why not change the world?®

Support Vector Machines (SVM) for classification

Ahmed Eleish

Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

November 5th 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

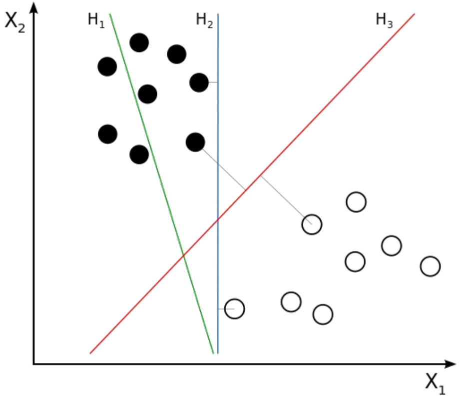
Support Vector Machines

- Rationale
- Hyperplanes, Margins and Support vectors
- Classification using SVM
- Linear Separability of classes (or not)
  - Transformation of input
  - Kernel Trick
- Soft Margin SVM
- SVM Regression

Rationale

- In *p*-dimensional space, if data points (*p*-dimensional vectors) belonging to 2 different classes can be separated by a (*p*-1)-dimensional hyperplane, this hyperplane can be used as a linear classifier.

- Example: in 2d space, a line could be linear classifier..



- The hyperplane representing the largest separation or "margin" between the classes maximizes the distance to the nearest data point from each class.

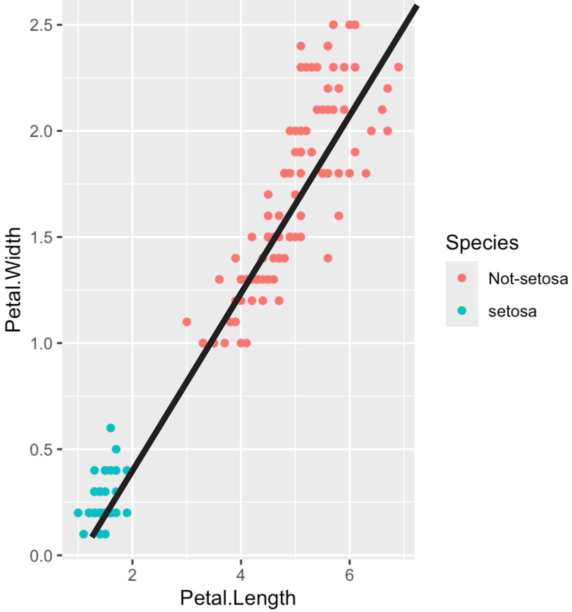
- SVM can be used for classification, regression and outlier detection.

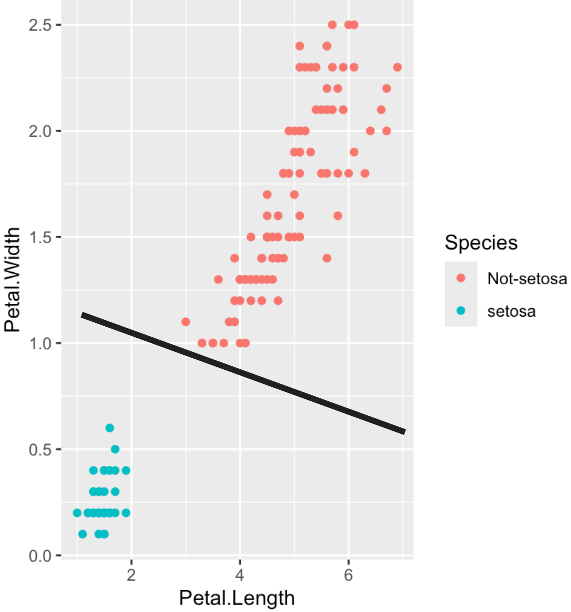
Hyperplane

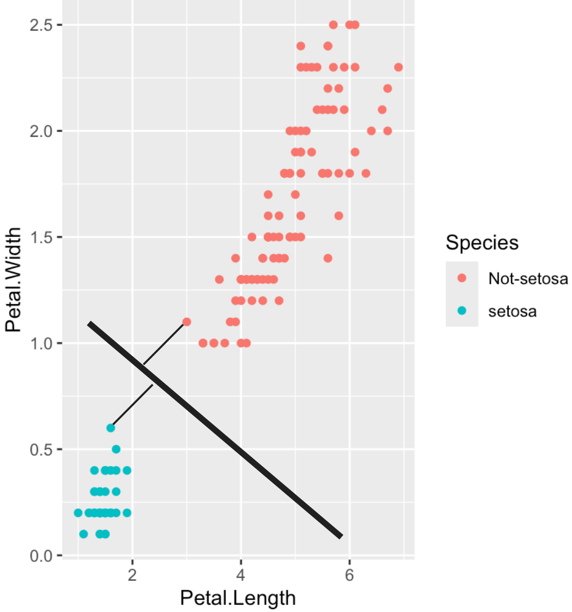
- A hyperplane is a plane of dimension *p-1* in a p dimensional space
- “a flat hypersurface, a subspace whose dimension is one less than that of the ambient space”
- “any codimension-1 vector subspace of a vector space”

https://en.wikipedia.org/wiki/Hyperplane
https://mathworld.wolfram.com/Hyperplane.html

Hyperplanes





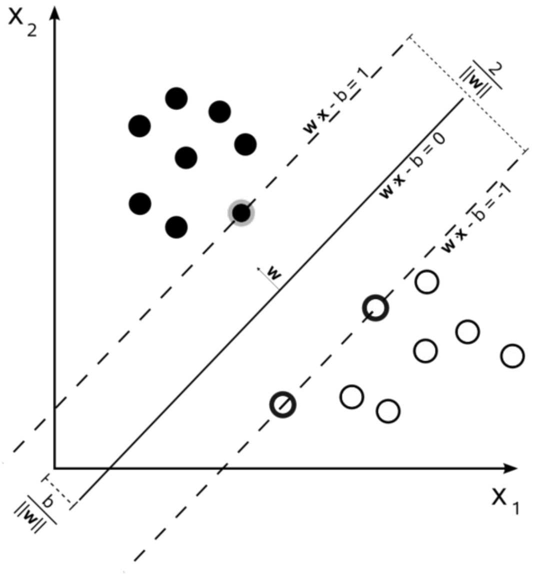


Margin

- The distance between the hyperplane (decision boundary) and the nearest points from each class.

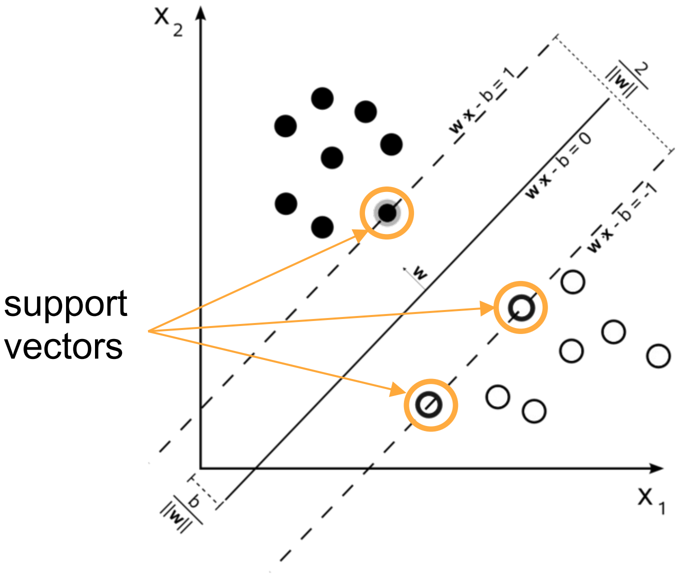
- larger margin = greater confidence in the classifier

- SVMs find the hyperplane that maximizes the margin
    - “maximum-margin classifiers”



Support Vectors

- The points closest to the decision boundary.
- They determine the position and orientation of the hyperplane, i.e. define the decision boundary.
- They are used to calculate the margin.



Support Vector Machines

- Given training dataset of points (x_1,y_1) where y_i is equal to 1 or -1

* Find the maximum-margin-hyperplane that divides the points x_i for which $y_i = 1$ from the points for which $y_i = -1$

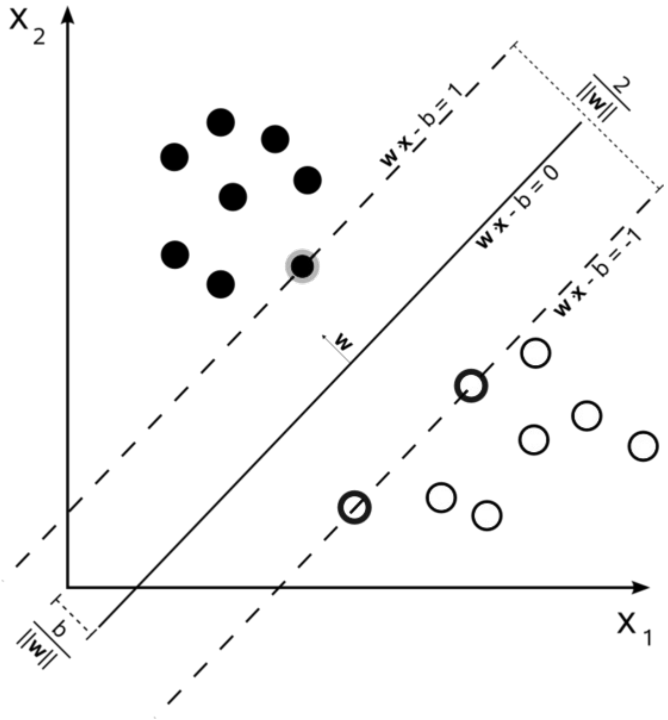
Hyperplane:

$$W^T X - b = 0$$

- To find *W* and *b*:

$$\underset{\mathbf{w},\, b}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

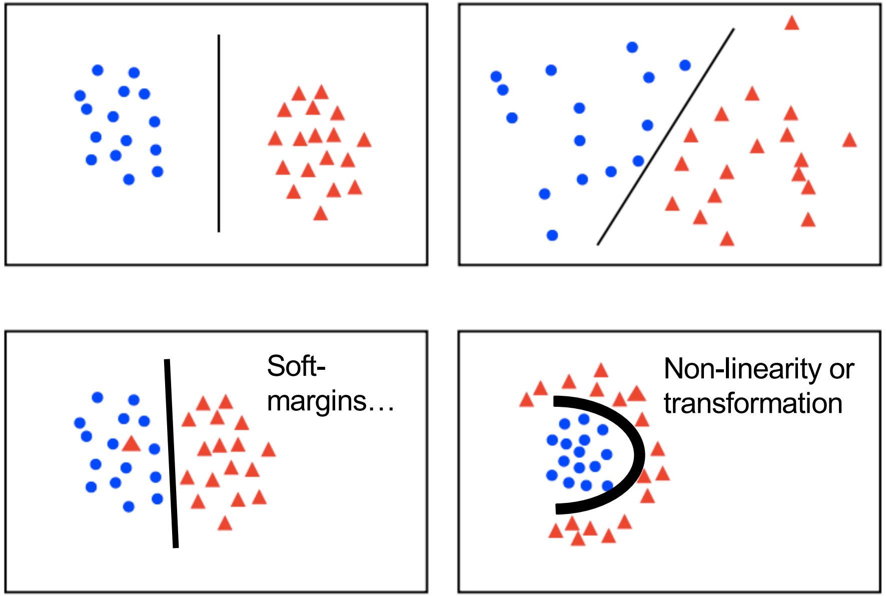
$$\text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 \quad \forall i \in \{1, \ldots, n\}$$



Linear Separability

linearly separable

not linearly separable



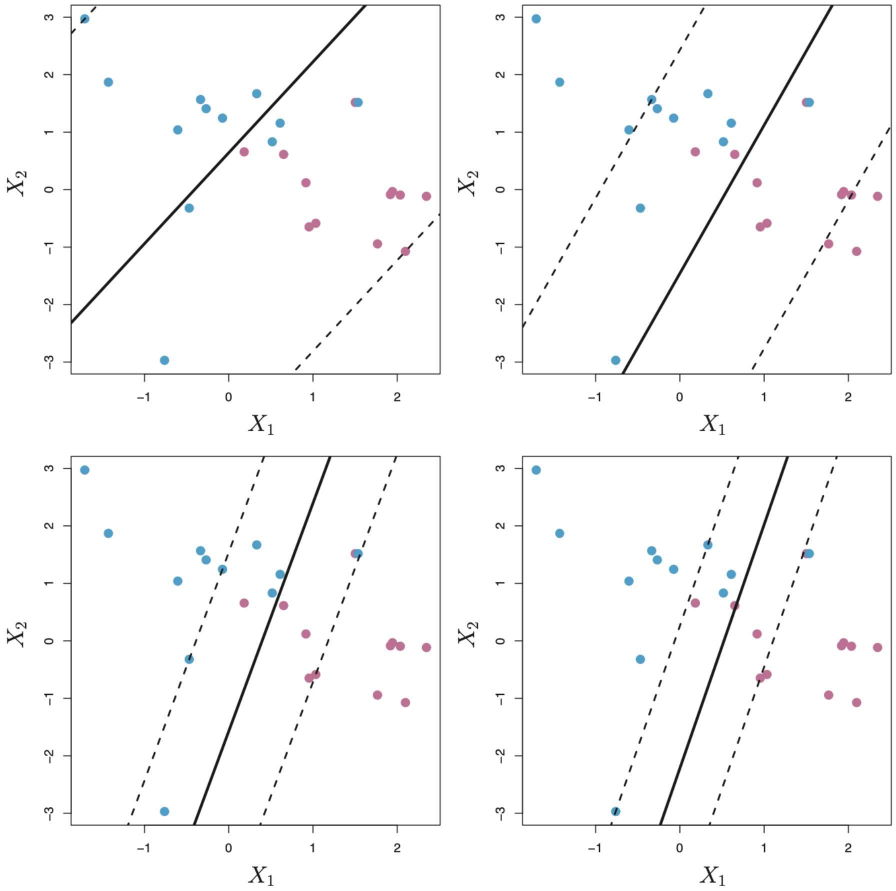
Soft-margin SVM

Allow for some margin violations controlled by the parameter *C*, the *regularization parameter*

$$\min_{w,b,z} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell} z_i$$

$$s.t \quad z_i \geq 1 - y_i\left(x_i \cdot w + b\right)$$

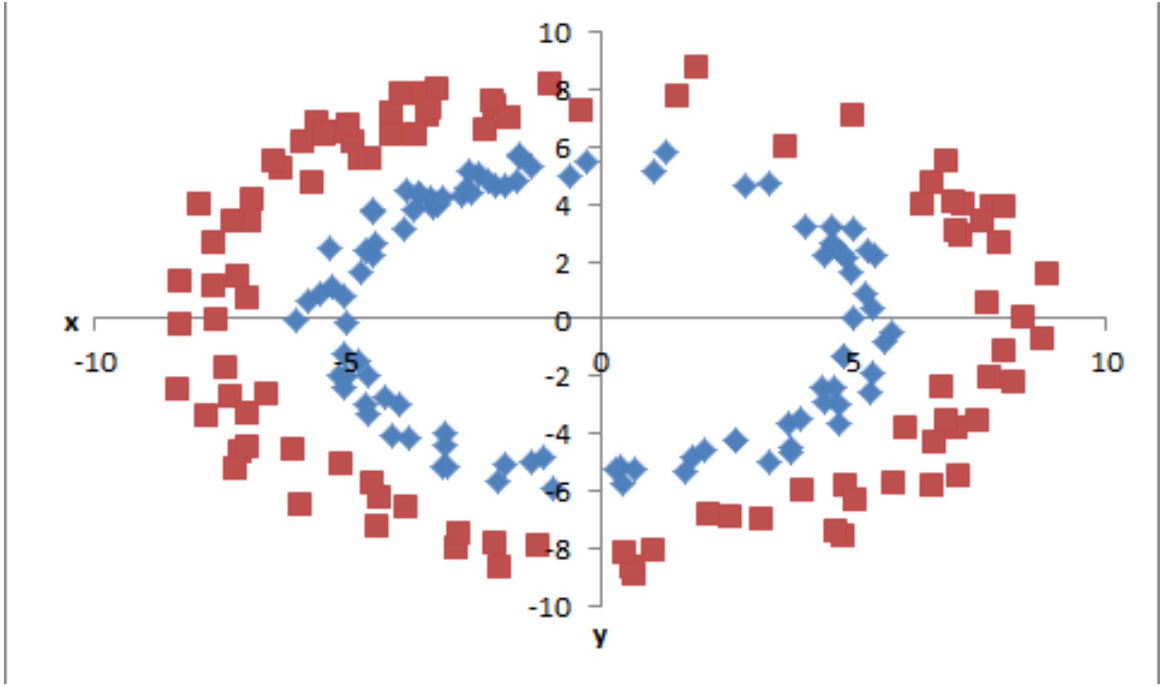
$$z_i \geq 0 \quad i = 1,..,N$$



Top left: Highest C value, decreasing C narrows the margin

Non-linearity

What to do??



Non-linearity

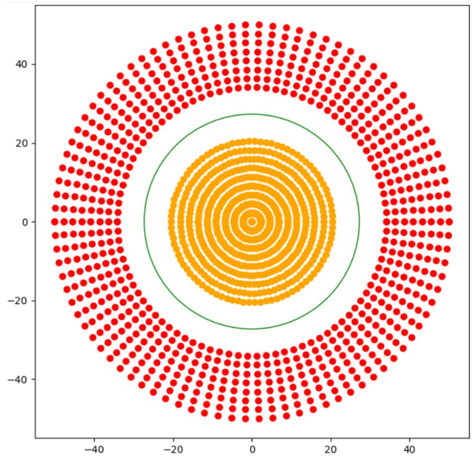
Transform the input:

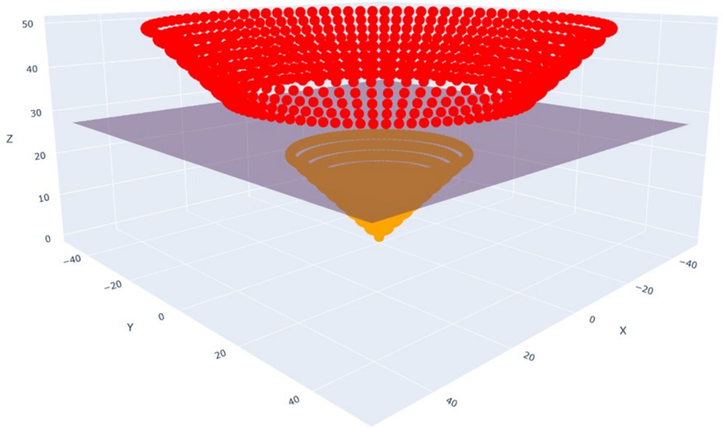
- Add a new dimension where the data are linearly separable

If are dataset contains variables X1, X2:
we can add X3 = f(X1,X2)

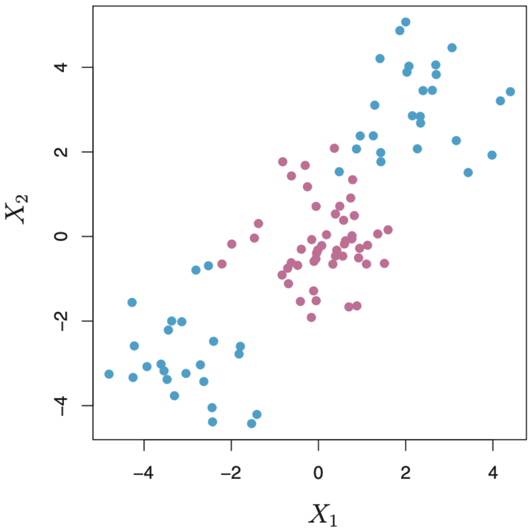
e.g. $X3 = (X1^2 + X2^2)^{(1/2)}$

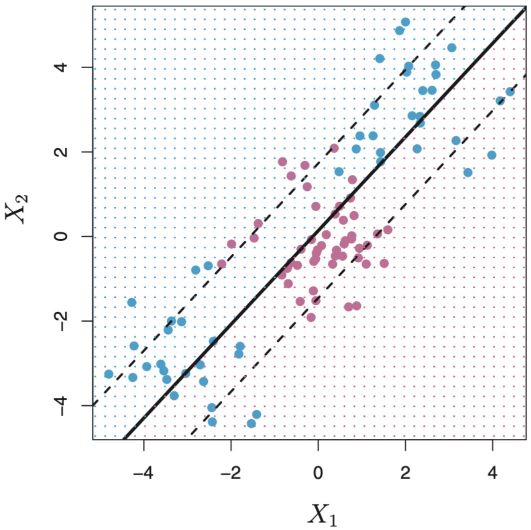
- Computationally expensive





More practically…





The Kernel Trick

- Instead of adding dimensions, find similarity between points.

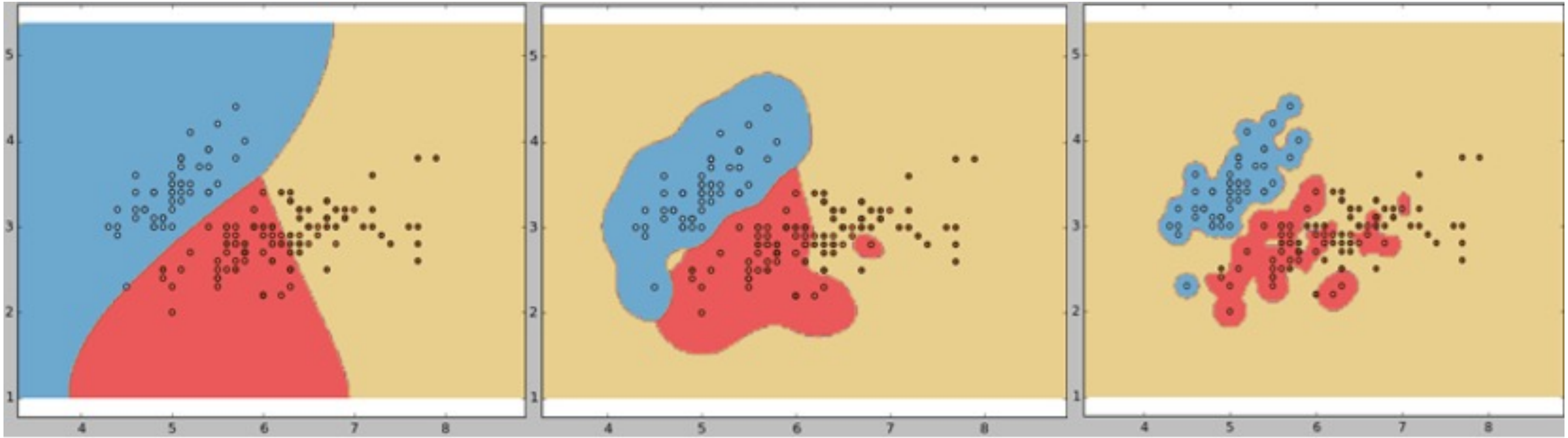
- similarity between points x_1 = ($x1_1$,$x2_1$) and x_2 = ($x1_2$,$x2_2$) using a function f(x_1,x_2)

e.g. Radial Basis Function (RBF) Kernel:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$$

γ is a hyperparameter controlling the *linearity* of the model

Gamma (γ)

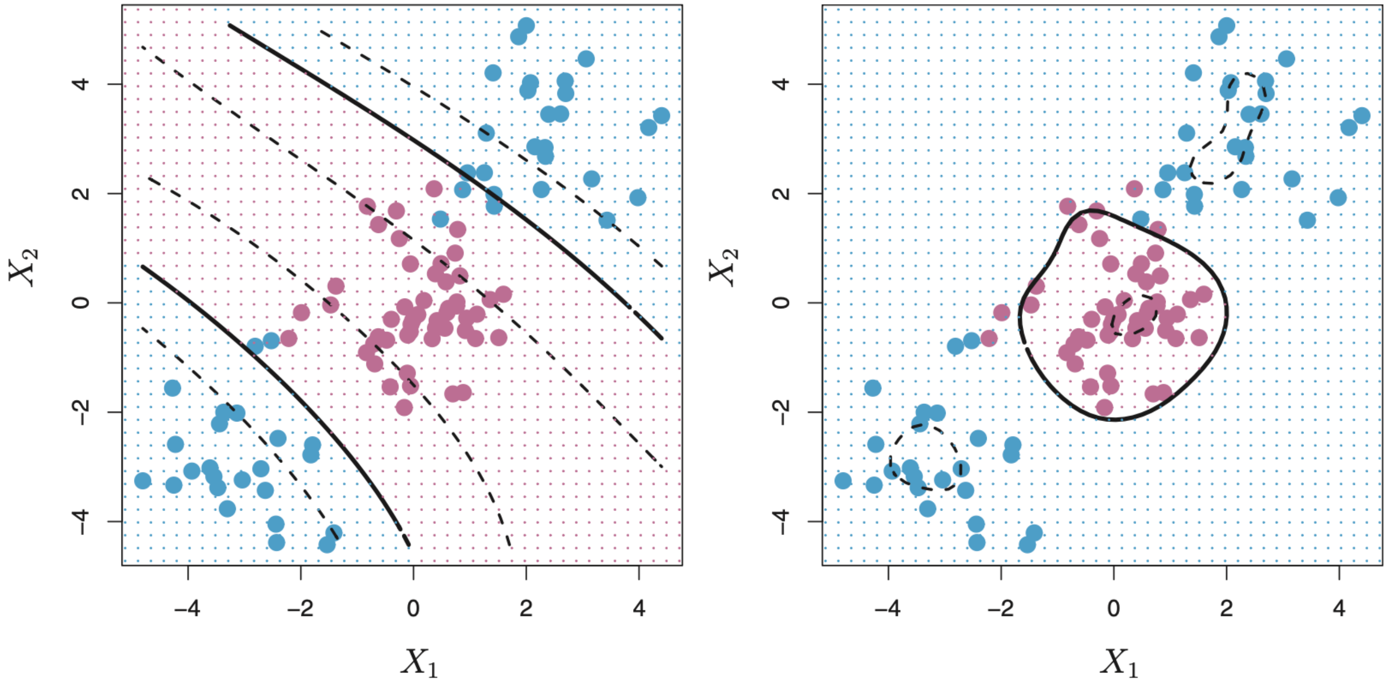


γ = 0.1 γ = 10 γ = 100

Kernels

- Polynomial Kernel
- Gaussian Kernel
- Gaussian RBF Kernel
- Laplace RBF Kernel
- Hyperbolic Tangent Kernel
- Sigmoid Kernel
- Bessel function of first kind Kernel
- ANOVA radial basis Kernel
- Linear Splines Kernel

Applying Kernels



In-class exercise

- https://rpi.box.com/s/0oqzzebiiu3z0yfqqnveclrkpnsf1c8q

Thanks!