ISLAMIC UNIVERSITY OF TECHNOLOGY

Organization of Islamic Cooperation

Board Bazar, Gazipur

Support Vector Machines

CSE 4621

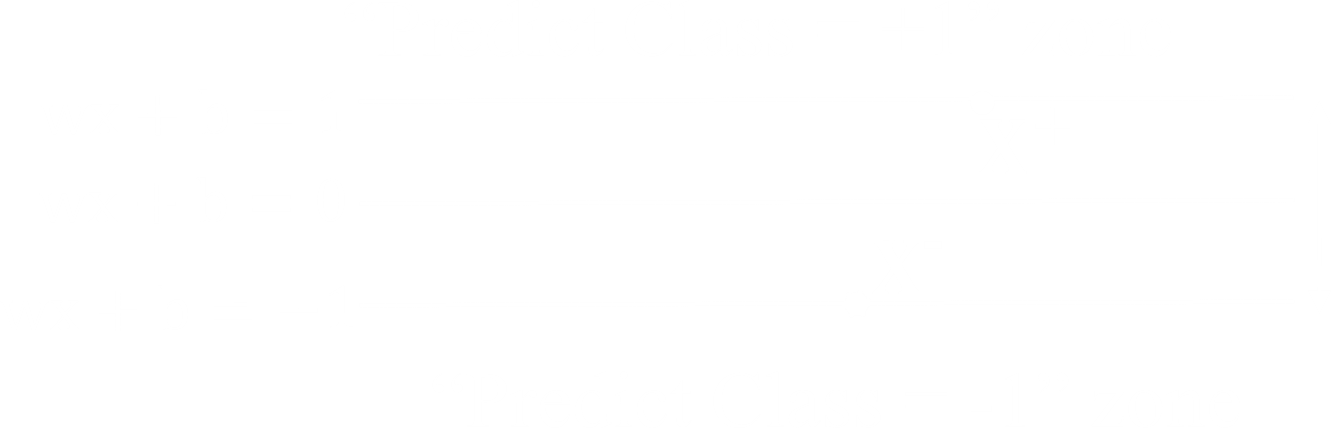
CSE'18

19th April, 2022

**Support Vector Machines** allow us to find a linear decision boundary which minimizes the possibility of misclassifications for future data points. To do this, it uses a **classifier margin**, which is the width by which the decision boundary could be increased before it hits a data point. Without going into the theoretical explanation of this, in short, the larger the classifier margin, the better. SVMs thus attempt to find the decision boundary which has the maximum classifier margin. The data points at which the margin stops are called the **support vectors**.

## Mathematical Interpretation

The good part about linear SVMs is that we only need to worry about the support vectors. Consider that we have two lines parallel to the decision boundary, one going through the support vector on the positive side and the other going through the support vector on the negative side. The distance between these two lines defines the classifier margin.



We have previously seen that the magnitude of the response for a line is given by . Since we want to find the distance between the two lines, which is the **margin**, it is given by:

For a given data point , if the correct label is , we know and if , then . This can be written as . Given this information, our goal is to maximize . can also be written as . For simplicity, we can omit the square root (since the maximum value will still be the maximum value, regardless of the square root), which gives us . Another way of looking at this is that we want to minimize . We now have a **constrained optimization problem** in our hands.

## Lagrange Multipliers

We must find values of and such that

1. is minimized

This is a **constrained optimization problem**, since we are trying to optimize one equation while being constrained by another. In our case, the constrain applies for different samples, so we have constraints. One way of solving such problems is to use the **Lagrange multiplier**.

The equation is said to be in its **primal form**. We convert this to its **dual form**. Since we have constraints, we will also have different Lagrange multipliers (i.e. we are trying to find ). Without going into the details, this can be brought down to:

1. Maximize

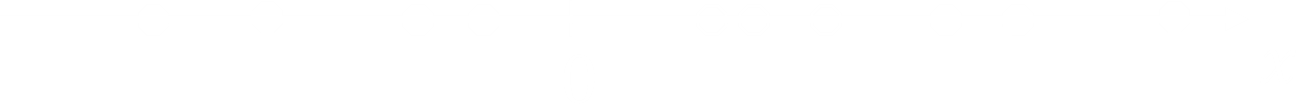
The fun part is that for **non-support vectors**, the values will be , while for support vectors, . From this, we can derive that

for any such that

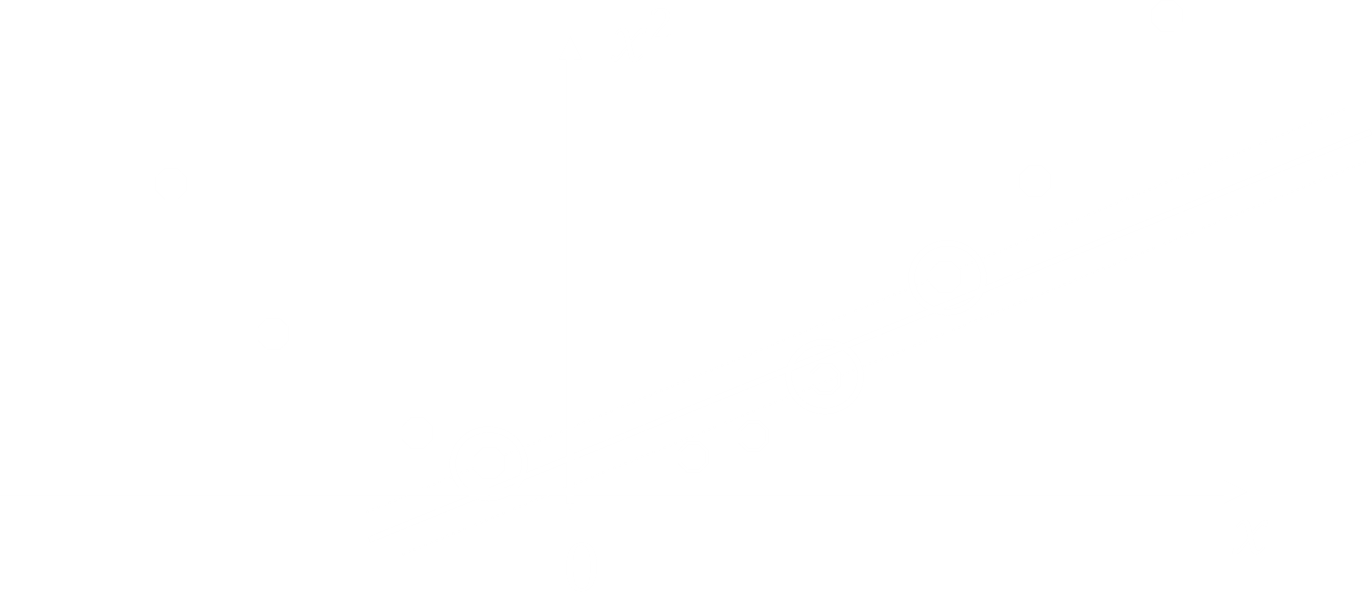
For new data points, , we compute as .

## Non-Linear SVMs

There are some datasets that are impossible separate linearly.



For datasets like this, it might be easier for us to find a linear decision boundary if we map the dataset onto a **higher dimensional plane**.



This mapping has been done by simply using as the -axis. By doing this mapping, we are now able to find a linear decision boundary. The original feature space can always be converted to a higher-dimensional one in which a linear decision boundary can be found. The linear SVM we find is actually non-linear in the original feature space.

### **The Kernel Trick**

For a new data point , we previously saw that we can calculate as . As can be seen, the output depends on the **dot product** of two vectors, . If we map each data point onto a higher-dimension, i.e. , then we get .

It is expensive to map all the datapoints to a higher dimension space, find the SVM and then bring the SVM back to the original space. A **kernel function** gives us the same value in both spaces.

For a 2-dimensional vector , let . We need to show that .

There are several other kernel functions like this.

## Strengths and Weaknesses of SVMs

* Works well when datapoints are clearly separated into classes.
* Efficient in high dimensions, even when the number of dimensions exceeds the number of samples.
* Memory efficient.
* Does not perform well on noisy data sets.
* No probabilistic explanation for classifications.

## Acknowledgements

The contents of this report and the diagrams used are adapted from lecture notes provided by Professor Mingyue Tan from the University of British Columbia.