BINUS UNIVERSITY

People
Innovation
Excellence

Course : COMP6577 – Machine Learning
Effective Period : February 2020

Support Vector Machine (SVM)

Session 19 & 20

Learning Outcome

- LO3: Student be able to experiment classification and clustering algorithm from given dataset

People
Innovation
Excellence

BINUS UNIVERSITY

Outline

- Introduction
- Linearly Separable Classes
- Non-separable classes
- Case Study

People
Innovation
Excellence

BINUS UNIVERSITY

People
Innovation
Excellence

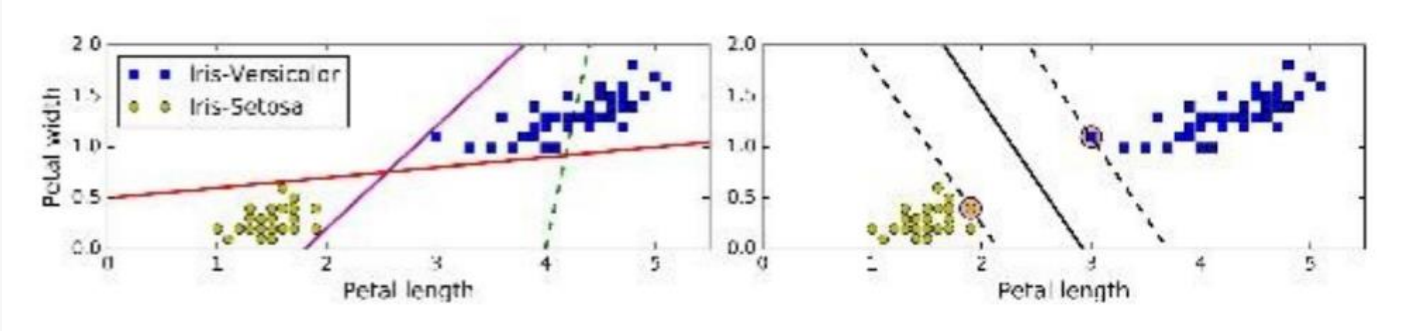
Introduction

- A Support Vector Machine (SVM) is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression, and even outlier detection.
- It is one of the most popular models in Machine Learning, and anyone interested in Machine Learning should have it in their toolbox. SVMs are particularly well suited for classification of complex but small- or medium-sized datasets.

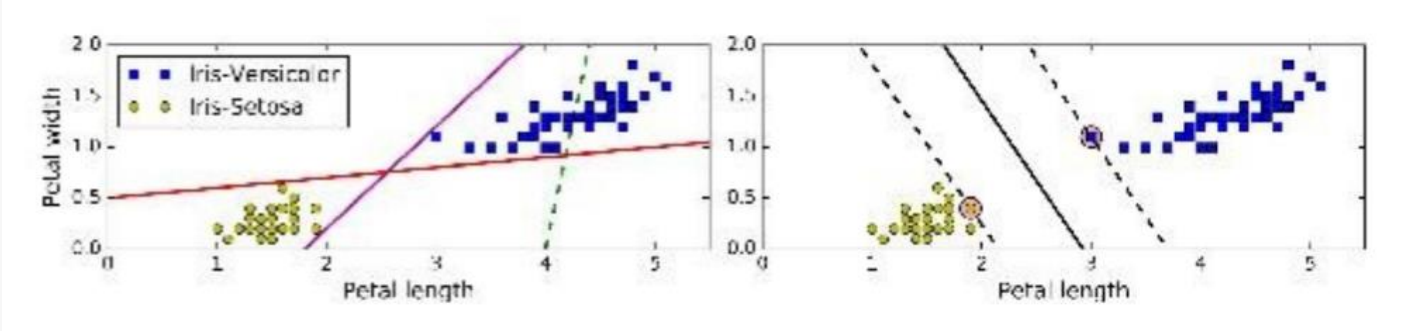
BINUS UNIVERSITY

Linear SVM Classification

- Look at the figure below.



- The two classes can clearly be separated easily with a straight line (they are linearly separable).
- The left plot shows the decision boundaries of three possible linear classifiers. The model whose decision boundary is represented by the dashed line is so bad that it does not even separate the classes properly. The other two models work perfectly on this training set, but their decision boundaries come so close to the instances that these models will probably not perform as well on new instances.

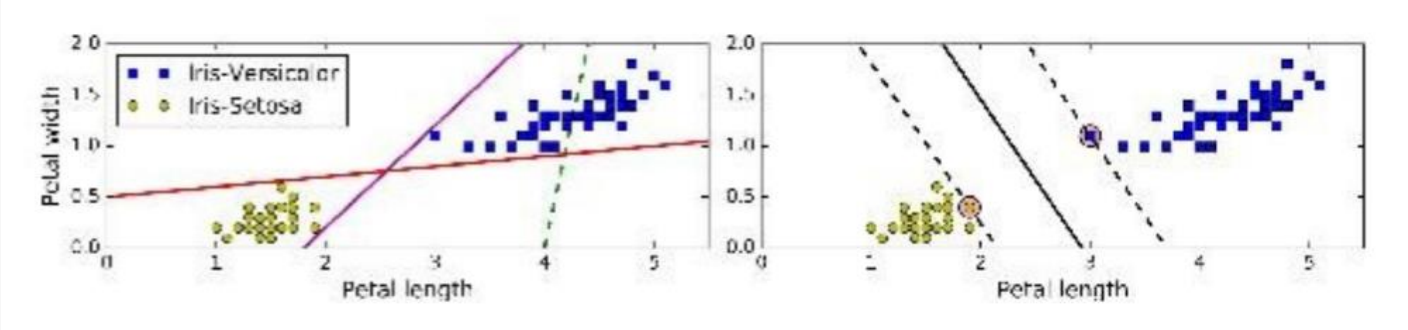


- In contrast, the solid line in the plot on the right represents the decision boundary of an SVM classifier; this line not only separates the two classes but also stays as far away from the closest training instances as possible.
- SVM classifier is fitting the widest possible street (represented by the parallel dashed lines) between the classes. This is called large margin classification.

BINUS UNIVERSITY

People
Innovation
Excellence

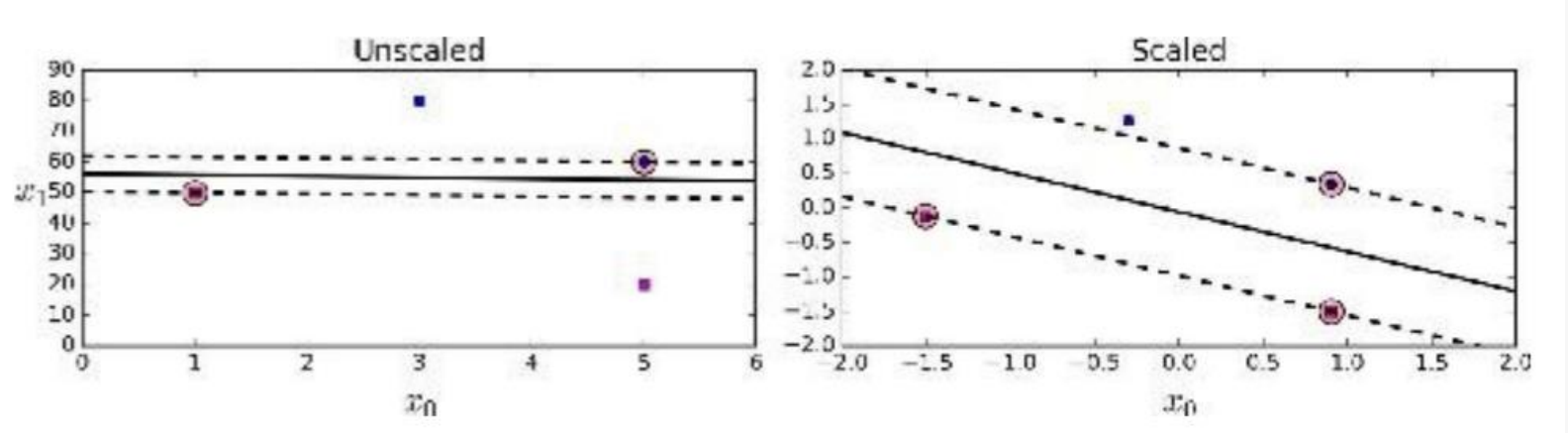
- Notice that adding more training instances “off the street” will not affect the decision boundary at all: it is fully determined (or “supported”) by the instances located on the edge of the street. These instances are called the support vectors.



BINUS UNIVERSITY

People
Innovation
Excellence

- SVMs are sensitive to the feature scales, as you can see in the previous figure: on the left plot, the vertical scale is much larger than the horizontal scale, so the widest possible street is close to horizontal. After feature scaling (e.g., using Scikit-Learn's **`StandardScaler`**), the decision boundary looks much better (on the right plot).



BINUS UNIVERSITY

Linearly Separable Classes: Maximum Margin Classifiers

- Assuming linearly separable classes, there is an infinity of linear classifiers that solve the classification task exactly, without committing errors on the training set.
- It is easy to see, and it will become apparent very soon that from this infinity of hyperplanes that solve the task, we can always identify a subset such as

$$y_n(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}_n + \theta_0) \geq 1, \quad n = 1, 2, \ldots, N$$

BINUS UNIVERSITY

People
Innovation
Excellence

- Hence, for linearly separable classes, the optimization task is equivalent to:

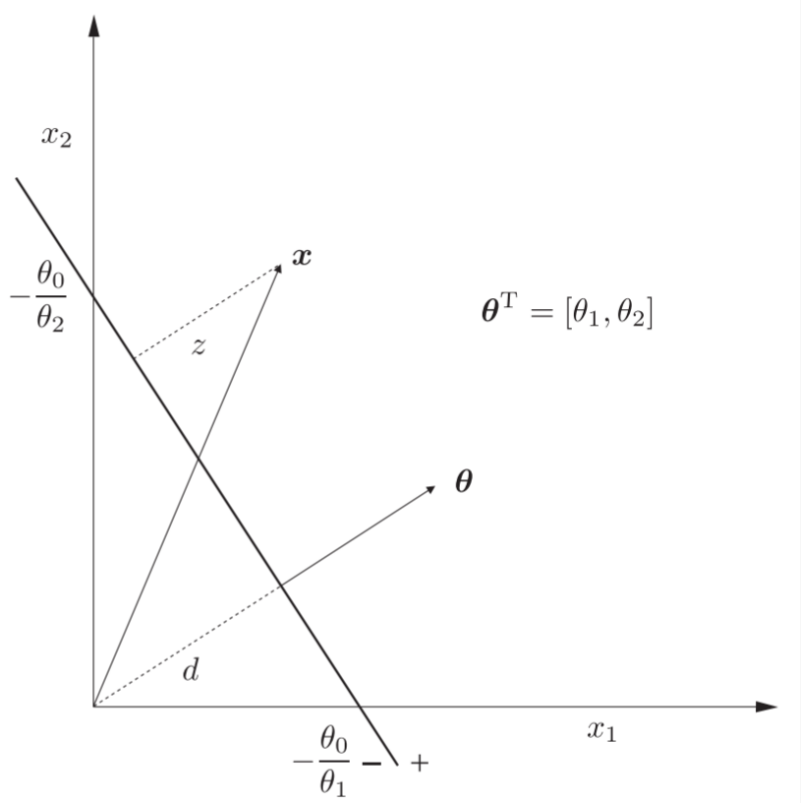
$$\begin{aligned} &\text{minimize with respect to } \boldsymbol{\theta} && \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ &\text{subject to} && y_n(\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0) \geq 1,\ n = 1, 2, \ldots, N. \end{aligned}$$

- In other words, from this infinity of linear classifiers, which can solve the task and classify correctly all training patterns, our optimization task selects the one that has minimum norm.
- the norm ||θ|| is directly related to the margin formed by the respective classifier.
- Each hyperplane $f(\boldsymbol{x}) = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x} + \theta_0 = 0$ d by the equation

BINUS UNIVERSITY

People
Innovation
Excellence

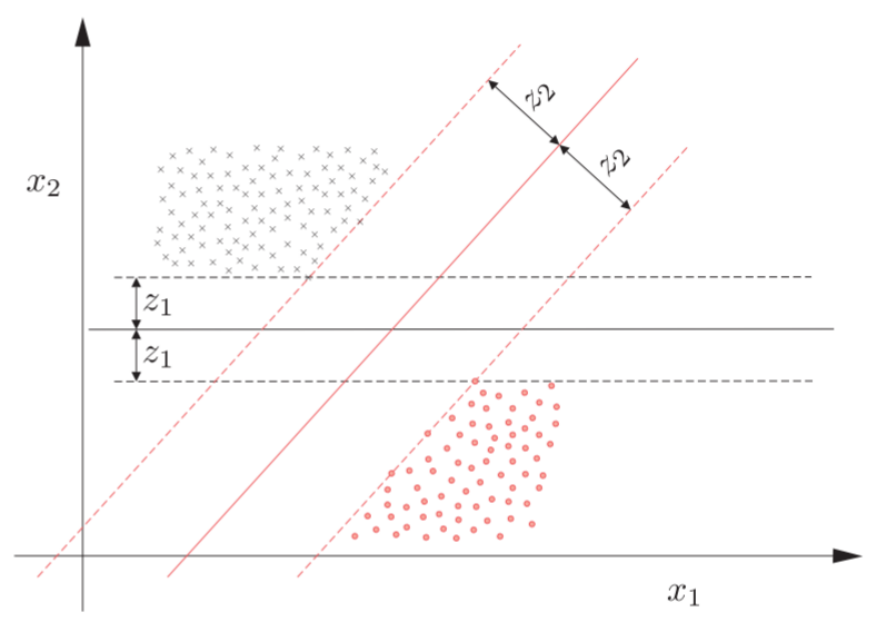
- From classical geometry, its direction in space is controlled by θ (which is perpendicular to the hyperplane) and its position is controlled by θ_0.



BINUS UNIVERSITY

People
Innovation
Excellence

- For each direction, θ, “red” and “gray,” the (linear) hyperplane classifier, $\theta^T x + \theta_0 = 0$, (full lines) is placed in between the two classes and normalized so that the nearest points from each class have a distance equal to one. The dotted lines, $\theta^T x + \theta_0 = \pm 1$, which pass through the nearest points, are parallel to the respective classifier, and define the margin. The width of the margin is determined by the direction of the corresponding classifier i

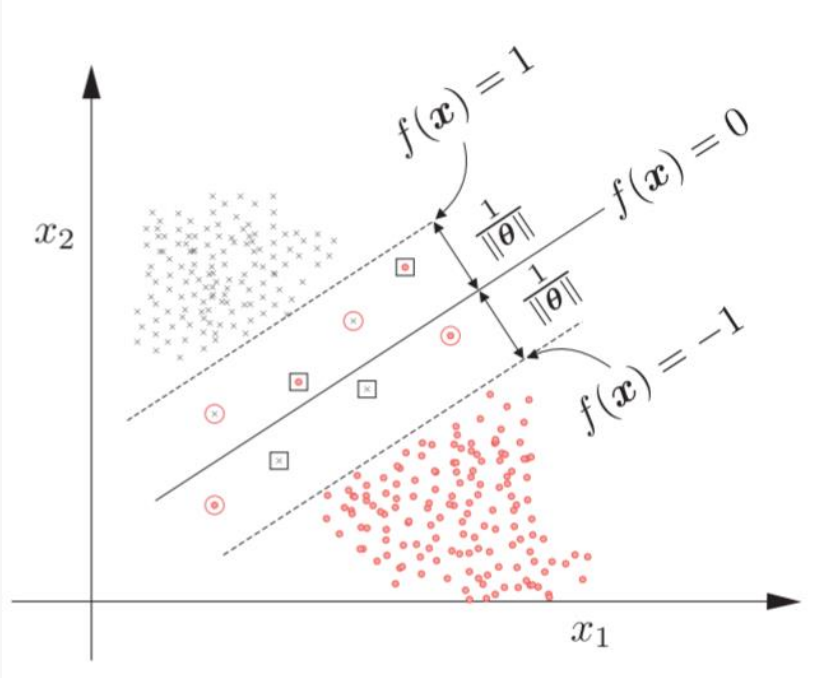


BINUS UNIVERSITY

People
Innovation
Excellence

Non-separable classes

- Although linear SVM classifiers are efficient and work surprisingly well in many cases, many datasets are not even close to being linearly separable.
- This figure shows the respective geometry for a linear classifier.



BINUS UNIVERSITY

People
Innovation
Excellence

- There are three types of points.
  - Points that lie on the border or outside the margin and in the correct side of the classifier, that is, $y_n f(x_n) \geq 1$. These points commit no (margin) error, that is $\xi n = 0$.
  - Points which lie on the correct side of the classifier, but lie inside the margin (circled points), that is, $0 < y_n f(x_n) < 1$. These points commit a margin error, and $0 < \xi n < 1$.
  - Points that lie on the wrong side of the classifier (points in squares), that is, $y_n f(x_n) \leq 0$. These points commit an error and $1 \leq \xi n$.

BINUS UNIVERSITY

- Our desire would be to estimate a hyperplane classifier, so as to **maximize the margin and at the same time to keep the number of errors (including margin errors) as small as possible**.
- This goal could be expressed via the optimization task in

$$\begin{aligned} &\text{minimize with respect to } \boldsymbol{\theta},\ \theta_0,\ \boldsymbol{\xi} && J(\boldsymbol{\theta},\boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{n=1}^{N}\xi_n, \\ &\text{subject to} && y_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0) \geq \rho - \xi_n, \end{aligned}$$

if in place of ξn we had the indicator function, I(ξn), where

$$I(\xi) = \begin{cases} 1, & \text{if } \xi > 0, \\ 0, & \text{if } \xi = 0. \end{cases}$$

However, in such a case the task becomes a combinatorial one. So, we relax the task and use ξn in place of the indicator function, leading to those equation.

BINUS UNIVERSITY

- Note that optimization is achieved in a trade-off rationale; the user-defined parameter, *C*, controls the influence of each of the two contributions to the minimization task.
- If *C* is large, the resulting margin (the distance between the two hyperplanes defined by $f(x) = \pm 1$) will be small, in order to commit a smaller number of margin errors.
- If *C* is small, the opposite is true. As we will see from the simulation examples, the choice of *C* is very critical.

Case Study

Given data of Singapore Airbnb which can be downloaded in this link

https://www.kaggle.com/jojoker/singapore-airbnb

- Analyze the data and classify the data by applying SVM method.

End of Session 19 & 20

References

- Sergios Theodoridis. (2015). *Machine Learning: a Bayesian and Optimization Perspective*. Jonathan Simpson. ISBN: 978-0-12-801522-3. Chapter 11.
- Aurélien Géron. (2017). 01. *Hands-on Machine Learning with Scikit-Learn and Tensorflow*. O'Reilly Media, Inc..LSI: 978-1-491-96229-9. Chapter 5.
- https://www.kaggle.com/jojoker/singapore-airbnb