Revision

HARD MARGIN SVM

References

- Machine Learning for Intelligent Systems, Kilian Weinberger, Cornell, Lecture 9,
  - https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote09.html
- Support Vector Machine: how it really works,
  - https://www.youtube.com/watch?v=A7FeQekjd9Q , Victor Lavrenko, Assistant Professor at the University of Edinburgh
- Support Vector Machine Python Example, Cory Maklin,
  - https://towardsdatascience.com/support-vector-machine-pythonexample-d67d9b63f1c8
- Support Vector Machine: Complete Theory, Saptashwa Bhattacharyya,
  - https://towardsdatascience.com/understandingsupport-vector-machine-part-1-lagrange-multipliers-5c24a52ffc5e
- Support Vector Machine: Digit Classification with Python; Including my Hand Written Digits, Saptashwa Bhattacharyya
  - https://towardsdatascience.com/support-vector-machine-mnist-digitclassification-with-python-including-my-hand-written-digits-83d6eca7004a
- Support Vector Machine: Kernel Trick; Mercer's Theorem, Saptashwa Bhattacharyya,
  - https://towardsdatascience.com/understanding-supportvector-machine-prt-2-kernel-trick-mercers-theorem-e1e6848c6c4d

Hard Margin Support Vector Classifier

$$\min_{w,b} w^T w$$
$$such\ that \quad \forall i\ y_i(W^T x_i + b) \geq 0$$
$$\min_i |w^T x_i + b| = 1$$

- Hard constraints! Luckily, we have an equivalent formulation for the optimal solution:

$$\min_{w,b} w^T w$$
$$such\ that \quad \forall i\ y_i(W^T x_i + b) \geq 1$$

- This is a linearly constrained quadratic optimization problem, that could be solved using quadratic programming, e.g., using Lagrange Multipliers.

Some Intuition

The width of the "road" that we are building have 1 distance to the edges on the both sides from the hyperplane, thus $\frac{2}{||w||_2}$ is the total width, which is maximized by minimizing $||w||_2$.

$$\max_{w,b} \frac{1}{||\vec{w}||_2} . 1 = \min_{w,b} ||w||_2$$

$$\min_{w,b} ||w||_2$$
$$such\ that \quad \forall i\ y_i(W^T x_i + b) \geq 1$$

$$\min_{w,b} w^T w$$
$$such\ that \quad \forall i\ y_i(W^T x_i + b) \geq 1$$

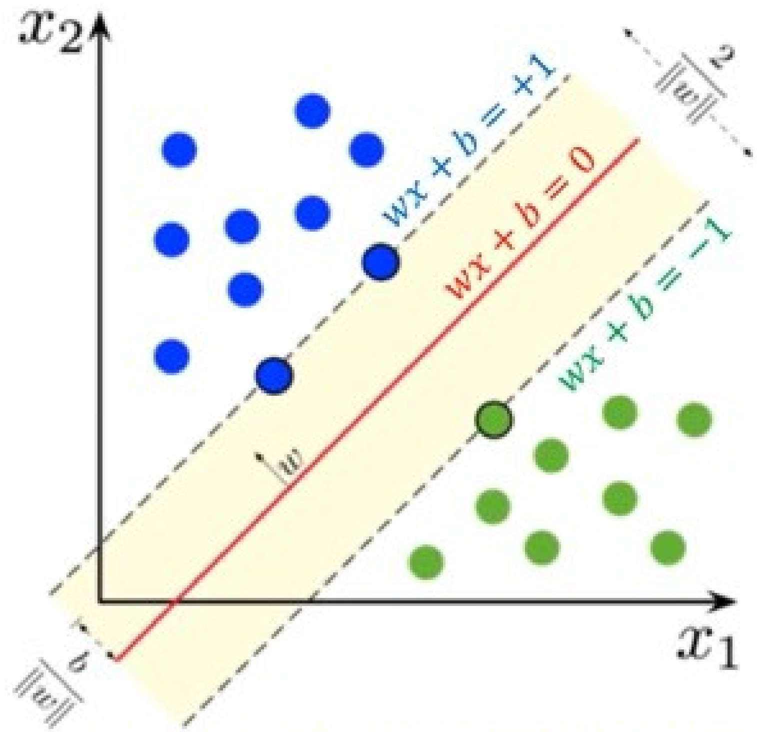


Image ref: https://en.wikipedia.org/wiki/Support-vector_machine

Soft Margin SVMs

- For low-dimensional data, there might be no separating hyperplane between the two classes
- In this case, there is no solution to the optimization problems for the hard margin SVMs.
- **Solution:** Allow the constraints to be slightly violated using slack variables:

C determines how amplified the slack would be!

Width of the "road" is no cumulative number of actual "road" plus some slack!

$$\min_{w,b} w^T w + C \sum_{i=1}^{n} \xi_i$$

$$\text{such that} \quad \forall i \; y_i\left(W^T x_i + b\right) \geq 1 - \xi_i$$

$$\forall \xi_i \geq 0$$

Allow point i to be on the road

It will be forced to be positive integer

Soft Margin SVMs

$$\min_{w,b} w^T w + C \sum_{i=1}^{n} \xi_i$$

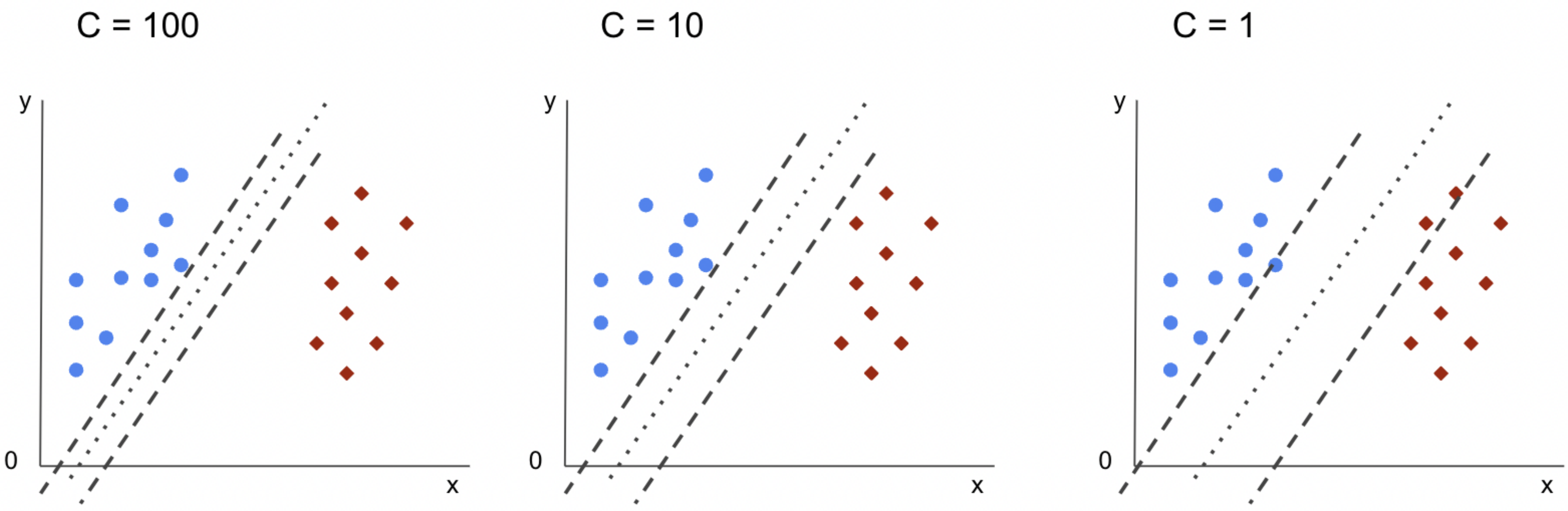
$$\text{such that} \quad \forall i \; y_i(W^T x_i + b) \geq 1 - \xi_i$$

$$\forall \xi_i \geq 0$$

- The slack variable ξ_i allows the input x_i to be closer to the hyperplane (or even on the wrong side)
  - There is a penalty in the objective function for such “slack”
  - **If C is very small:** the SVM becomes very lenient and may sacrifice some points to obtain a simpler (i.e., lower $||w||_2$) solution.
  - **If C is very large:** the SVM becomes very strict and tries to get all pints to beo n the correct side of the hyperplane.

Soft Margin Example

- C adds penalty to each misclassified point.
- If the C value is small, then essentially, the penalty for misclassified points is also small, thus resulting in a larger margin-based boundary.
- If the C value is large, then SVM tries to minimize the number of misclassified points by reducing the margin width.



Unconstrained Formulation

- Assuming $C \neq 0$

Slack amount equals how closer the data point is than 1 from hyperplane

$$\boldsymbol{\xi_i = \begin{cases} 1 - y_i\left(W^T x_i + b\right) & if\ y_i\left(W^T x_i + b\right) < 1 \\ 0 & if\ y_i\left(W^T x_i + b\right) \geq 1 \end{cases}}$$

No slack if the data point i is already far away from hyperplane.

- Which is equivalent to:

$$\boldsymbol{\xi_i = max\left(1 - y_i\left(W^T x_i + b\right), 0\right)}$$

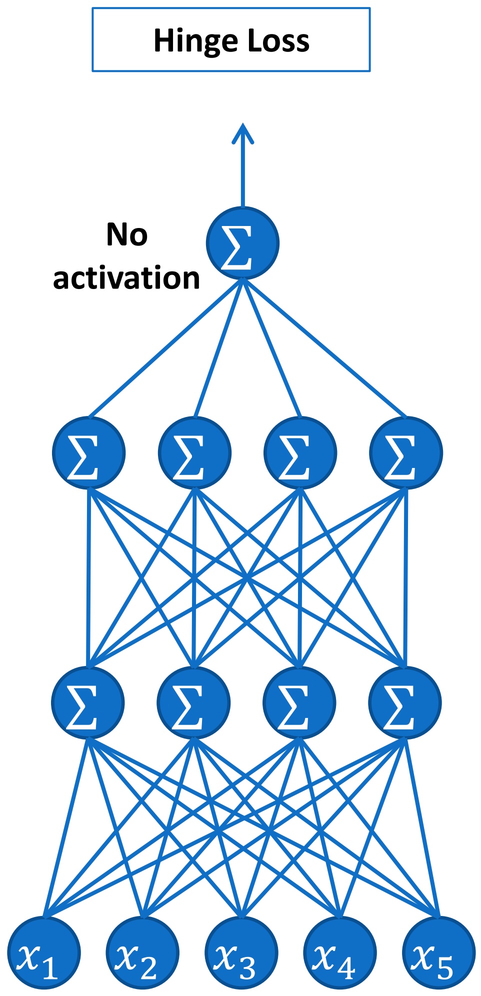
Unconstrained Formulation

$$\boxed{\xi_i = max\left(1 - y_i(W^T x_i + b), 0\right)}$$

- Plugging this closed form into the objective of our SVM optimization problem, we obtain the following unconstrained version as loss function and regularizer:

$$\underbrace{\min_{w,b} w^T w}_{\text{Regularizer}} + \underbrace{C \sum_{i=1}^{n} max\left(1 - y_i(W^T x_i + b), 0\right)}_{\text{Loss (Hinge-loss)}}$$

- This formulation allows us to optimize the SVM parameters (w, b) just like logistic regression (e.g., through gradient descent)

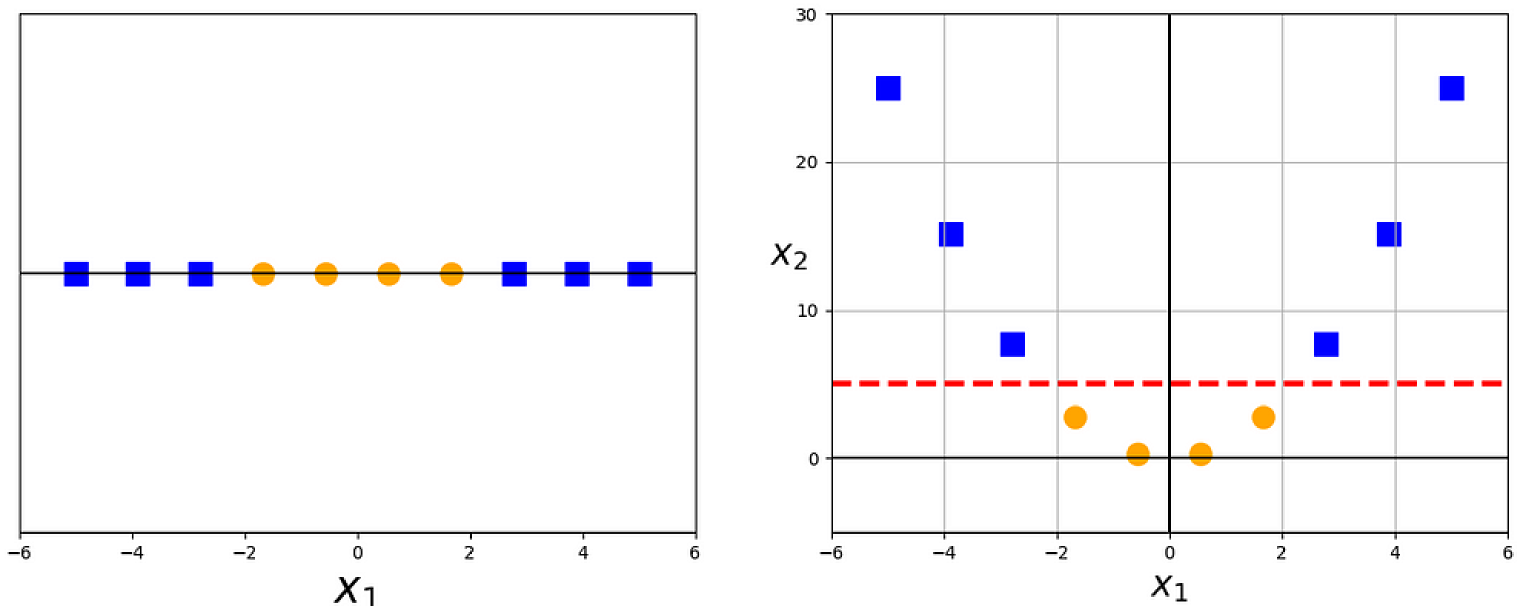

Hinge Loss
No activation
Σ
Σ Σ Σ Σ
Σ Σ Σ Σ
x_1 x_2 x_3 x_4 x_5

What if data is not linearly separable?

Kernel SVM

Kernel Functions

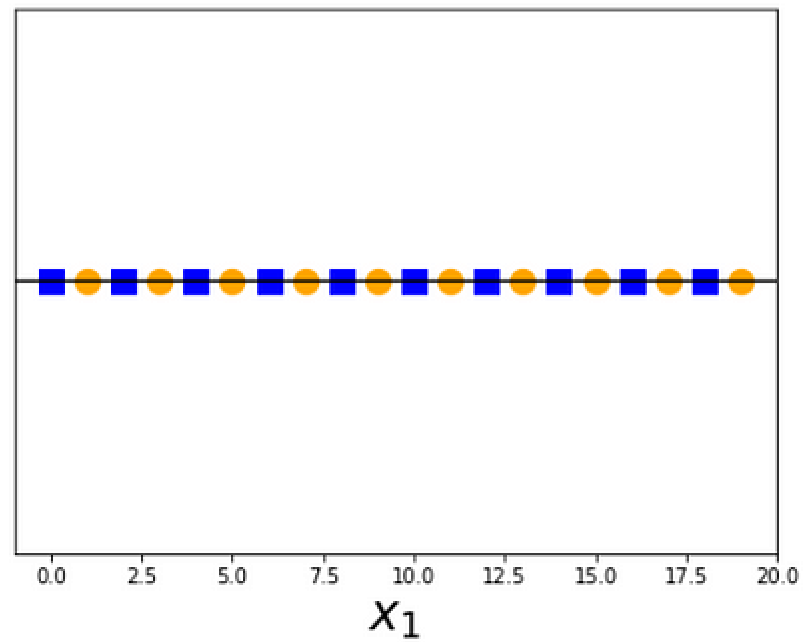
- Kernel Functions Φ are applied to increase the dimensions of the data, thus making it linearly separable, before applying SVM.

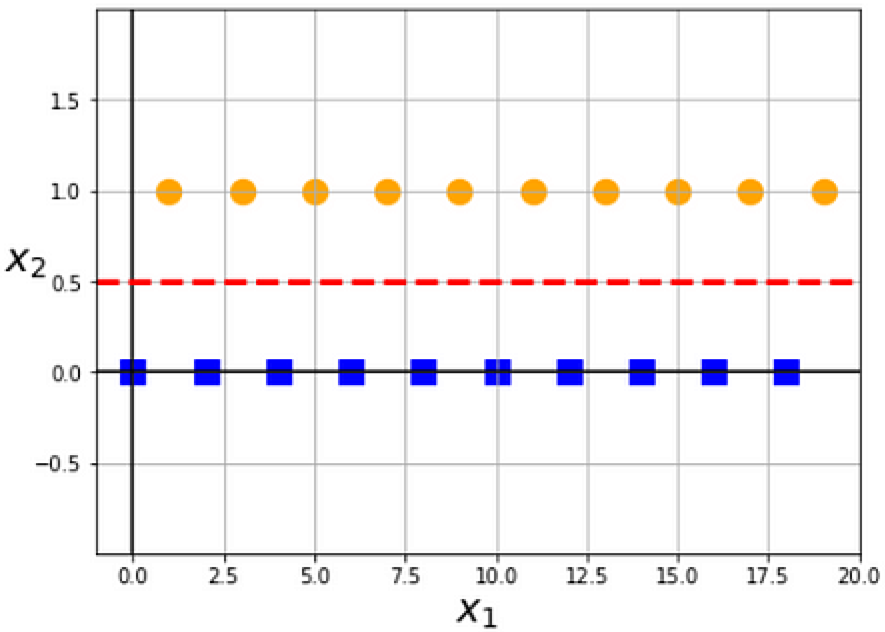


$$\Phi(X) = X^2$$

Kernel Functions

- Class A = even numbers
- Class B = odd numbers

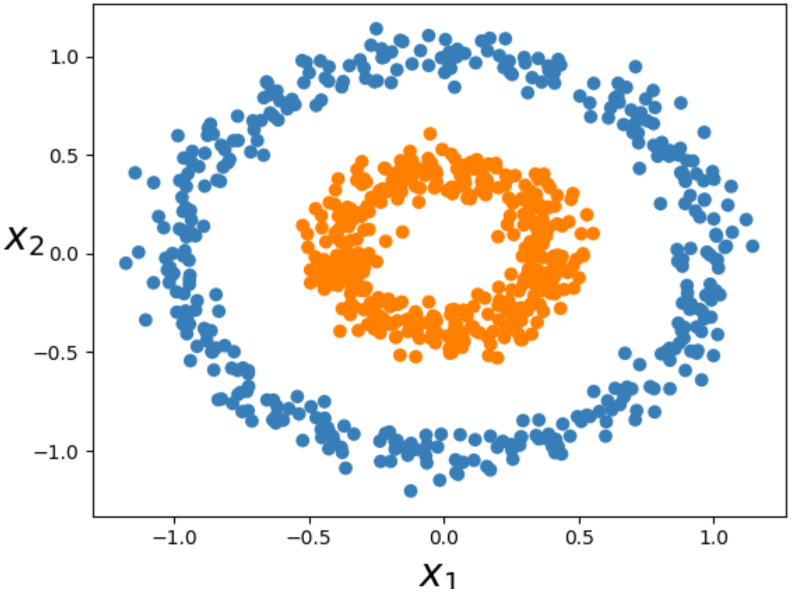


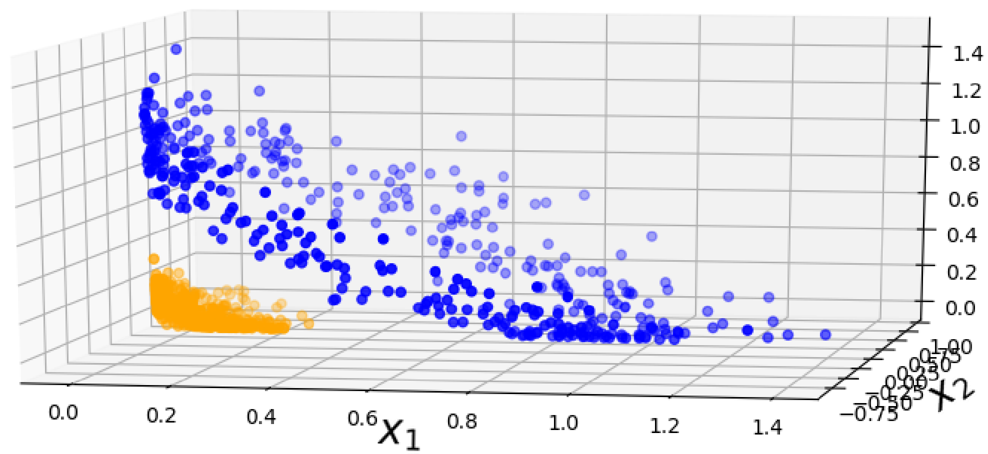


$$\boxed{\Phi(X) = X \bmod 2}$$

Credit: https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

Kernel Functions





$$\Phi(X) = \begin{pmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{pmatrix}$$

2nd-degree polynomial

Side Note

- "Kernel Trick" allows using kernels, without actually computing them!
- Example: 2nd degree polynomial mapping

$$\phi(\mathbf{a})^T \cdot \phi(\mathbf{b}) = \begin{pmatrix} {a_1}^2 \\ \sqrt{2}\, a_1 a_2 \\ {a_2}^2 \end{pmatrix}^T \cdot \begin{pmatrix} {b_1}^2 \\ \sqrt{2}\, b_1 b_2 \\ {b_2}^2 \end{pmatrix} = {a_1}^2 {b_1}^2 + 2 a_1 b_1 a_2 b_2 + {a_2}^2 {b_2}^2$$

$$= (a_1 b_1 + a_2 b_2)^2 = \left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^T \cdot \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right)^2 = \left(\mathbf{a}^T \cdot \mathbf{b} \right)^2$$

Credit: https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

Famous Kernels

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^p; \text{ polynomial kernel.} \tag{1.22}$$

$$K(x_i, x_j) = e^{\frac{-1}{2\sigma^2}(x_i - x_j)^2}; \text{ Gaussian kernel; Special case of Radial Basis Function.}$$

$$K(x_i, x_j) = e^{-\gamma(x_i - x_j)^2}; \text{ RBF Kernel}$$

$$K(x_i, x_j) = \tanh(\eta\, x_i \cdot x_j + \nu); \text{ Sigmoid Kernel; Activation function for NN.}$$

Credit: https://towardsdatascience.com/understanding-support-vector-machine-part-2-kernel-trick-mercers-theorem-e1e6848c6c4d

A Comprehensive Lecture

- A great explanation on the concepts
- https://www.youtube.com/watch?v=ny1iZ5A8ilA&list=PLUE9cBml08yjxtiDUgPRIsSL_f8K7zBKd&index=72&t=180s&ab_channel=IntuitiveMachineLearning

- Smaller explanation
- https://www.youtube.com/watch?v=Q7vT0--5VII&list=PLUE9cBml08yjxtiDUgPRIsSL_f8K7zBKd&index=70&ab_channel=VisuallyExplained

Project: Explanation

- If you will do fine-tuning of existing pretrained models available on your dataset, you will get a higher performance!

- List of available models with TensorFlow
  - https://keras.io/api/applications/

Assignment 4: Task 2

- Implement SVM Classifier on your me/not me dataset

- Implement SVM Classifier on your multi-class labels
  - How would a binary classifier work on multi-class labels?

- Use best possible value of C
- Use best possible kernel

- Compute relevant metrics on test split

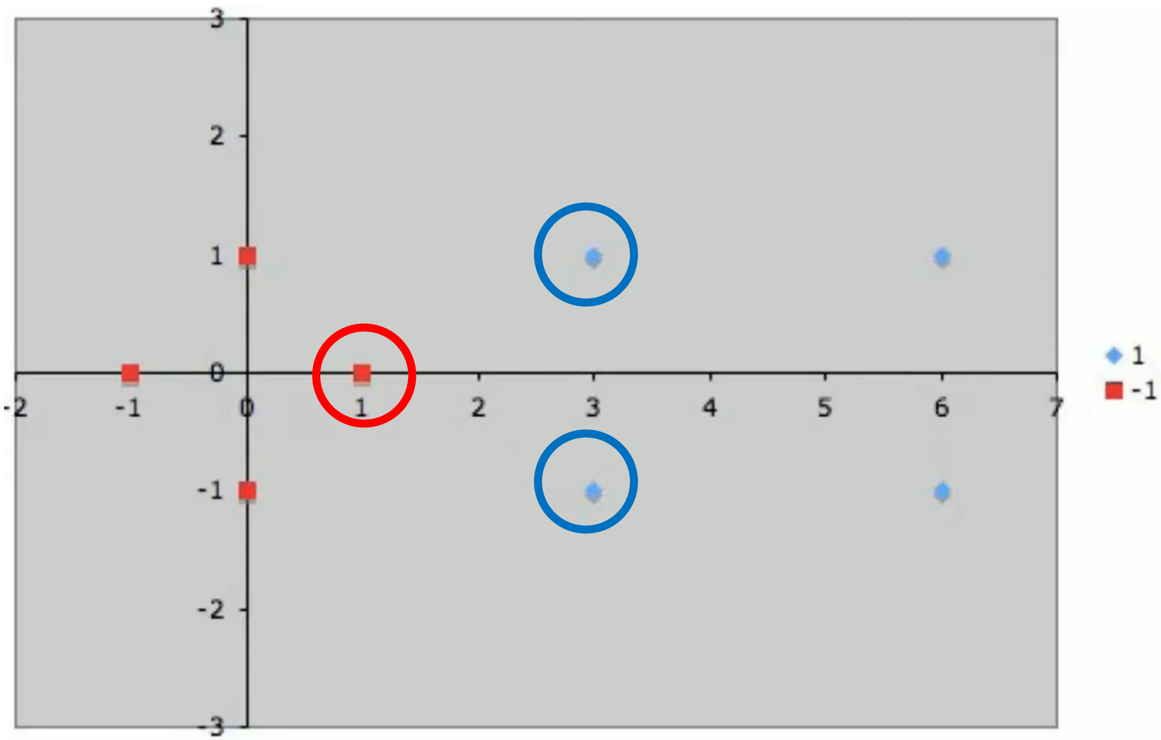
Linear SVM Example

Linear SVM: Example

- Suppose you are given the following 2D dataset with labels

Step 1: Identify Support Vectors

x_1	x_2	y
3	1	1
3	−1	1
6	1	1
6	−1	1
1	0	−1
0	1	−1
0	−1	−1
−1	0	−1



Linear SVM: Example

- Suppose you are given the following 2D dataset with labels

Step 2: Augment Each SV with Bias

x_1	x_2	b	y
3	1	1	1
3	−1	1	1
6	1		1
6	−1		1
1	0	1	−1
0	1		−1
0	−1		−1
−1	0		−1

$$S_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad S_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad S_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$$

$$\widetilde{S_1} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \widetilde{S_2} = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \quad \widetilde{S_3} = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

−1 1 1

Linear SVM: Example

- Suppose you are given the following 2D dataset with labels

Step 3: Write the equations needed to calculate the weight vector.

Solve for unknown value α for dot product of every support vector with all other support vectors!

$$\widetilde{S_1} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \qquad \widetilde{S_2} = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \qquad \widetilde{S_3} = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$-1 \qquad\qquad 1 \qquad\qquad 1$$

For each SV, do
$\alpha_i \, \widetilde{s_i} . \widetilde{S}$

$$\alpha_1 \widetilde{S_1} \cdot \widetilde{S_1} + \alpha_2 \widetilde{S_2} \cdot \widetilde{S_1} + \alpha_3 \widetilde{S_3} \cdot \widetilde{S_1} = -1$$

$$\alpha_1 \widetilde{S_1} \cdot \widetilde{S_2} + \alpha_2 \widetilde{S_2} \cdot \widetilde{S_2} + \alpha_3 \widetilde{S_3} \cdot \widetilde{S_2} = +1$$

$$\alpha_1 \widetilde{S_1} \cdot \widetilde{S_3} + \alpha_2 \widetilde{S_2} \cdot \widetilde{S_3} + \alpha_3 \widetilde{S_3} \cdot \widetilde{S_3} = +1$$

Linear SVM: Example

- Suppose you are given the following 2D dataset with labels

Step 4: Put in known values of SVs

$$\widetilde{S_1} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \widetilde{S_2} = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \quad \widetilde{S_3} = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$\alpha_1 \widetilde{S_1} \cdot \widetilde{S_1} + \alpha_2 \widetilde{S_2} \cdot \widetilde{S_1} + \alpha_3 \widetilde{S_3} \cdot \widetilde{S_1} = -1$$

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = -1$$

$$\alpha_1 \widetilde{S_1} \cdot \widetilde{S_2} + \alpha_2 \widetilde{S_2} \cdot \widetilde{S_2} + \alpha_3 \widetilde{S_3} \cdot \widetilde{S_2} = +1$$

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = +1$$

$$\alpha_1 \widetilde{S_1} \cdot \widetilde{S_3} + \alpha_2 \widetilde{S_2} \cdot \widetilde{S_3} + \alpha_3 \widetilde{S_3} \cdot \widetilde{S_3} = +1$$

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = +1$$

Linear SVM: Example

Step 5: Perform dot product

$$\alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix} \cdot \begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix} \cdot \begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}1\\0\\1\end{pmatrix} = -1$$

$$\alpha_1(1+0+1) + \alpha_2(3+0+1) + \alpha_3(3+0+1) = -1$$

$$\alpha_1(2) + \alpha_2(4) + \alpha_3(4) = -1$$

$$\alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix} \cdot \begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix} \cdot \begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}3\\1\\1\end{pmatrix} = +1$$

$$\alpha_1(3+0+1) + \alpha_2(9+1+1) + \alpha_3(9-1+1) = +1$$

$$\alpha_1(4) + \alpha_2(11) + \alpha_3(9) = +1$$

$$\alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix} \cdot \begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}3\\1\\1\end{pmatrix} \cdot \begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}3\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}3\\-1\\1\end{pmatrix} = +1$$

$$\alpha_1(3+0+1) + \alpha_2(9-1+1) + \alpha_3(9+1+1) = +1$$

$$\alpha_1(4) + \alpha_2(9) + \alpha_3(11) = +1$$

Linear SVM: Example

Step 6: Solve simultaneous equations

$\alpha_1(2) + \alpha_2(4) + \alpha_3(4) = -1$	$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$
$\alpha_1(4) + \alpha_2(11) + \alpha_3(9) = +1$	$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$
$\alpha_1(4) + \alpha_2(9) + \alpha_3(11) = +1$	$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = +1$

How to solve simultaneous equations:
https://www.youtube.com/watch?v=NlpykbGDzF8&ab_channel=tecmath

Online Calculator

Side Notes:

- **Simultaneous equations are two or more algebraic equations that share variables.**
- **Linear simultaneous equations contain terms that are raised to a power that is no higher than one.**
- **Can be solved using elimination method.**

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$$

$$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = +1$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$$

$$-(4\alpha_1 + 9\alpha_2 + 11\alpha_3 = +1)$$

$$\Rightarrow \quad 2a_2 - 2a_3 = 0$$

$$2a_2 - 2a_3 = 0$$

$$2(-3a_2 - 1a_3 = -3)$$

$$2a_2 - 2a_3 = 0$$

$$-6a_2 - 2a_3 = -6$$

$$2a_2 - 2a_3 = 0$$

$$-(-6a_2 - 2a_3 = -6)$$

$$\Rightarrow \quad 8a_3 = 6$$

$$2(2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1)$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$$

$$4\alpha_1 + 8\alpha_2 + 8\alpha_3 = -2)$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$$

$$4\alpha_1 + 8\alpha_2 + 8\alpha_3 = -2)$$

$$-(4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1)$$

$$\Rightarrow \quad -3a_2 - 1a_3 = -3$$

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$$

$$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = +1$$

$$2a_2 - 2a_3 = 0$$

$$-3a_2 - 1a_3 = -3$$

$$8a_3 = 6$$

$$8a_3 = 6$$

$$a_3 = 6/8 = 3/4 = 0.75$$

$$2a_2 - 2a_3 = 0$$

$$2a_2 - 2(0.75) = 0$$

$$2a_2 - 1.5 = 0$$

$$2a_2 = 0 + 1.5$$

$$2a_2 = 1.5$$

$$a_2 = 1.5/2 = 0.75$$

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$$

$$2a_1 + 4(0.75) + 4(0.75) = -1$$

$$2a_1 + 3 + 3 = -1$$

$$2a_1 = -1 - 3 - 3$$

$$2a_1 = -7$$

$$a_1 = -7/2 = -3.5$$

Linear SVM: Example

Step 7: Calculate Weight Vector

$$\tilde{w} = \sum_i \alpha_i \tilde{S}_i$$

$$\tilde{S_1} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \qquad \tilde{S_2} = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \qquad \tilde{S_3} = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$a_1 = -3.5$$

$$a_2 = 0.75$$

$$a_3 = 0.75$$

$$\tilde{w} = -3.5 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$\tilde{w} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}$$

Linear SVM: Example

Step 7: Interpretation

Recall that last entry in $\tilde{w}$ is a result of bias value which can be used as hyperplane offset

$$\tilde{w} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}$$

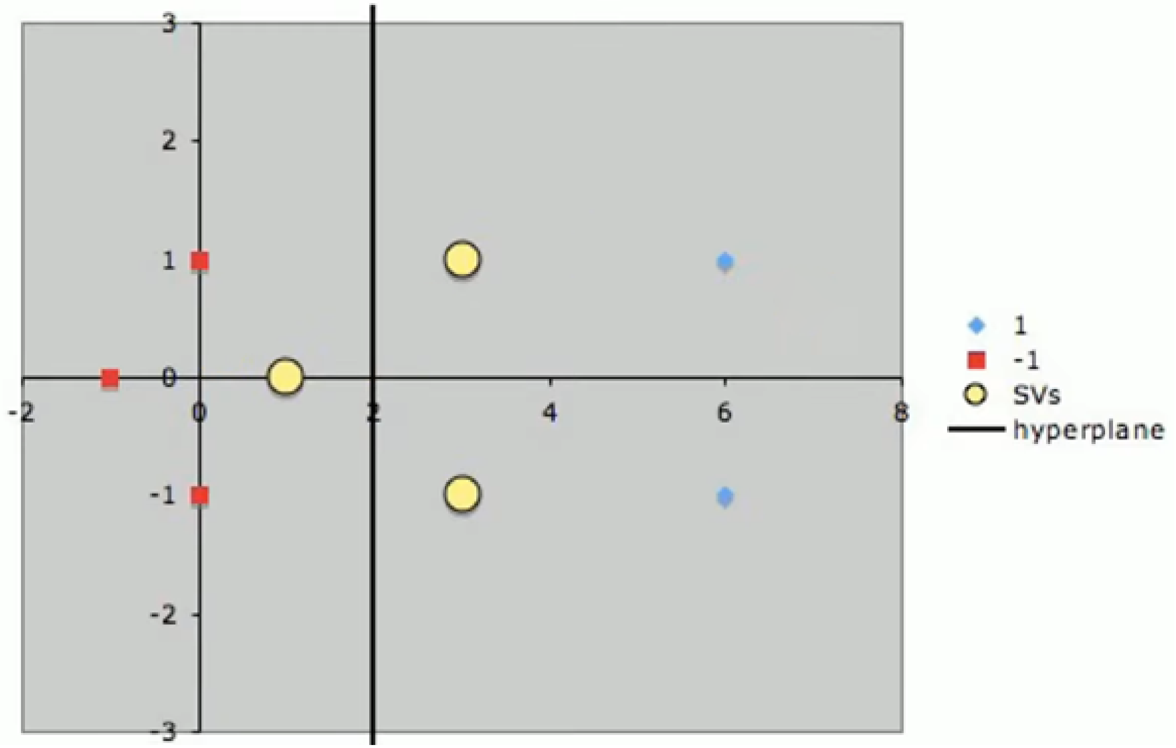
$$y = wx + b$$

Where $\tilde{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $b - 2 = 0$

Line is vertical when $\tilde{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

Line is horizontal when $\tilde{w} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

What if $\tilde{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$?



Book Reading

- Murphy – Chapter 1, Chapter 14
- Handouts