constraints, are there in our SVM formulation?

N (look at g above: *i = 1,2,…,N*)

x and w are vectors and not scalars!

What do we do now?

The SVM Constraint is an inequality!

1. Instead of g(x, y) = k, we have g(x,y) <= 0
2. We impose Karush-Kuhn-Tucker (KKT) conditions on x, y and α as follows

L(x, y, α□□□□f(x, y) + αg(x, y).....(1)
∇L = 0......................................(2)
g(x, y) <= 0................................(3)
α□□□□□□□.......................□□□□□□□□□□
αg(x,y) = 0 ....(5)

(5) => α□□□□□□□for non-support vectors; g(x,y) = 0 for support vectors

If there are multiple constraints, g_i, each will have a Lagrange multiplier α_ι and the 2nd term of L above will be a summation.

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Leading Questions

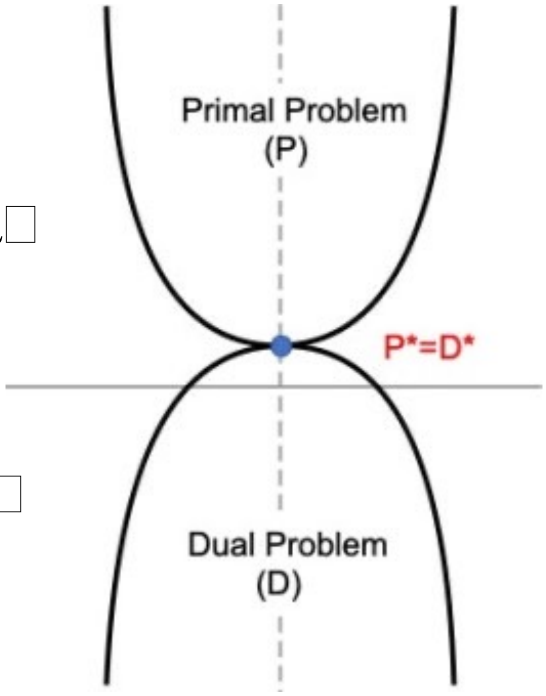
What is the goal of x and y w.r.t. L(x, y, α) ?	What is the goal of α w.r.t. L(x, y, α) ?	Where did you encounter a similar scenario?	When do minmax problems stabilize?
Minimize L(x, y, α)	Penalty term does the opposite – maximize L(x, y,□α□	GANs – minmax problems	Minimizing L w.r.t. x, y is same as maximizing L w.r.t. α => $\mathrm{argmin}_{x,y}L(x,y,\alpha) = \mathrm{argmax}_{\alpha}L(x,y,\alpha)$

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Strong Duality

$\min_{w,b} L(w, b, \alpha)$

$\max_{\alpha} L(w, b, \alpha)$

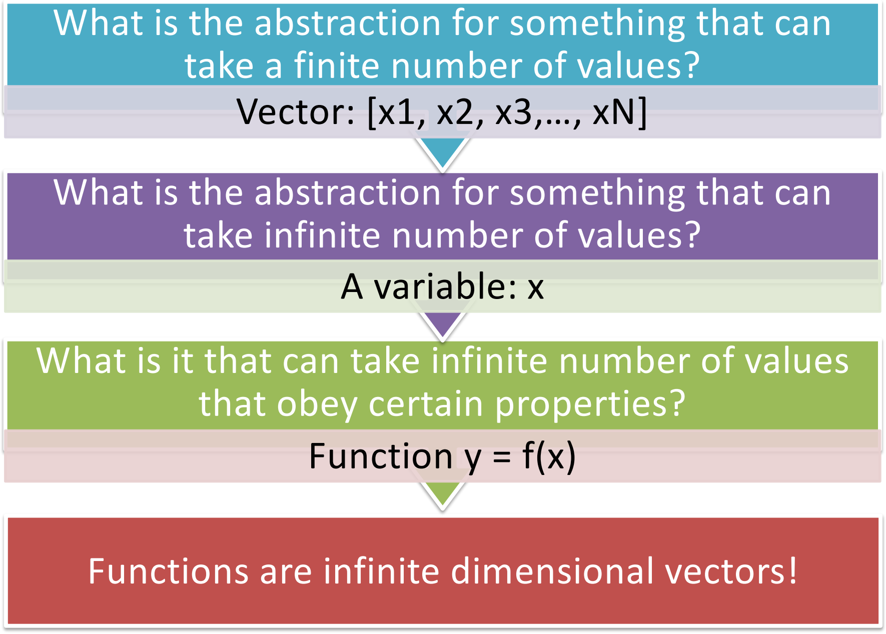


11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Leading Questions



What is the abstraction for something that can take a finite number of values?

Vector: [x1, x2, x3,…, xN]

What is the abstraction for something that can take infinite number of values?

A variable: x

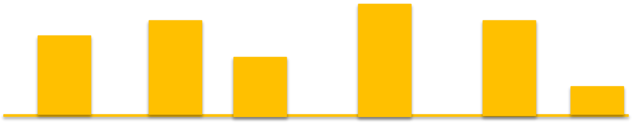
What is it that can take infinite number of values that obey certain properties?

Function y = f(x)

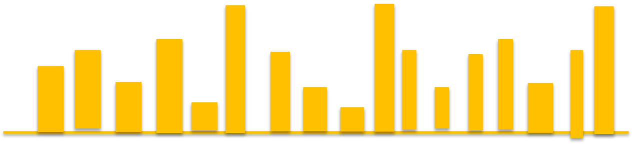
Functions are infinite dimensional vectors!

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

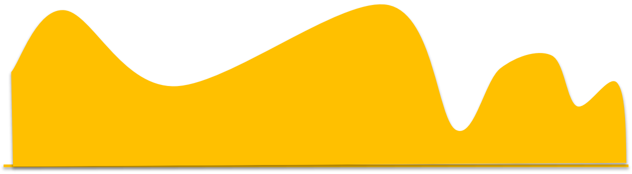
Remember discrete vs continuous random variables and PMF / PDF?



Bar chart of a vector with 6 dimensions



Bar chart of a vector with 17 dimensions



Bar chart of a vector with infinite dimensions

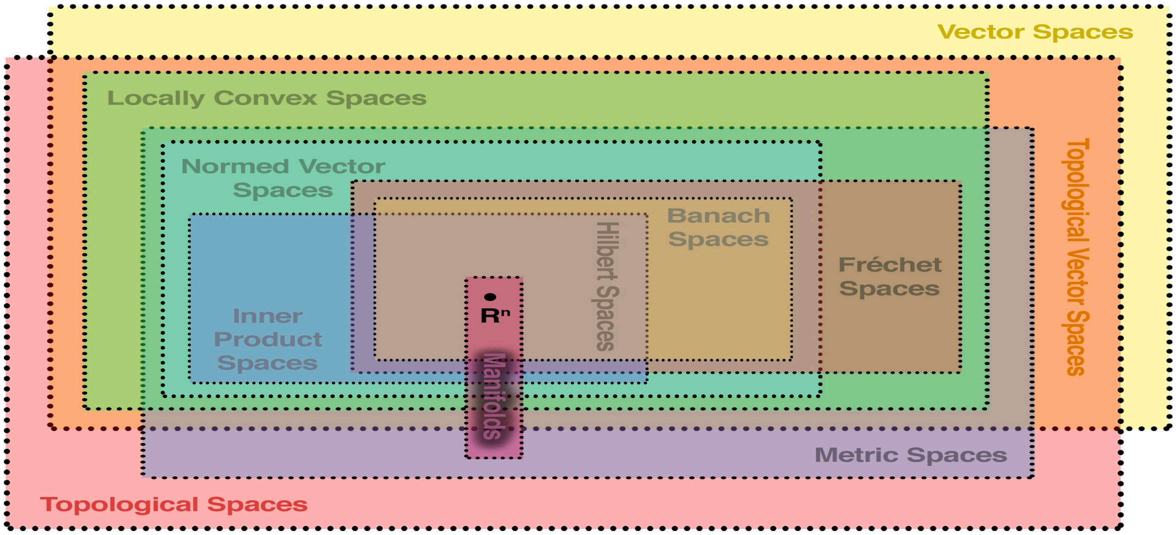
Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Abstraction	Vector Space	Function Space
Summation	Σ	$\int$
Norm	$\parallel x \parallel_p = (\Sigma\|x\|^p)^{1/p}$	$\parallel f \parallel_p = \left(\int \|f(x)\|^p dx\right)^{1/p}$
Inner Product	$< u, v > = \ \Sigma u_i . v_i$	$< f, g > = \int f(t). g(t)dt$
Orthogonality	$< u, v >$ = 0	$< f, g > = 0$
Bases	Orthonormal vectors	Orthonormal functions

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

The Dual is in Hilbert Space



Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Solving for w, b, and α Part-1: Primal (w and b)

$L(x,y,\alpha) = f(x,y) + \alpha(g(x,y) - c)$

$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i(w^T x_i + b))$

$$\nabla L = 0 \Rightarrow \delta_w L = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Classifier: $\mathrm{h(x)} = \mathrm{sgn}(w \cdot x + b) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x_i x} + b)$

For support vectors, $w.x_j + b = y_j$

$\Rightarrow b = y_j - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x_i x_j}$

$$\nabla L = 0 \Rightarrow \delta_b L = \sum_{i=1}^{n} \alpha_i y_i = 0$$

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Solving for w, b, and α□□Part-2: Dual (α□

L(w, b, α) = $\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i(w^T x_i + b))$

$$= \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i y_i w^T x_i - \sum_{i=1}^{n} \alpha_i y_i b + \sum_{i=1}^{n} \alpha_i$$

Substitute $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ *and* $\sum_{i=1}^{n} \alpha_i y_i = 0$ *from previous slide*

$$L = \frac{1}{2} w^T w - w^T w \; - b(0) + \sum_{i=1}^{n} \alpha_i = -\frac{1}{2} w^T w + \sum_{i=1}^{n} \alpha_i$$

= $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} (\sum_{i=1}^{n} \alpha_i y_i x_i)(\sum_{j=1}^{n} \alpha_j y_j x_j)$

= $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \Sigma_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{(x_i^T x_j})$

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Next Steps

- We use Quadratic Programming software or algorithms like SMO (Sequential Minimal Optimization) to solve for alphas
- At optimal solution, the data items corresponding to non-zero alphas are the support vectors
- Support vectors and the corresponding alphas can be saved as a model
- The feature vectors in the training set appear only inside dot products => ready for the kernel trick for non-linear data!

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Notes

1. We **converted** a problem in terms of the weight vector w and bias b into a problem in terms of the Lagrange multipliers α_i
2. Feature vectors x_i and labels y_i are **already known** from the training dataset.
3. Since the variables in our problem changed resulting in a dual, instead of minimizing L w.r.t. w□and b , **we must now maximize L w.r.t α_i**
4. The original problem is actually equivalent to $\min_{w,b} \max_{\alpha}$□L(w, b, α□□

□□□$\min_{w,b} \max_{\alpha}$□f(w, b) + αg(w, b))

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Observation

- When we use Lagrange "trick", QP, and duality, we notice that all the equations have only **dot products** of the data instances
- Nowhere will we need the individual values of the features
- Classification is anyway all about **pairwise similarity** of the data points, which the dot product indicates
- This finding comes in **handy** when we consider **non-linear** classification

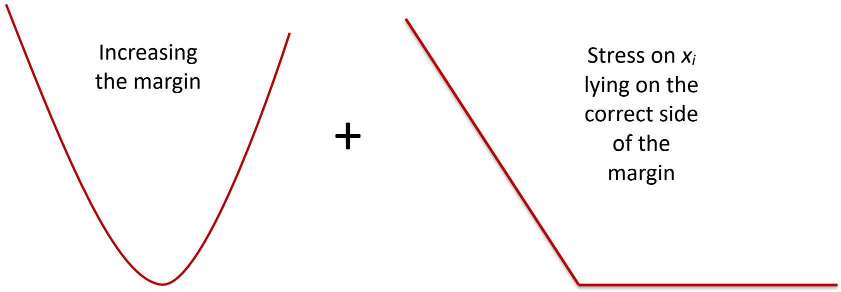
11/12/23

SVM Objective Function is Convex

$$\arg\min_w \underline{\frac{1}{2}\|w\|^2 + \sum}_{i=1}^{n} \alpha_i \max[0,1 \underline{- y_i w^T x_i]}$$

Regularization

(Hinge) Loss Function



Increasing the margin

Stress on x_i lying on the correct side of the margin

Sum of Convex Functions is Convex

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

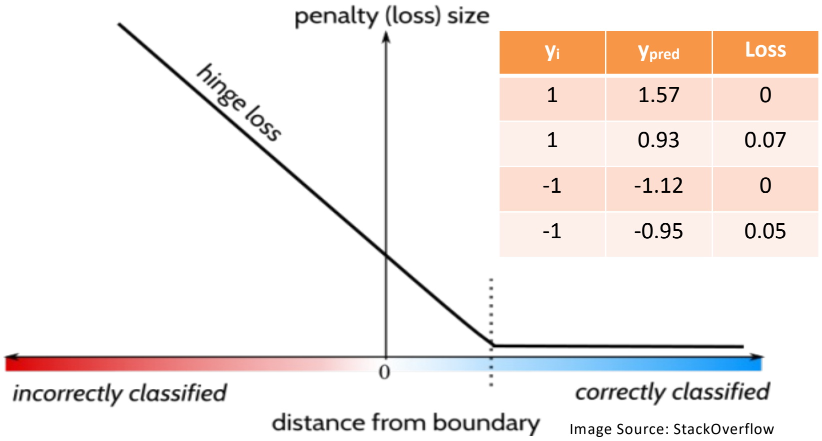
To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Hinge Loss

$$\arg\min_w \underline{\frac{1}{2}\|w\|^2 + \sum}_{i=1}^{n} \alpha_i \max[0,1 \underline{- y_i w^T x_i]}$$

Regularization

(Hinge) Loss Function



y_i	y_{pred}	Loss
1	1.57	0
1	0.93	0.07
-1	-1.12	0
-1	-0.95	0.05

Image Source: StackOverflow

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Various Loss Functions

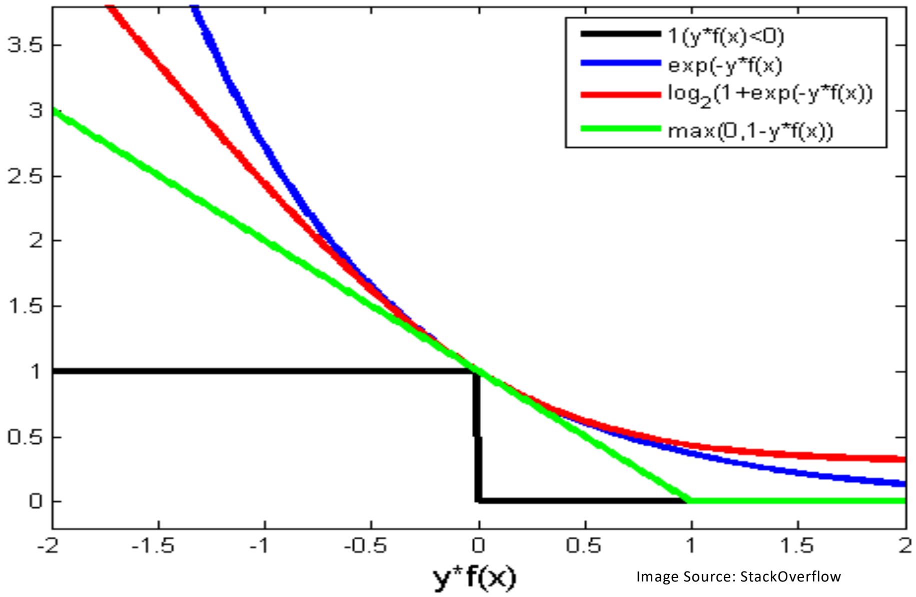
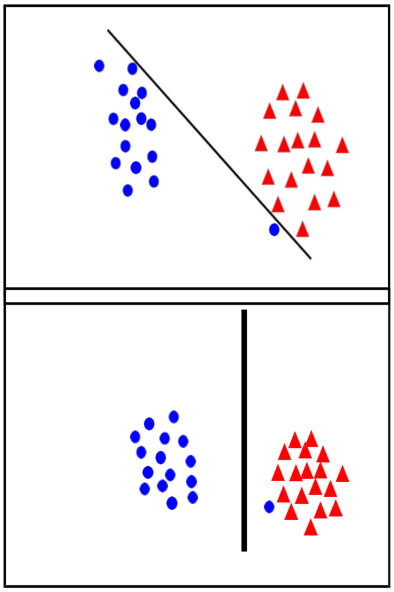


Image Source: StackOverflow

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Linear separability: What is the best w?



•the points can be linearly separated but there is a very **narrow margin**

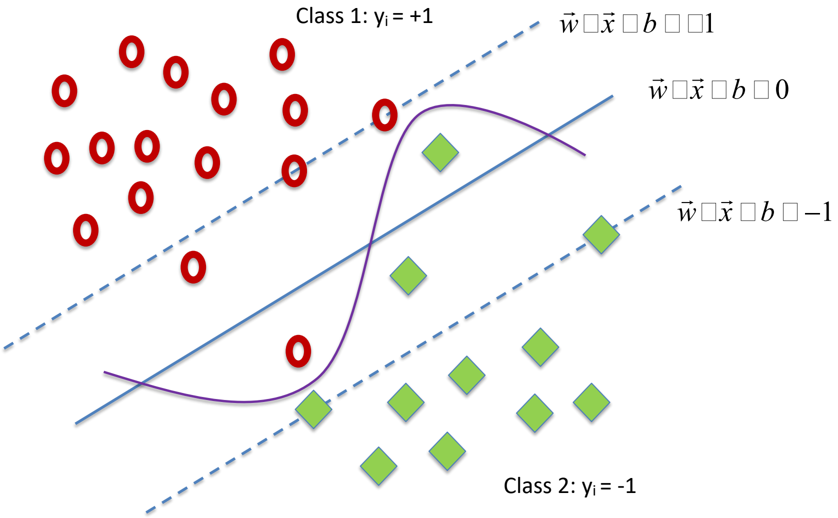
•but possibly the large margin solution is better, even though **one constraint is violated**

In general there is a trade off between the margin and the number of mistakes on the training data

Source: A. Zisserman

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

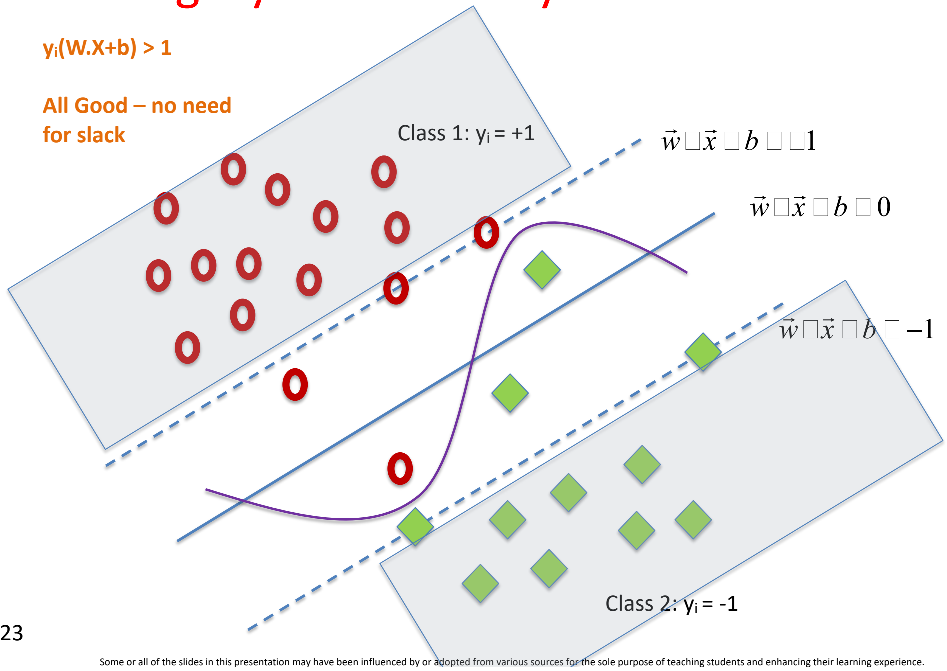
Slightly non-linear: Can we cut some slack?


Class 1: $y_i = +1$
$\vec{w} \Box \vec{x} \Box b \Box \Box 1$
$\vec{w} \Box \vec{x} \Box b \Box 0$
$\vec{w} \Box \vec{x} \Box b \Box -1$
Class 2: $y_i = -1$

We keep track of the "bad apples" using "slack" variables

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Category #1: Correctly Classified Points


Class 1: $y_i = +1$
$\vec{w} \Box \vec{x} \Box b \Box \Box 1$
$\vec{w} \Box \vec{x} \Box b \Box 0$
$\vec{w} \Box \vec{x} \Box b \Box -1$
Class 2: $y_i = -1$

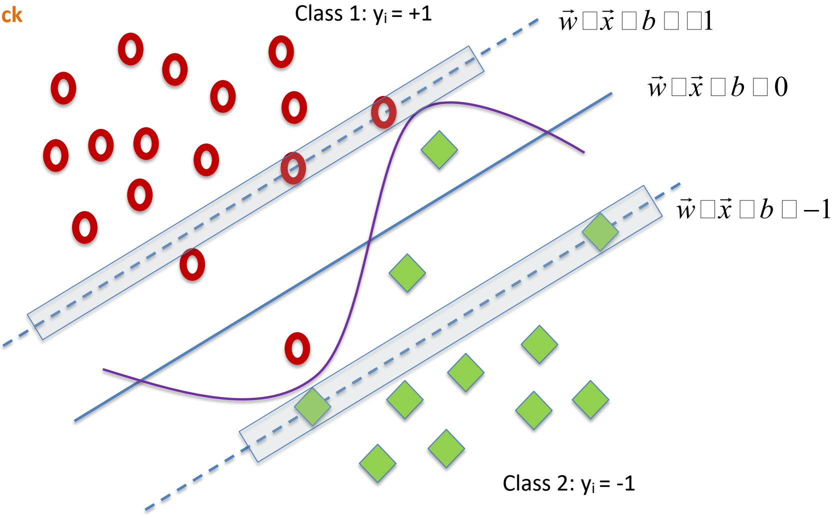
$y_i(W.X+b) > 1$

All Good – no need for slack

Category #2: Support Vectors

$y_i(W.X+b) = 1$

All Good – no need for slack


Class 1: $y_i = +1$
$\vec{w}\square\vec{x}\square b\square\square 1$
$\vec{w}\square\vec{x}\square b\square 0$
$\vec{w}\square\vec{x}\square b\square -1$
Class 2: $y_i = -1$

11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

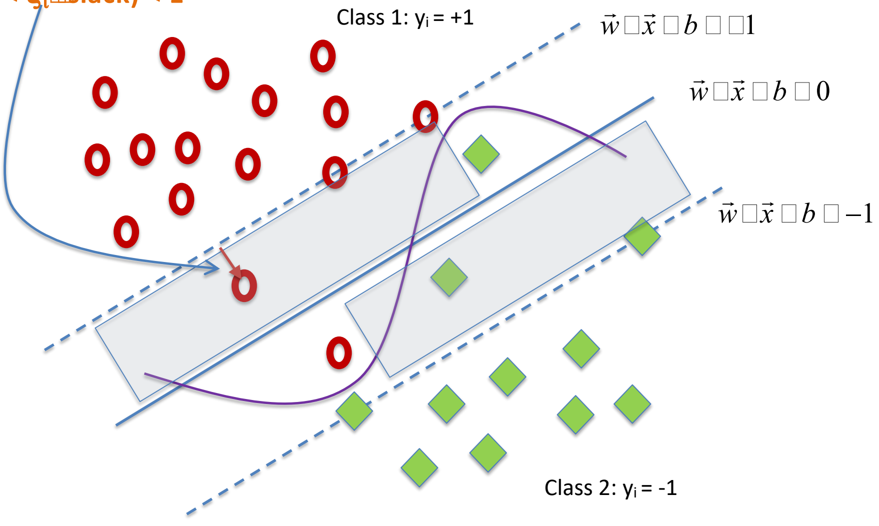
To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Category #3: Wrong Side of Margin

$y_i(W.X+b) < 1$

But classified correctly

$0 < \zeta_\iota$□□slack) < 1


Class 1: $y_i = +1$
$\vec{w}\square\vec{x}\square b\square\square 1$
$\vec{w}\square\vec{x}\square b\square 0$
$\vec{w}\square\vec{x}\square b\square -1$
Class 2: $y_i = -1$

11/12/23

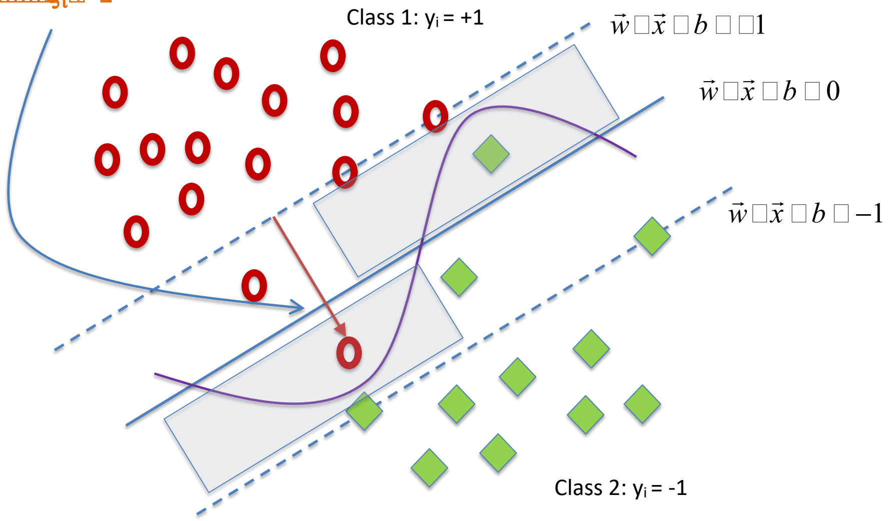
Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Category #4: Incorrectly Classified

...and wrong side of margin

y_i(W.X+b) < 1

□□□□□ζ_i□> 1


Class 1: y_i = +1
$\vec{w} \square \vec{x} \square b \square \square 1$
$\vec{w} \square \vec{x} \square b \square 0$
$\vec{w} \square \vec{x} \square b \square -1$
Class 2: y_i = -1

11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Soft Margin SVM: Objective Function

We add a slack term and a hyperparameter, C:

$$\text{Minimize } \frac{\|W\|^2}{2} + \text{C } \Sigma\square\zeta_i$$

and the constraints change as well

$$y_i(\text{W.X} + \text{b}) >= 1 - \zeta_i$$

Repeat the math with this additional set of variables, ζ_i – Gradients now include w.r.t ζ_i

- Large C => close to hard margin, less points misclassified, smaller margin, overfitting
- C-->infinity => hard margin

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Hyperparameter C and the trade-off

$$\underline{\arg\min_w \frac{1}{2}\|w\|^2} \quad + \quad \underline{C\sum_{i=1}^{n}\xi_i}$$

Increasing the margin

Vs

Stress on x_i lying on the correct side of the margin

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Non-linear SVM

Data cannot be separated by a single hyperplane

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

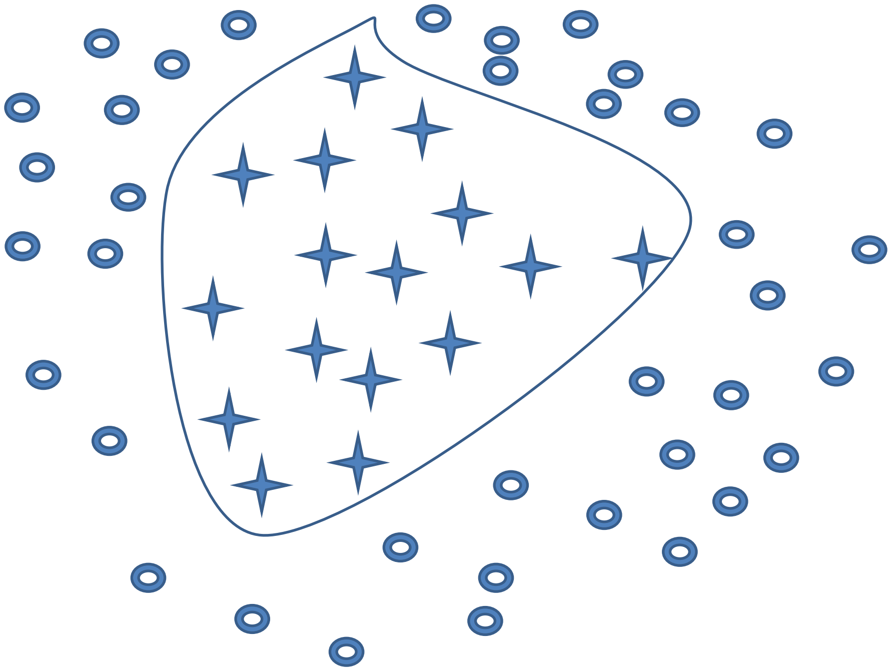
What do we do when our 1-bed apartment gets noisy and we can’t work?

We go to the library which is more spacious and quieter!

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

What if there was trick that could transport you from the 3D world to a space with infinite dimensions?

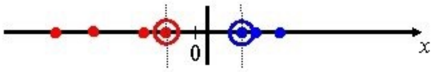
Or better still: Let you work in infinite dimensions without even visiting the space?



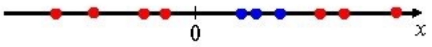
To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Nonlinear SVMs

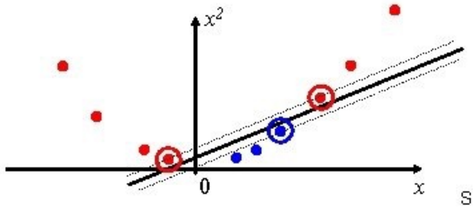
- Linearly separable dataset in 1D:



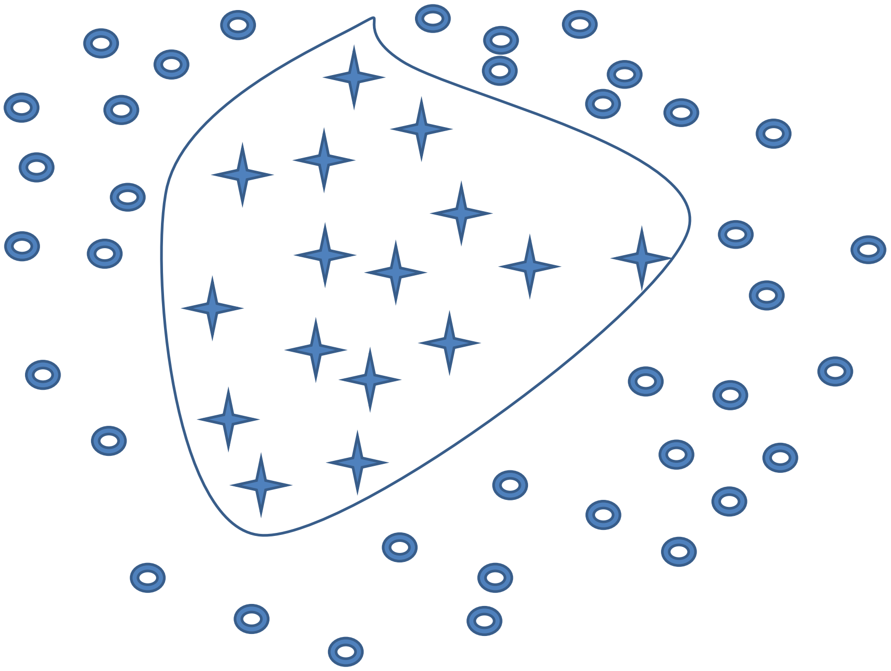
- Non-separable dataset in 1D:



- We can map the data to a *higher-dimensional space*:

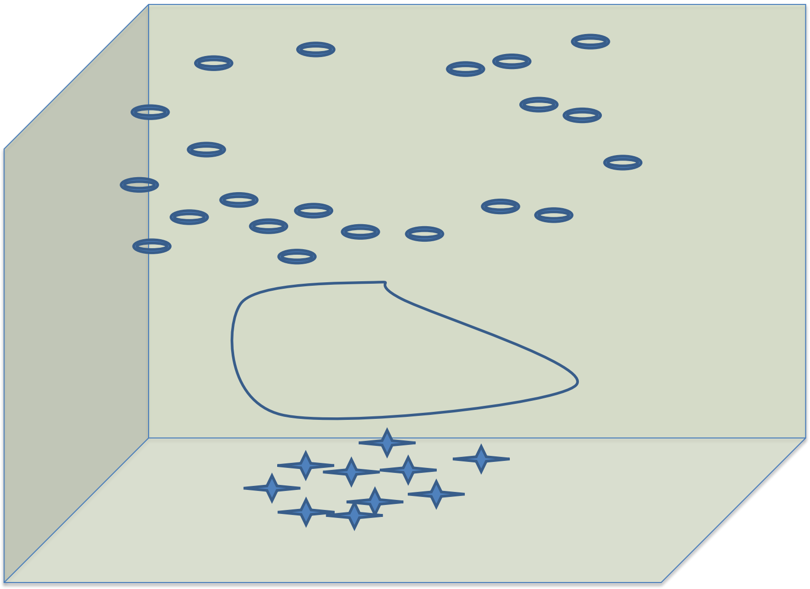


Slide credit: Andrew Moore



To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

WHEN MAPPED TO 3-D…

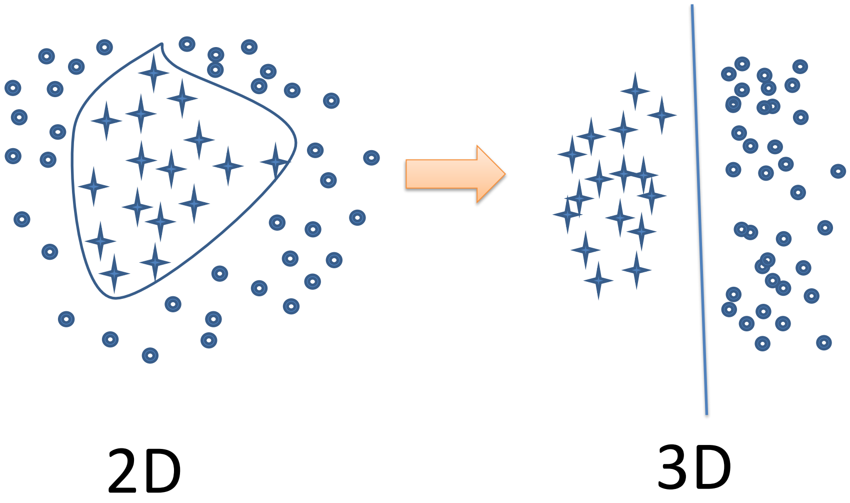


11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Map from 2D to 3D



11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Cover’s Theorem

“Given a set of training data that is not linearly separable, one can, with high probability transform it into a training set that is linearly separable by projecting it into a higher-dimensional space via some non-linear transformation.” Proof

Thomas M. Cover

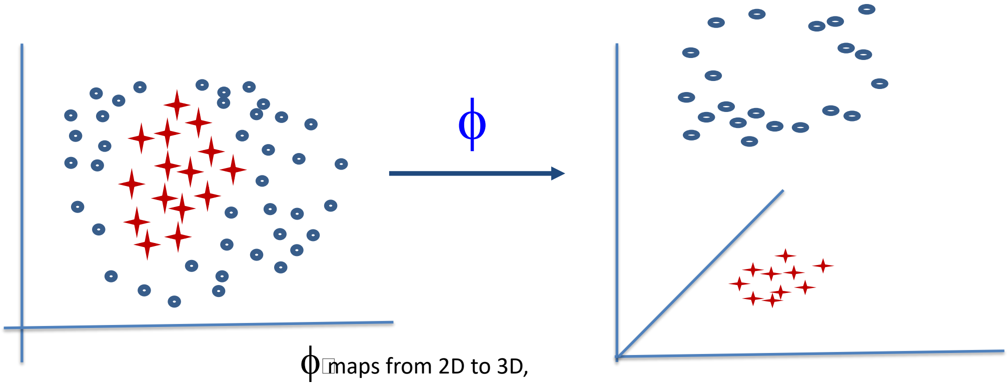
Photo credit: IEEE Information Theory Society

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Cover’s theorem illustrated

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \xrightarrow{\phi} \begin{pmatrix} z_1 = x_1^2 \\ z_2 = \sqrt{2}x_1x_2 \\ z_3 = x_2^2 \end{pmatrix}$$



ϕ maps from 2D to 3D,
making the problem linearly separable

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

What is ϕ

Mapping Function:

$$\begin{array}{llll} \phi : & \Re^2 & \longrightarrow & \Re^3 \\ & (x_1, x_2) & \longmapsto & (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \end{array}$$

Equation of the Classifying Hyperplane in 3D:

$$\omega^T z + b = 0$$

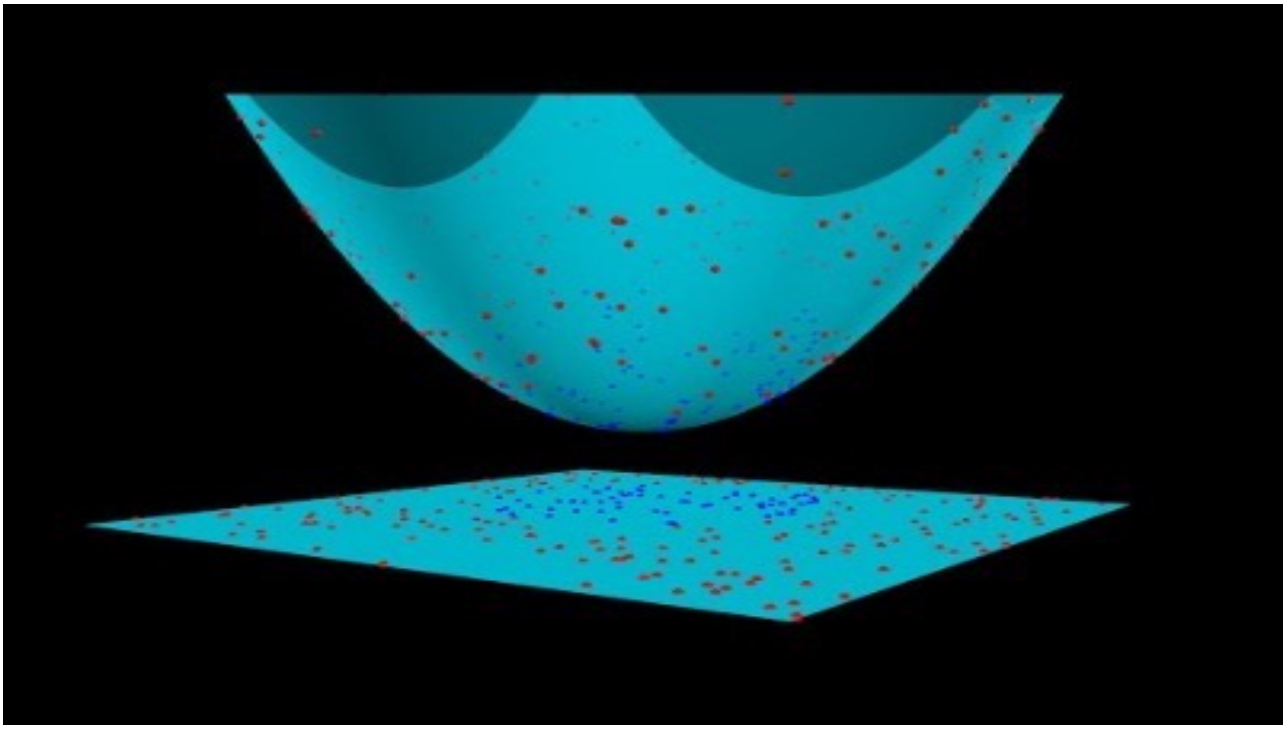
Substituting for z, we get the equation of an ellipse in 2D:

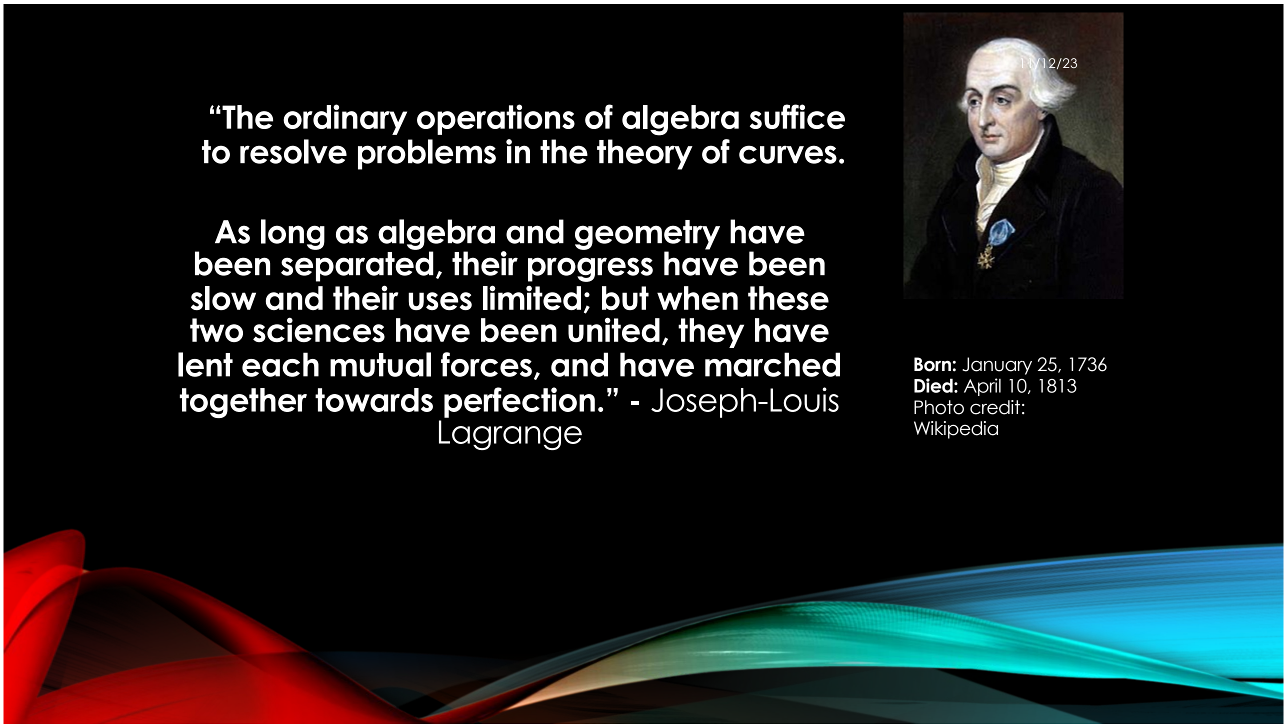
$$\omega_1 x_1^2 + \omega_2 \sqrt{2} x_1 x_2 + \omega_3 x_2^2 = 0$$

ϕ Maps x ∈ X to R^D where D can be potentially infinite

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.





"The ordinary operations of algebra suffice to resolve problems in the theory of curves.

As long as algebra and geometry have been separated, their progress have been slow and their uses limited; but when these two sciences have been united, they have lent each mutual forces, and have marched together towards perfection." - Joseph-Louis Lagrange

Born: January 25, 1736
Died: April 10, 1813
Photo credit:
Wikipedia

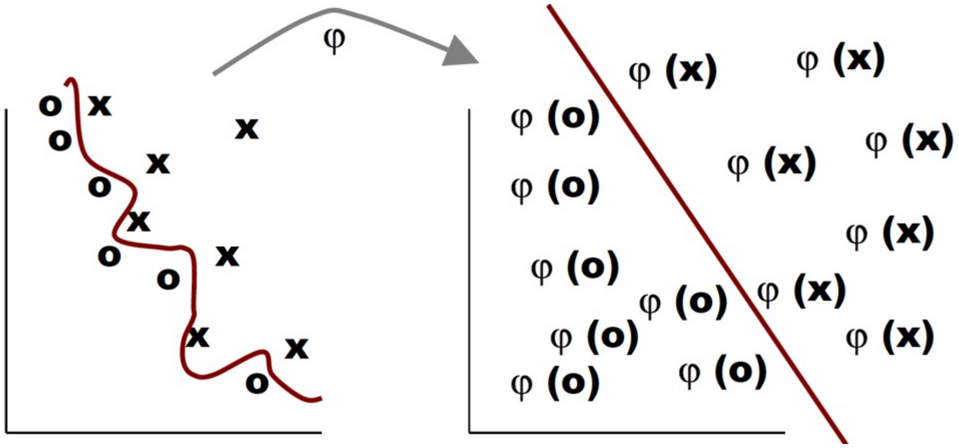
To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

A different perspective of Cover's theorem

1. Machine Learning is essentially is fitting a curve $y = f(x)$ to the data.
2. We fit a straight line if the dependencies among the data are linear, $y = mx + c$
3. How can we fit a straight line if the relationships in the data are non-linear, say modeled by $y = f(x) = mx^2 + c$?
4. Export the data to a different space where $x' = mx^2 + c \; then \; y' = x'$ fits the data!
5. SVM finds $y = f(x)$ by transforming the data to a different space.

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Transformation to Higher Dimension



Original Space   Higher Dimensional Space

Image Source: R. Berwick

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

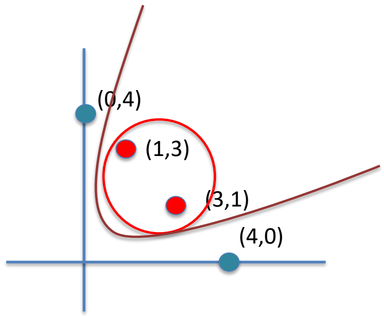
To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

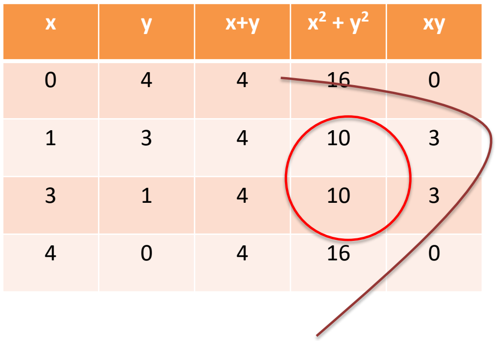
Problems with the transformation

- ✓ If there are 1M training samples, need to evaluate the function ϕ at 1M points
- If the transformed space has 1M dimensions, need to evaluate dot products across all the 1M dimensions for all pairs of samples
- => Computationally infeasible

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Not all transformations apply!





x	y	x+y	$x^2 + y^2$	xy
0	4	4	16	0
1	3	4	10	3
3	1	4	10	3
4	0	4	16	0

Which transformations apply and how do we solve all these problems?

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

How long does it take to train a model in self-driving car to recognize objects?

Lesson: Capture only the core (kernel) and leave the rest!

Only dot products are needed in the dual formulation

A “Gram” Matrix of inner products in the transformed space is all we need

$< \phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_1) >$	$< \phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2) >$	…	$< \phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_n) >$
$< \phi(\boldsymbol{x}_2), \phi(\boldsymbol{x}_1) >$	$< \phi(\boldsymbol{x}_2), \phi(\boldsymbol{x}_2) >$	…	$< \phi(\boldsymbol{x}_2), \phi(\boldsymbol{x}_n) >$
…	…	…	…
$< \phi(\boldsymbol{x}_n), \phi(\boldsymbol{x}_1) >$	$< \phi(\boldsymbol{x}_n), \phi(\boldsymbol{x}_2) >$	…	$< \phi(\boldsymbol{x}_n), \phi(\boldsymbol{x}_n) >$

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Simplifying the Gram matrix computation

- How can we avoid the computing cost of the (i) transformations and (ii) inner products?
- What if we found a function

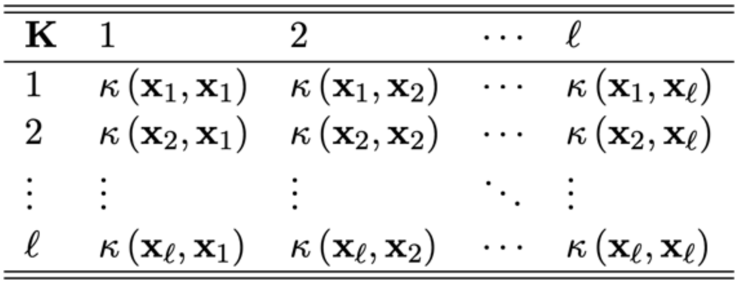
$$K(x_m, x_n) =< \phi(\boldsymbol{x}_m), \phi(\boldsymbol{x}_n) > \forall \boldsymbol{x}_m, \boldsymbol{x}_n$$

- We can compute K in the original feature space without computing ϕ’s or their inner products
- K should satisfy the properties of a Gram matrix
- Classifier is given by

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)^T \phi(x)) + b = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x)) + b$$

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Kernel Gram matrix as a function



$\mathbf{K}$	1	2	$\cdots$	ℓ
1	$\kappa(\mathbf{x}_1, \mathbf{x}_1)$	$\kappa(\mathbf{x}_1, \mathbf{x}_2)$	$\cdots$	$\kappa(\mathbf{x}_1, \mathbf{x}_\ell)$
2	$\kappa(\mathbf{x}_2, \mathbf{x}_1)$	$\kappa(\mathbf{x}_2, \mathbf{x}_2)$	$\cdots$	$\kappa(\mathbf{x}_2, \mathbf{x}_\ell)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
ℓ	$\kappa(\mathbf{x}_\ell, \mathbf{x}_1)$	$\kappa(\mathbf{x}_\ell, \mathbf{x}_2)$	$\cdots$	$\kappa(\mathbf{x}_\ell, \mathbf{x}_\ell)$

As $\ell \to \infty$, the above infinite dimensional matrix can be represented by a function under certain circumstances

First, we need an (i) an inner product space that is (ii) complete to which the function maps

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Kernel Functions

- Represent inner products in the transformed space => need to satisfy the properties of inner product, such as

symmetry $< x_1, x_2 > = < x_2, x_1 >$ and
positive semi-definiteness $< x, x > \geq 0$

- Implies that the Gram matrix has to be symmetric and positive semi-definite (=> Eigenvalues of the matrix are non-negative)

Mercer Conditions

For a kernel function to truly map to an inner product, the kernel gram matrix must also be

1. Symmetric => $K(x,y) \ = \ K(y,x)$ and
2. The diagonal elements represent the squared norms, therefore must be positive
3. Positive semidefinite =>

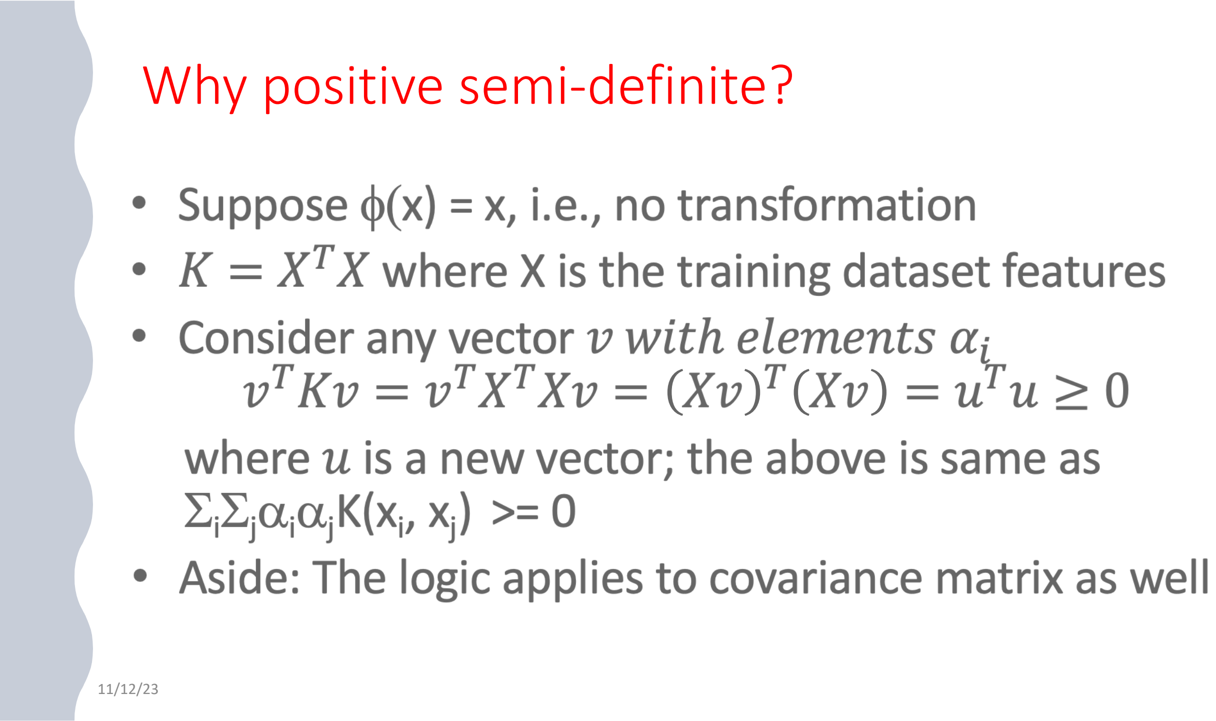
$\Sigma_i\Sigma_j\alpha_i\alpha_j K(x_i, x_j)$ >= 0 for all α_i and α_j in R

(or) $v^T K v \geq 0 \ \forall v$

(or) its eigenvalues are nonnegative. Why?

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.



Why positive semi-definite?

- Suppose $\phi(x) = x$, i.e., no transformation
- $K = X^T X$ where X is the training dataset features
- Consider any vector v *with elements* α_i

$$v^T K v = v^T X^T X v = (Xv)^T (Xv) = u^T u \geq 0$$

where u is a new vector; the above is same as $\Sigma_i\Sigma_j\alpha_i\alpha_j K(x_i, x_j)$ >= 0

- Aside: The logic applies to covariance matrix as well

11/12/23

Simplified Mercer's Theorem

"Let $K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be given.

Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any

$\{x^{(1)}, \dots, x^{(m)}\}, (m < \infty),$

the corresponding kernel matrix is symmetric positive semi-definite."

For Mercer kernels, there exists a transformation function φ□such that

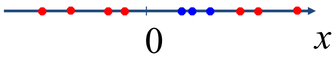
K(x,x') = φ□x)^T□□φ□x')

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

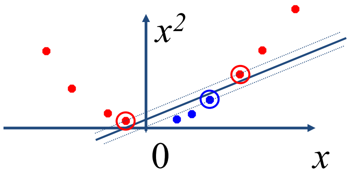
To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

The Kernel Trick

- Non-separable data in 1D:



- Apply mapping $\varphi(x) = (x, x^2)$:



Think of kernel as a function that directly gives you a similarity metric in the transformed space

•Dot Product of the new, transformed vectors

$\bullet\varphi(x)^T\varphi(x') = K(x, x') = xx' + x^2x'^2$

11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Kernel example 1: Polynomial

- Polynomial kernel with degree d and constant c:

$$K(x, x') = (x^T x' + c)^d$$

- What this looks like for $d = 2$:

$$x = (u, v), \qquad x' = (u', v')$$

$$K(x, x') = (uu' + vv' + c)^2$$

$$= u^2 u'^2 + v^2 v'^2 + 2uu'vv' + 2cuu' + 2cvv' + c^2$$

$$= \phi\ (x)^{\mathrm{T}}.\ \phi\ (x), where$$

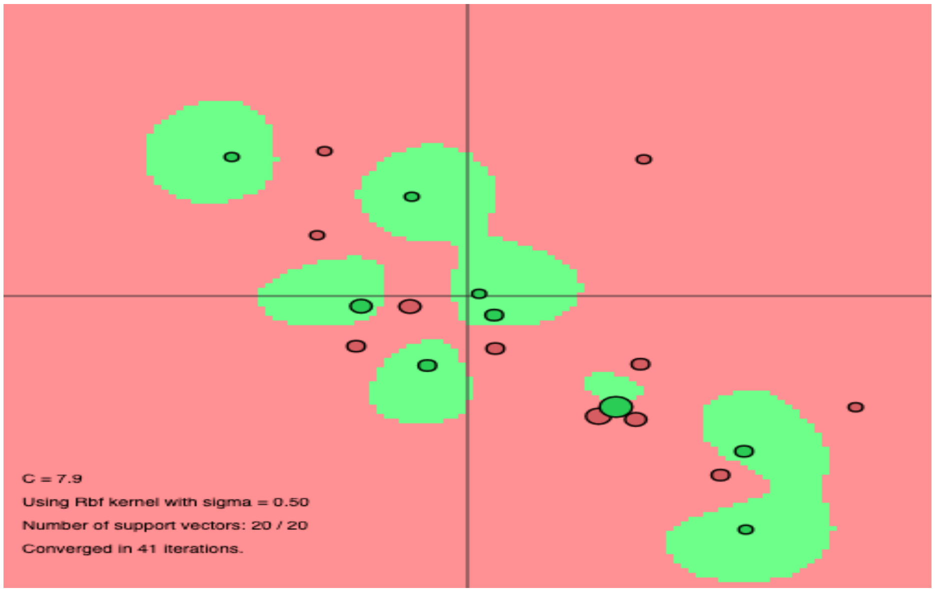
$$\phi\ (x) = (u^2, v^2, \sqrt{2}uv, \sqrt{2c}u, \sqrt{2c}v, c)$$

We mapped 2D to 6D space!

- Thus, the explicit feature transformation consists of all polynomial combinations of individual dimensions of degree up to d

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

What if the data is extremely nonlinear?



C = 7.9

Using Rbf kernel with sigma = 0.50

Number of support vectors: 20 / 20

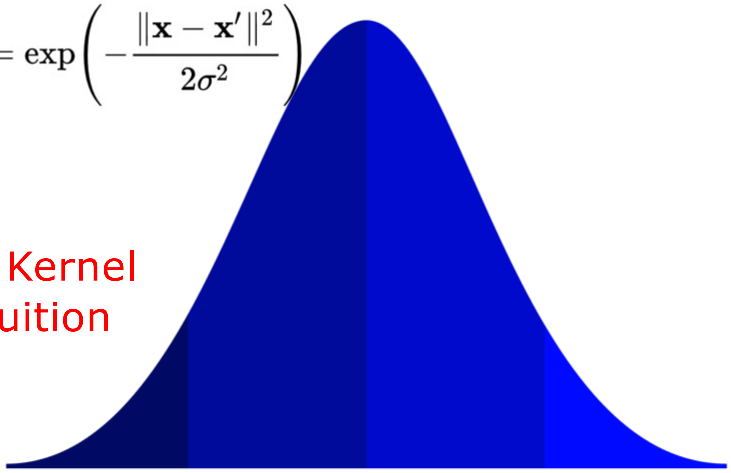
Converged in 41 iterations.

Source: Andrej Karpathy

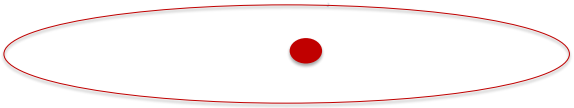
Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

RBF Kernel Intuition



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)



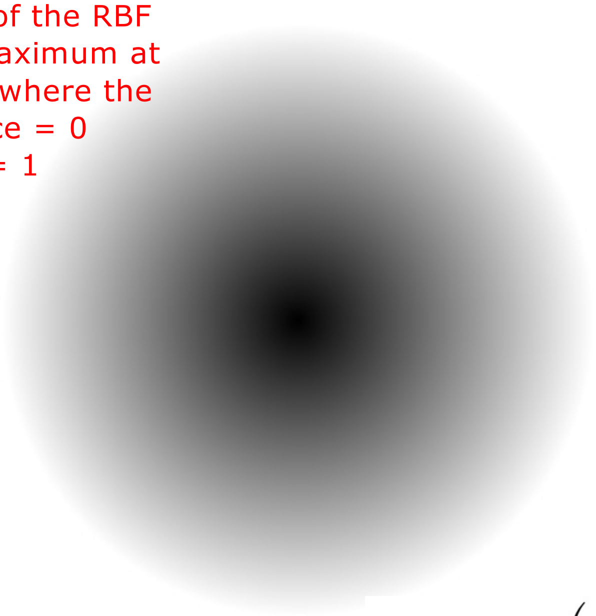
11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

The value of the RBF Kernel is maximum at the center where the distance = 0

$e^0 = 1$



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

11/12/23

RBF Kernel Space

1. Gaussian radial basis function (RBF) kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j)$ =

$$e^{-\|X_i - X_j\|^2 / 2\sigma^2}$$

1. Suppose there are 5 original 2-dimensional points:
   a) $x_1(0, 0)$, $x_2(4, 4)$, $x_3(-4, 4)$, $x_4(-4, -4)$, $x_5(4, -4)$
2. If we set σ to 4, we will have the following points in the kernel space
   a) E.g., $\left||x_1 - x_2|\right|^2 = (0-4)^2 + (0-4)^2 = 32$, thus, $K(x_1, x_2) = e^{-\frac{32}{2\cdot 4^2}} = e^{-1}$

Original Space

	x	y
x_1	0	0
x_2	4	4
x_3	-4	4
x_4	-4	-4
x_5	4	-4

RBF Kernel Space ($\sigma = 4$)

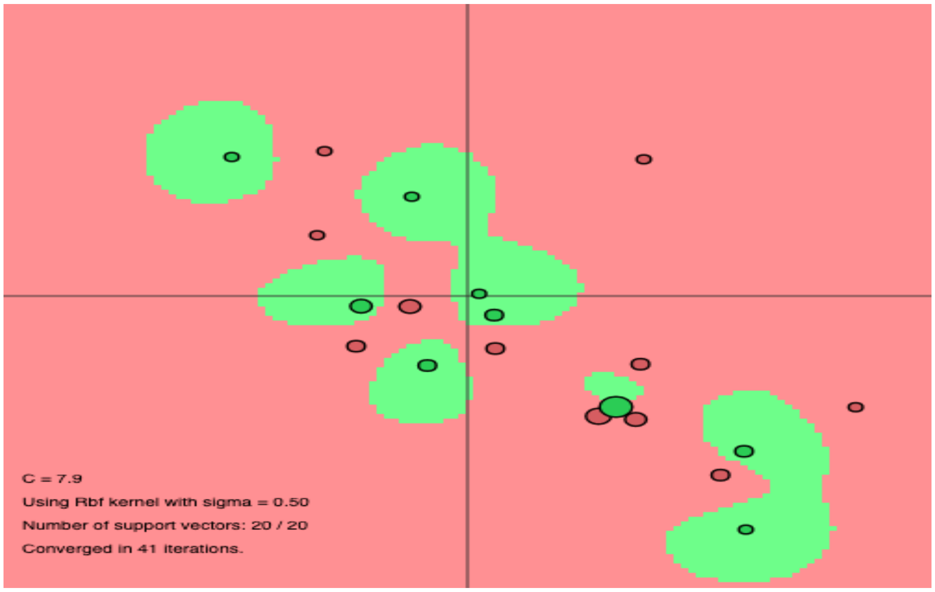
$K(x_i, x_1)$	$K(x_i, x_2)$	$K(x_i, x_3)$	$K(x_i, x_4)$	$K(x_i, x_5)$
0	$e^{-\frac{4^2+4^2}{2\cdot 4^2}}$ $= e^{-1}$	e^{-1}	e^{-1}	e^{-1}
e^{-1}	0	e^{-2}	e^{-4}	e^{-2}
e^{-1}	e^{-2}	0	e^{-2}	e^{-4}
e^{-1}	e^{-4}	e^{-2}	0	e^{-2}
e^{-1}	e^{-2}	e^{-4}	e^{-2}	0

Source: Jiawei Han

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

What if the data is extremely nonlinear?

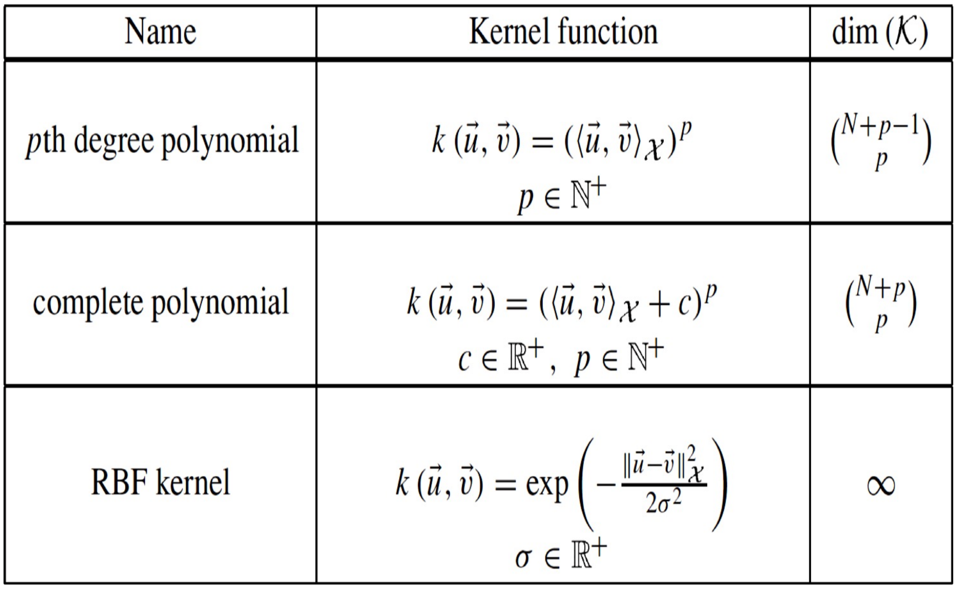


```
C = 7.9
Using Rbf kernel with sigma = 0.50
Number of support vectors: 20 / 20
Converged in 41 iterations.
```

Source: Andrej Karpathy

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Important Kernel Functions



Name	Kernel function	dim $(\mathcal{K})$
pth degree polynomial	$k(\vec{u},\vec{v}) = (\langle \vec{u},\vec{v}\rangle_{\mathcal{X}})^p$ $p \in \mathbb{N}^+$	$\binom{N+p-1}{p}$
complete polynomial	$k(\vec{u},\vec{v}) = (\langle \vec{u},\vec{v}\rangle_{\mathcal{X}} + c)^p$ $c \in \mathbb{R}^+,\ p \in \mathbb{N}^+$	$\binom{N+p}{p}$
RBF kernel	$k(\vec{u},\vec{v}) = \exp\left(-\frac{\lVert\vec{u}-\vec{v}\rVert_{\mathcal{X}}^2}{2\sigma^2}\right)$ $\sigma \in \mathbb{R}^+$	∞

Source: R. Herbrich

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

How Many Dimensions?

1. Polynomial Kernel when d is 2 and # of features is 2 (previous slide) results in a new feature space of 6 dimensions
2. In general, a polynomial kernel will result in a new feature space with ${}^{n+p}C_p$ dimensions
3. How about for RBF? Think Taylor’s series:

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \qquad \text{=> infinite dimensions}$$

$e^x = 1 + x + x^2 / \mathbf{2!} + x^3 / \mathbf{3!} + x^4 / \mathbf{4!} + x^5 / \mathbf{5!} + \ldots$ at a = 0

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

RBF Kernel as a sum of Polynomial Kernels

Follows from Taylor Series:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots , \quad -\infty < x < \infty$$

=> Infinite dimensions

11/12/23

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

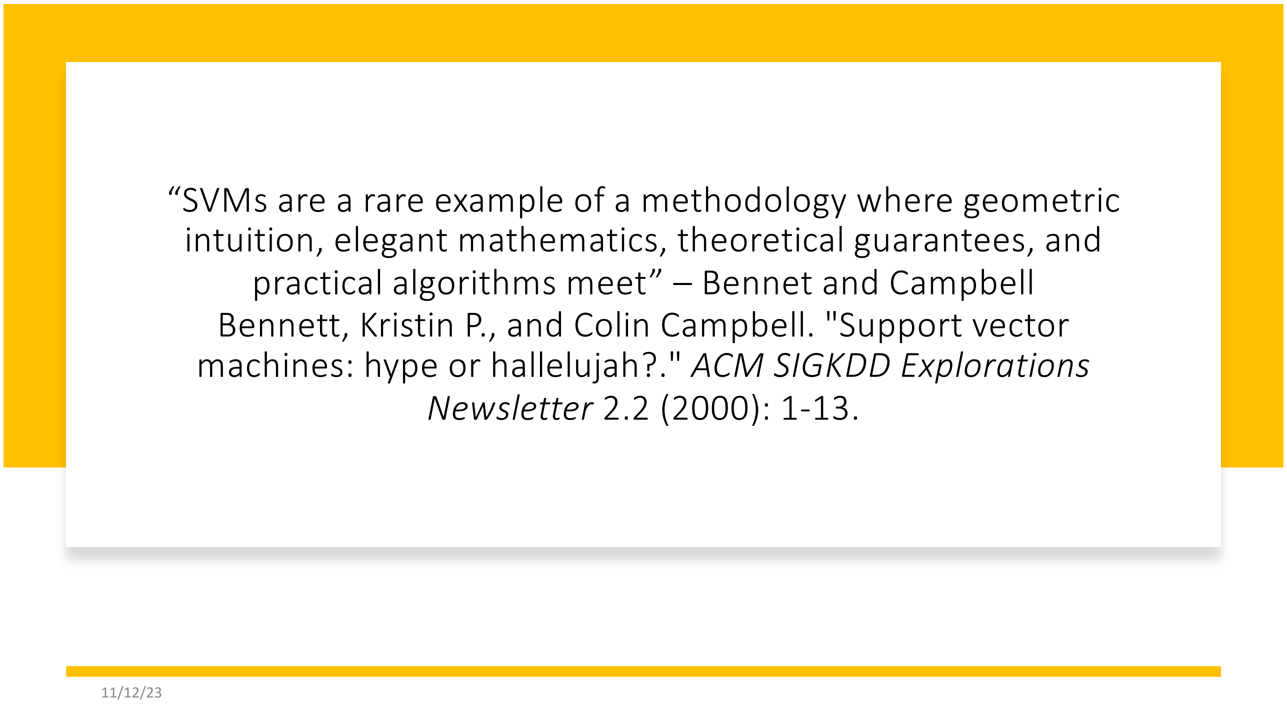
The Kernel Trick

1. Replace inner products in original feature space, <x,x’> with a call to the kernel function, K(x,x’)
2. K(x,x’) can be written as a dot product of transformation function ϕ□as□ϕ□x□^{T}□□ϕ□x’□
3. In effect, we have transformed

<x,x’> ➔ < ϕ□x□$^{□}$□ϕ□x’□□

=> Transforming the data in original feature space, X to data in a transformed feature space in higher dimensions

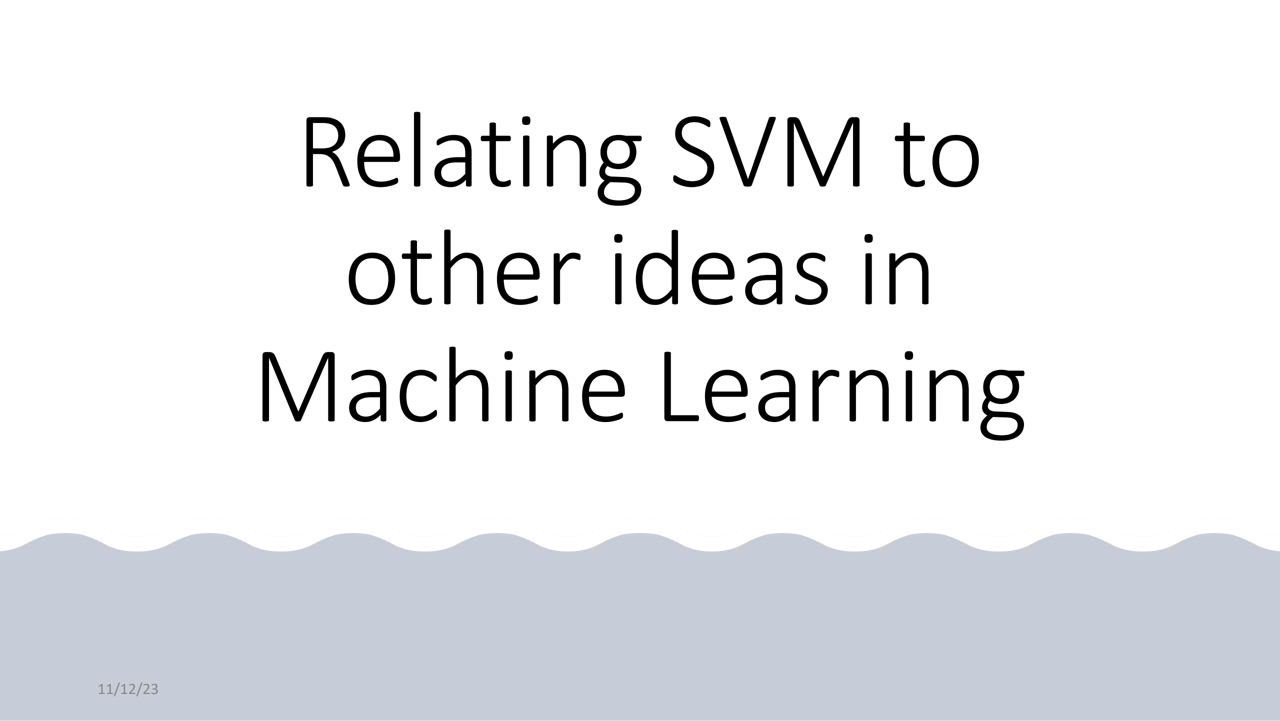
Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.



“SVMs are a rare example of a methodology where geometric intuition, elegant mathematics, theoretical guarantees, and practical algorithms meet” – Bennet and Campbell
Bennett, Kristin P., and Colin Campbell. "Support vector machines: hype or hallelujah?." *ACM SIGKDD Explorations Newsletter* 2.2 (2000): 1-13.

11/12/23

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.



Relating SVM to other ideas in Machine Learning

11/12/23

Kernel SVM and K-NN

$$f(x) = \operatorname{sgn}(\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)^T \phi(x)) = \operatorname{sgn}(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x))$$

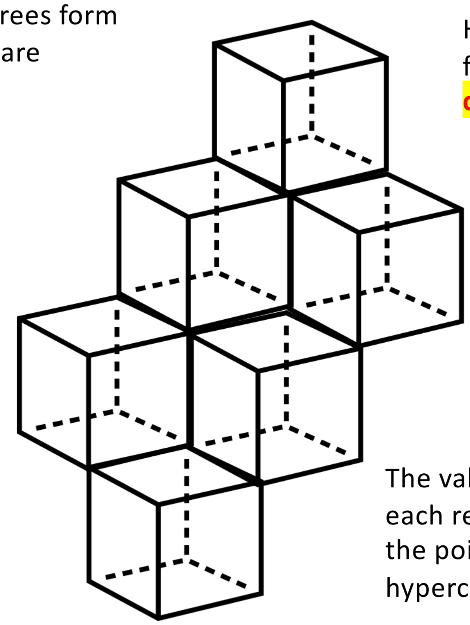
- If f(x) > 0, y = +1 otherwise, y = -1
- K is an inner product, which is a distance metric
- α_i = 0 for non-support vectors, by KKT
- α_i can be viewed as a weight or relative importance of each support vector
- This is instance-based learning, like K-NN
- Here the instances are support vectors

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

SVM and Regression Trees

Regression Trees form regions that are **hypercubes**



Hypercubes are formed based on **decision splits**

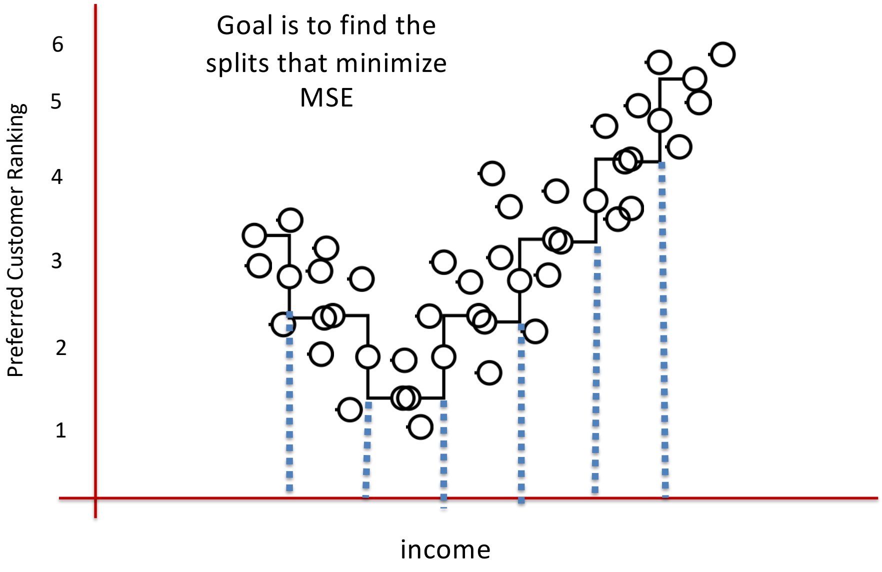
The value associated with each region is $Y_{average}$ of the points in that hypercube

11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

Single Feature - Binning



Goal is to find the splits that minimize MSE

11/12/23

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

Regression Trees

- In regression where y is continuous, a test item's expected target variable is predicted as

$$f(x) = \sum_{m=1}^{M} \left(\frac{1}{k} \left(\sum_{i=1}^{k} y_i \right) \mathbb{I}(X \in Rm) \right)$$

Each Region

Average of the target variable, y_i in Region R_m

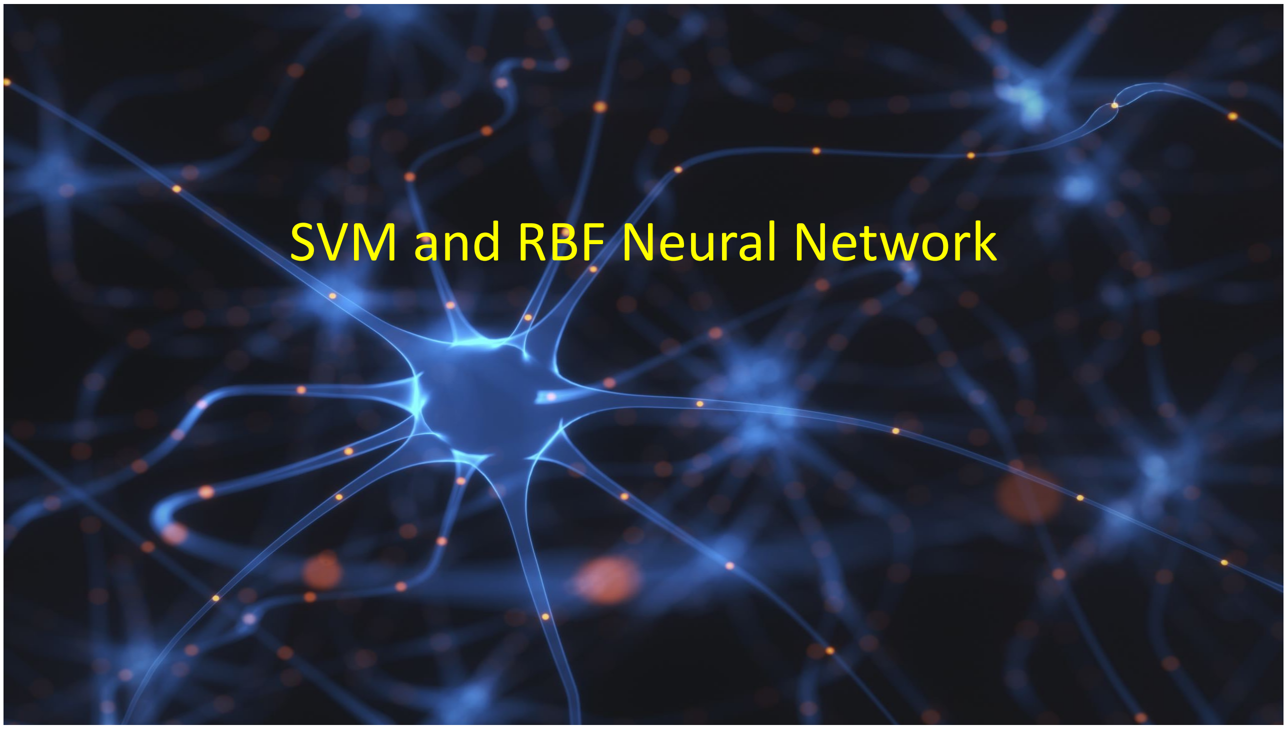
If the test feature vector is in Region R_m

Compare with SVM

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x)\right)$$

Some or all of the slides in this presentation may have been influenced by or adopted from various sources for the sole purpose of teaching students and enhancing their learning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.

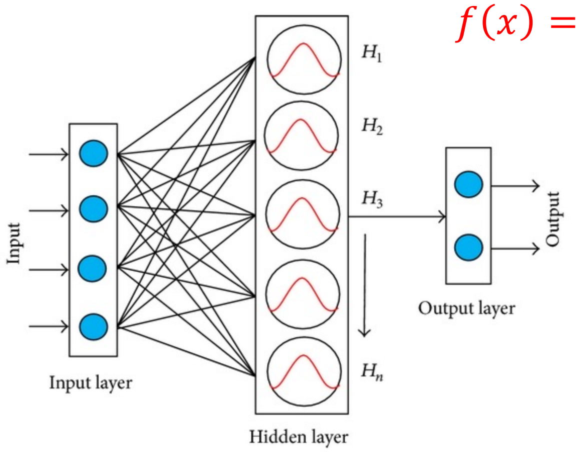


SVM and RBF Neural Network

$$RBNN:\ f(x) = \sum_i w_i e^{-(\frac{||x-x_i||^2}{2\sigma^2})}$$

SVM RBF Kernel

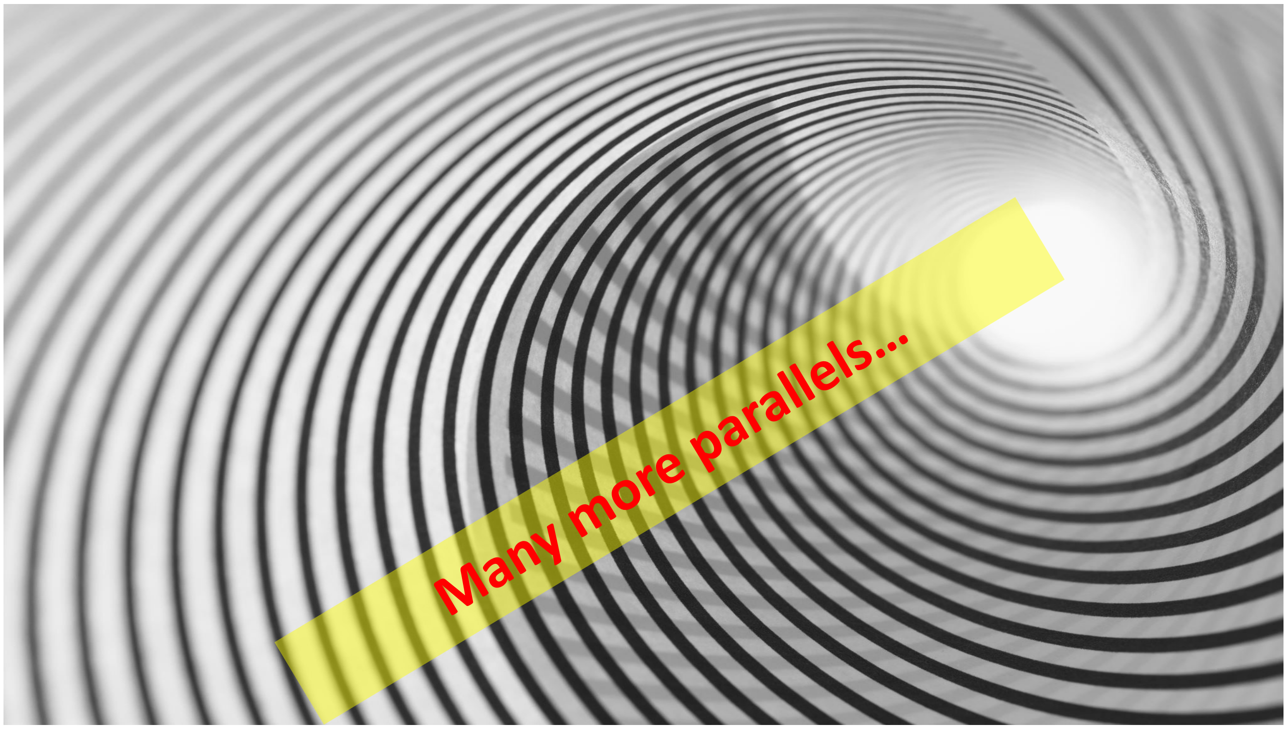
$$f(x) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i\, e^{-(\frac{||x-x_i||^2}{2\sigma^2})})$$



Support vectors are the hidden nodes!

This Photo by Unknown Author is licensed under CC BY ning experience.

To cite this presentation: Pendyala, V.S. (2022) "Exploring the math in Support Vector Machines". IEEE Computer Society, Santa Clara Valley Chapter Webinar.



Many more parallels…

https://www.sjsu.edu/people/vishnu.pendyala/

@vishnupendyala

11/12/23

This Photo by Unknown Author is licensed under CC BY-SA

To cite this presentation: Pendyala, V.S. (2022) “Exploring the math in Support Vector Machines”. IEEE Computer Society, Santa Clara Valley Chapter Webinar.

11/12/23