Support Vector Machines (SVMs)

Bharat Yadav
MA15w1-M

Hochschule Mittweida

03/11/2016

Outline

1. Introduction
2. SVM: Idea
3. Prerequisites
4. SVM: Non-linear case
5. SVM: Polynomial mapping
6. SVM: Kernel trick
7. SVM: Case Study
8. SVM: Overview
9. References

Motivaton of Machine Learning

Suppose we have 50 photographs of elephants and 50 photos of tigers.

vs.

We digitize them into 100 x 100 pixel images, so we have $x \in \mathbb{R}^n$, where n $= 10{,}000$.

Now, given a new (different) photograph we want to answer the question: **is it an elephant or a tiger?** [we assume it is one or the other]

Introduction

What are SVMs?

SVMs are **supervised learning** models with associated learning algorithms that analyze data used for **classification** and **regression analysis**.

Goal of SVM?

To find the **optimal separating hyperplane** which maximizes the margin of the **training data**.

Introduction

Supervised Learning

A machine learning task of inferring a function from labeled training data.

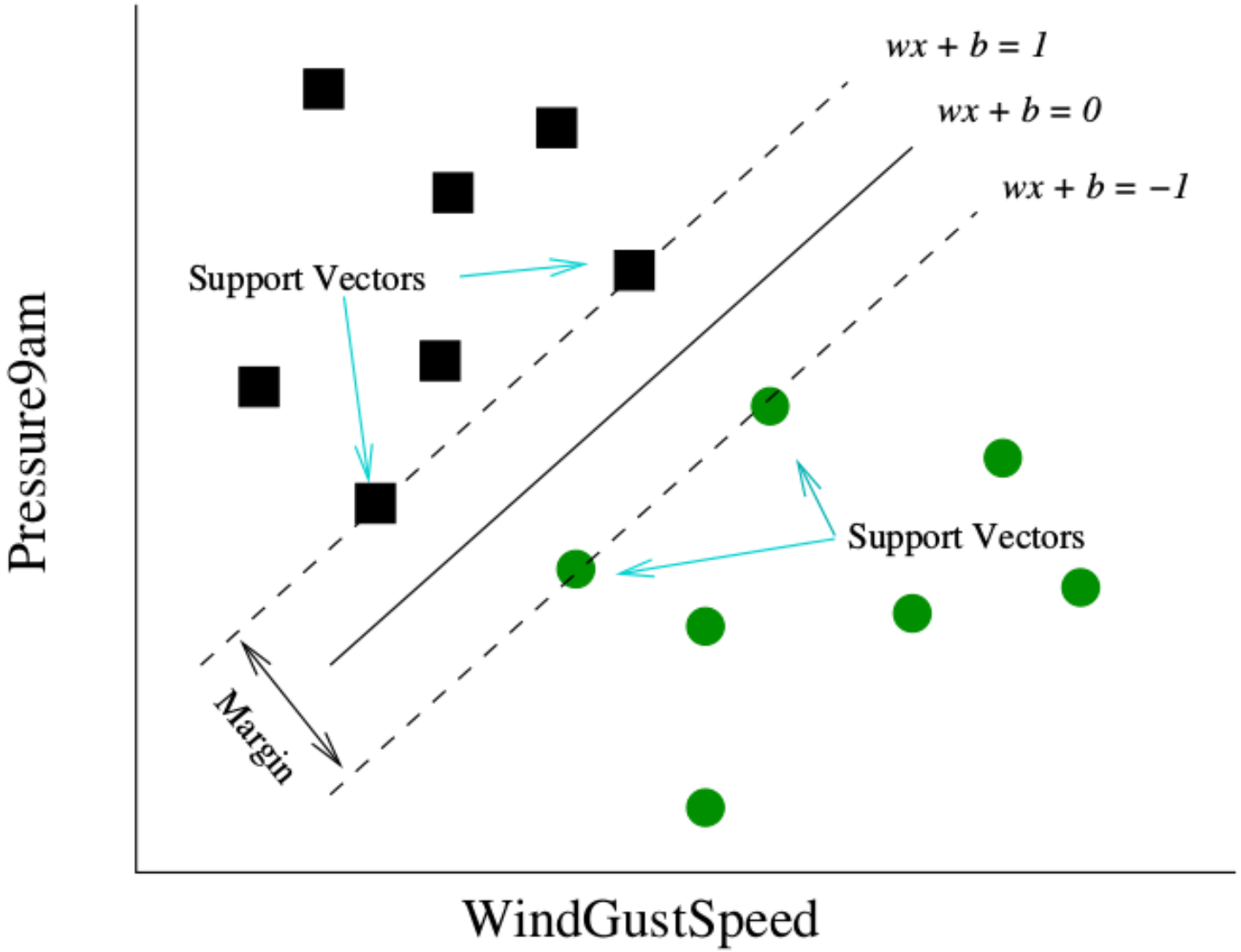
Classification

The problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

Regression Analysis

A statistical process for estimating the relationships among variables.

SVM: Idea


wx + b = 1
wx + b = 0
wx + b = −1
Support Vectors
Support Vectors
Margin
Pressure9am
WindGustSpeed

Prerequisites

Training Data

$$(x_1, y_1), ..., (x_n, y_m) \in \mathcal{X} \times \{\pm 1\}$$

where,
$\mathcal{X}$ is some non-empty set
x_i are inputs/patterns/cases
y_i are outputs/labels/targets
$\{\pm 1\}$ is binary classification/pattern recognition

Prerequisites

Similarity Measure

$$k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$$

$$(x, x') \mapsto k(x, x')$$

where, k is **kernel** (similarity function) and unless stated, is assumed symmetric, i.e.,

$$k(x, x') = k(x', x) \; \forall x, x' \in \mathcal{X}$$

Kernels are used to calculate dot product in **feature space**, $\mathcal{H}$.

Prerequisites

Types of Kernels

- Polynomial kernel

$$k(x, x') = \langle x, x' \rangle^d$$

- Gaussian

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

- Sigmoid

$$k(x, x') = tanh(\kappa \langle x, x' \rangle + \Theta)$$

with suitable choices of $d \in \mathbb{N}$ and $\sigma, \kappa, \Theta \in \mathbb{R}$

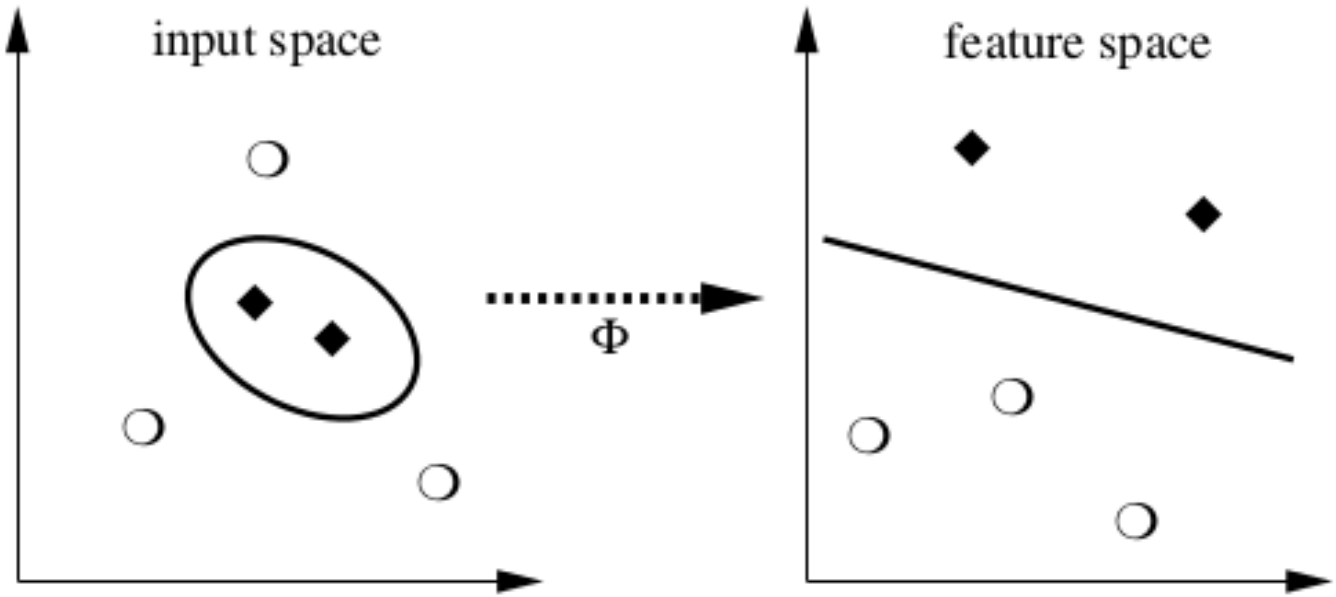
Prerequisites

Feature Space

$$\Phi : \mathcal{X} \mapsto \mathcal{H}$$

$$x \mapsto \mathbf{x} := \Phi(x)$$

where $\mathcal{H}$ is **Feature Space** and $\mathbf{x}$ is vector representation of x in $\mathcal{H}$.



Prerequisites

Advantages of Feature Space

- Easy to define a similarity measure from dot product in $\mathcal{H}$

$$k(x, x') := \langle \mathbf{x}, \mathbf{x'} \rangle = \langle \Phi(x), \Phi(x') \rangle$$

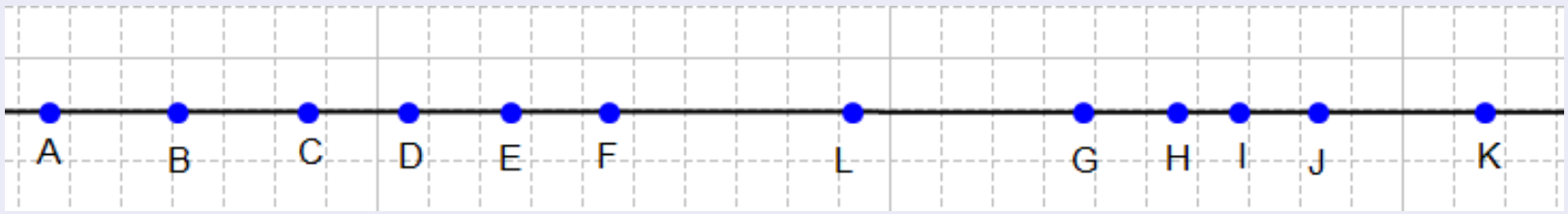
- Easy to deal with the patterns geometrically.
- Freedom to choose Φ allows us to design a large variety of similarity measures and learning algorithms.

Prerequisites

Hyperplane

An hyperplane is a generalization of a plane

- in one dimension, an hyperplane is called a point
- in two dimensions, it is a line
- in three dimensions, it is a plane
- in more dimensions you can call it an hyperplane



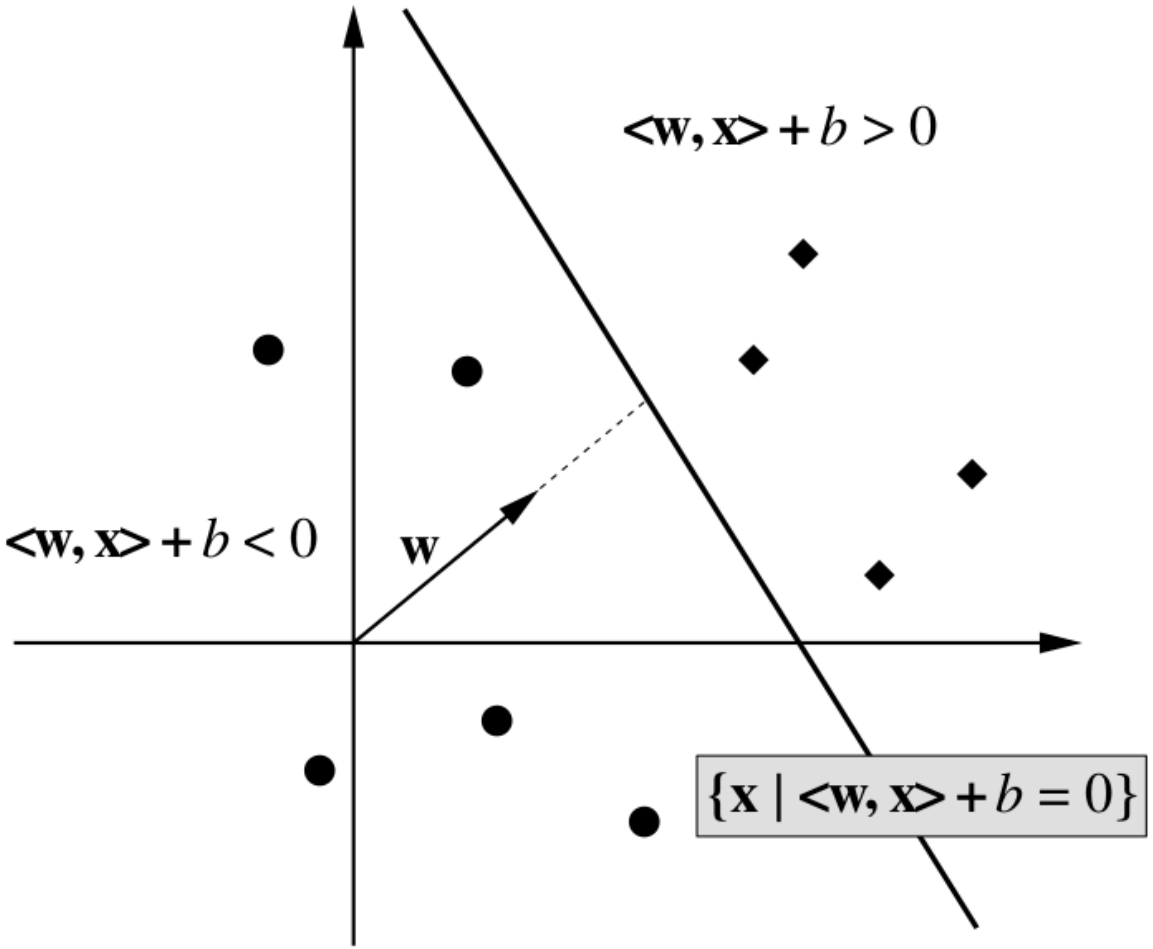
The point L is a separating hyperplane in one dimension

Prerequisites

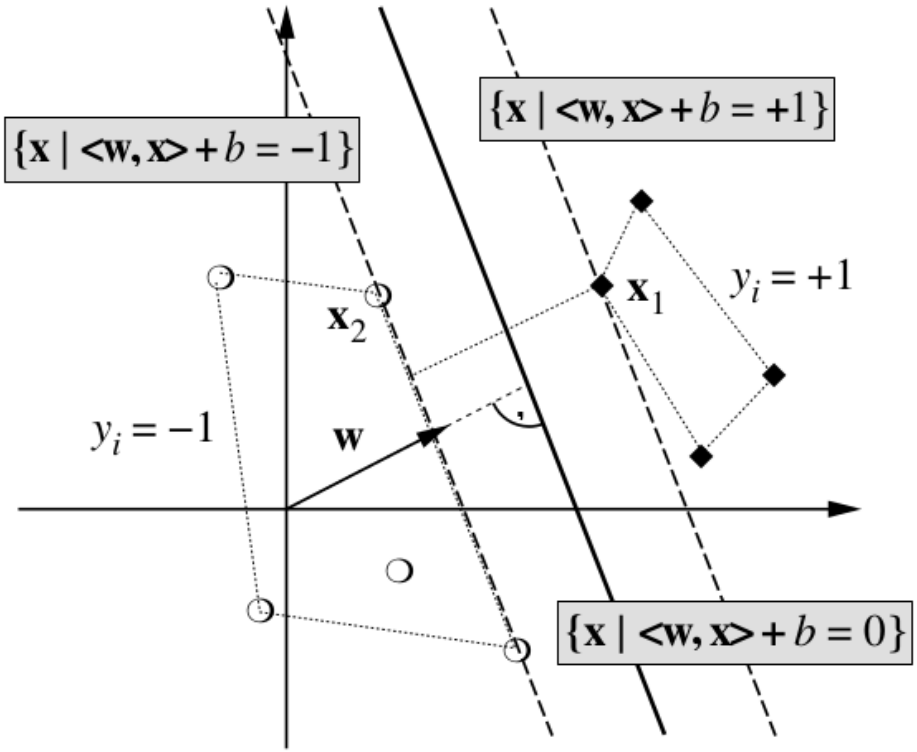
Optimal Separating Hyperplane

- Maximum margin of separation between any training point and hyperplane
- $\max_{\mathbf{w},b} \min \left\{ \|\mathbf{x} - \mathbf{x}_i\| : \mathbf{x} \in \mathcal{H}, \langle \mathbf{w},\mathbf{x} \rangle + b = 0, i = 1, ..., m \right\}$
- where $\langle \mathbf{w},\mathbf{x} \rangle + b = 0$ is the class of hyperplanes in some dot product space $\mathcal{H}$ such that $\mathbf{w} \in \mathcal{H}, b \in \mathbb{R}$.

Prerequisites



Prerequisites



Note:

$$\langle \mathbf{w}, \mathbf{x}_1 \rangle + b = +1$$
$$\langle \mathbf{w}, \mathbf{x}_2 \rangle + b = -1$$

$$\Rightarrow \quad \langle \mathbf{w}, (\mathbf{x}_1 - \mathbf{x}_2) \rangle = 2$$

$$\Rightarrow \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, (\mathbf{x}_1 - \mathbf{x}_2) \right\rangle = \frac{2}{\|\mathbf{w}\|}$$

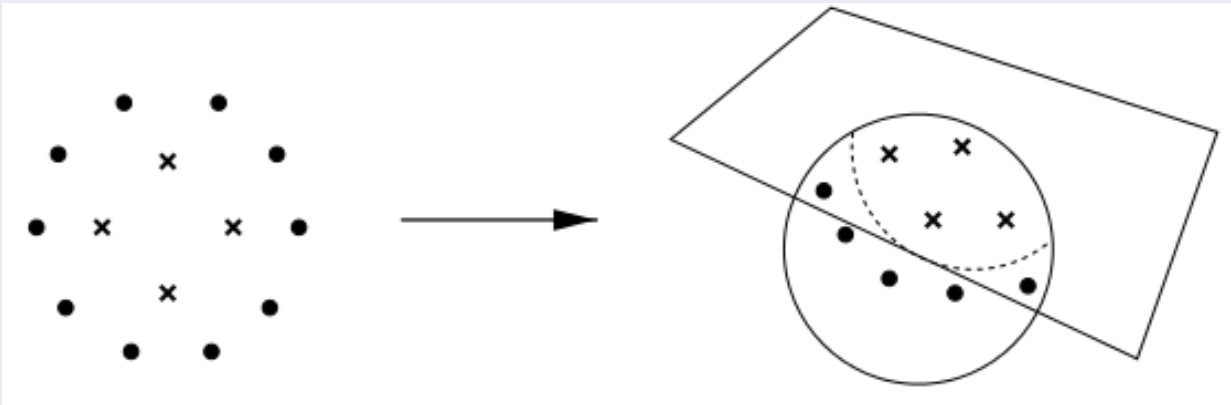
SVM: Non-linear case

- Map data into a richer feature space including nonlinear features
- construct a hyperplane in that space so all other equations are the same
- Formally, pre-process the data with:

$$x \mapsto \Phi(x)$$

- then learn the map from $\Phi(x)$ to y:

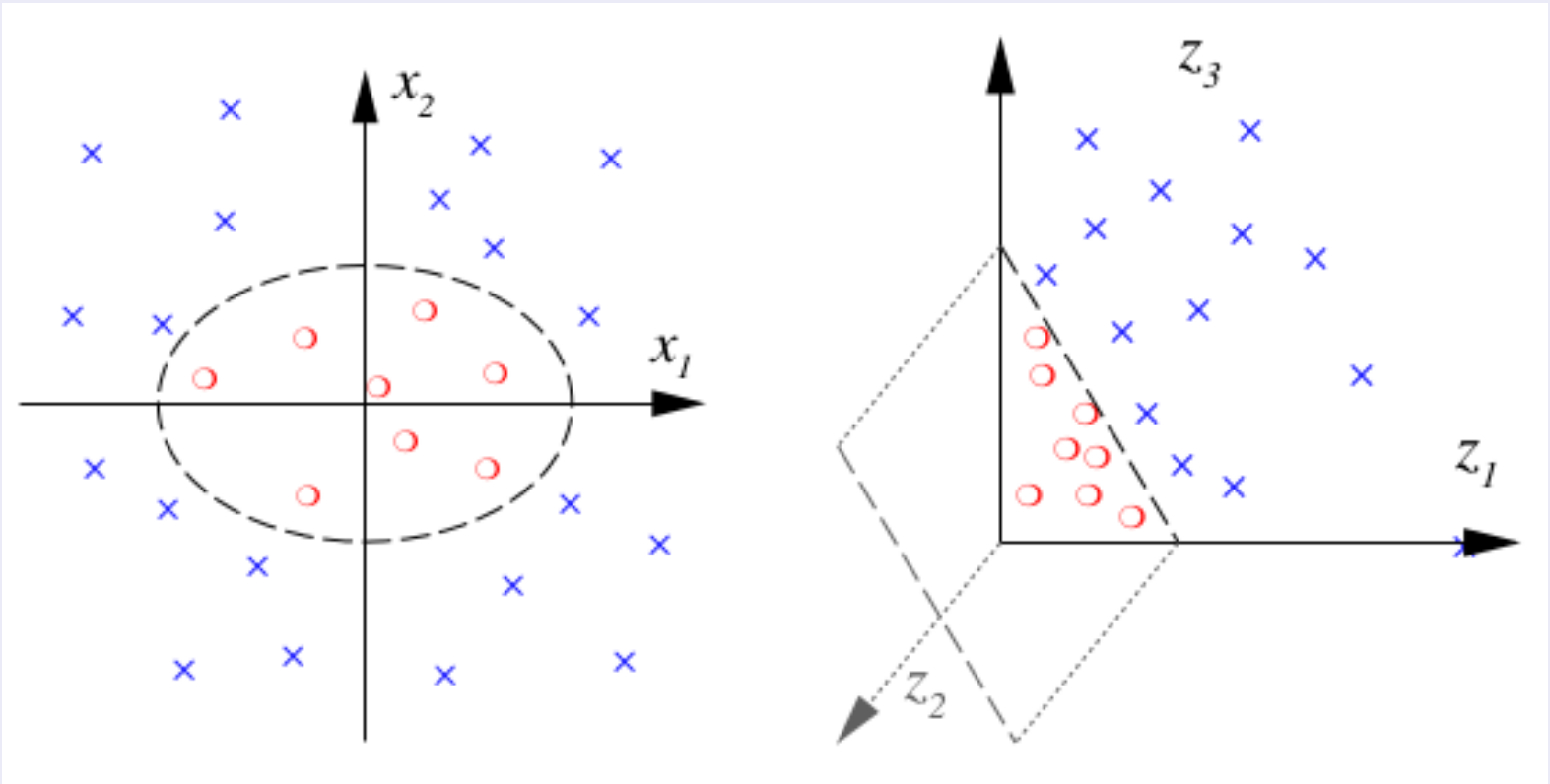
$$f(x) = w.\Phi(x) + b$$



SVM: Polynomial mapping

$$\Phi : \mathbb{R}^2 \mapsto \mathbb{R}^3$$

$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$



SVM: Kernel trick

- Problem: dimensionality of $\Phi(x)$ can be very large, making w hard to represent explicitly in memory.
- The Representer theorem (Kimeldorf and Wahba, 1971) shows that (for SVMs as a special case):

$$w = \sum_{i=1}^{m} \alpha_i \Phi(x_i)$$

- for some α, instead of optimizing w directly we can optimize α.

$$f(x) = \sum_{i=1}^{m} \alpha_i \Phi(x_i).\Phi(x) + b$$

- Where $K(x_i, x) = \Phi(x_i).\Phi(x)$ is the kernel function.

SVM: Case Study

Selected Observations

Two input variables (i.e.,two-dimensional space):Pressure3pm and Sunshine

```
> library(rattle)
> obs <- with(weather, Pressure3pm+Sunshine > 1032 |
              (Pressure3pm+Sunshine < 1020 &
               RainTomorrow == "Yes"))
> ds <- weather[obs,]
> with(ds, plot(Pressure3pm, Sunshine,
                pch=as.integer(RainTomorrow),
                col=as.integer(RainTomorrow)+1))
> lines(c(1016.2, 1019.6), c(0, 12.7))
> lines(c(1032.8, 1001.5), c(0, 12.7))
> legend("topleft", c("Yes", "No"), pch=2:1, col=3:2)
```

SVM: Case Study

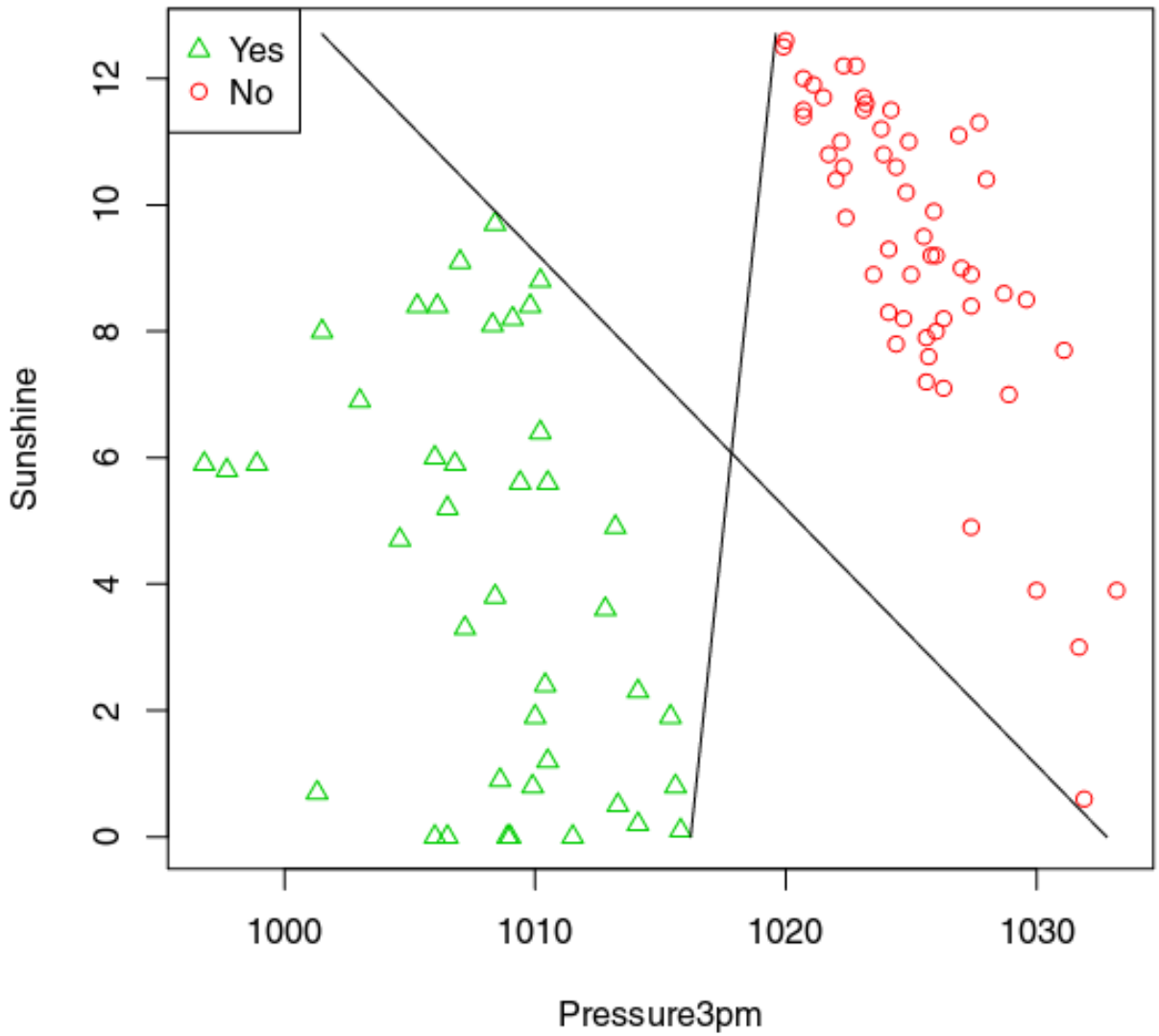


Figure: A simple and easily linearly separable collection of observations

SVM: Case Study

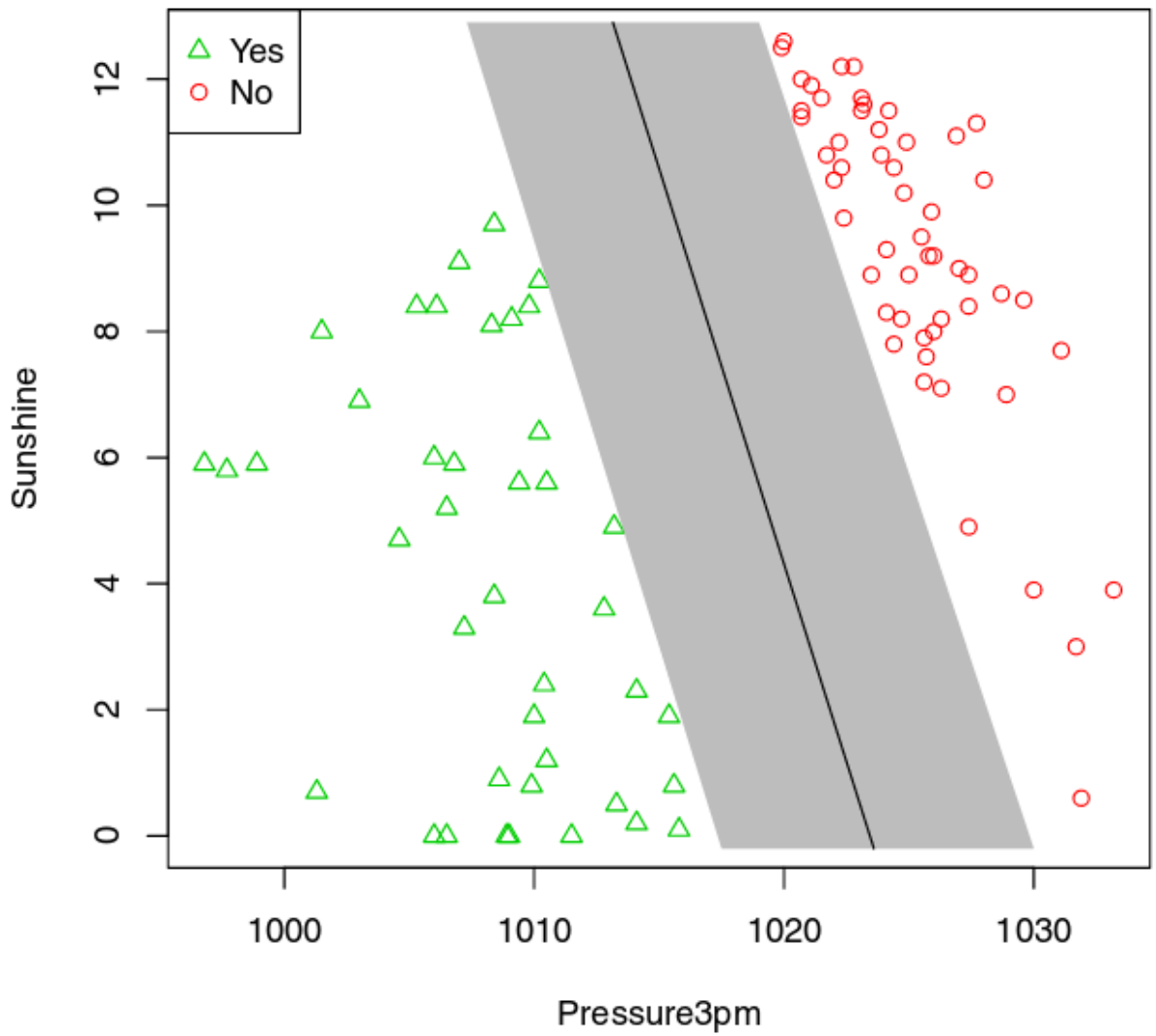


Figure: Maximal region or margin between the two classes of observations

SVM: Case Study

Algorithm(Original Observations)

```
> ds <- weather
> with(ds, plot(Pressure3pm, Sunshine,
                pch=as.integer(RainTomorrow),
                col=as.integer(RainTomorrow)+1))
> legend("topleft", c("Yes", "No"), pch=2:1, col=3:2)
```

SVM: Case Study

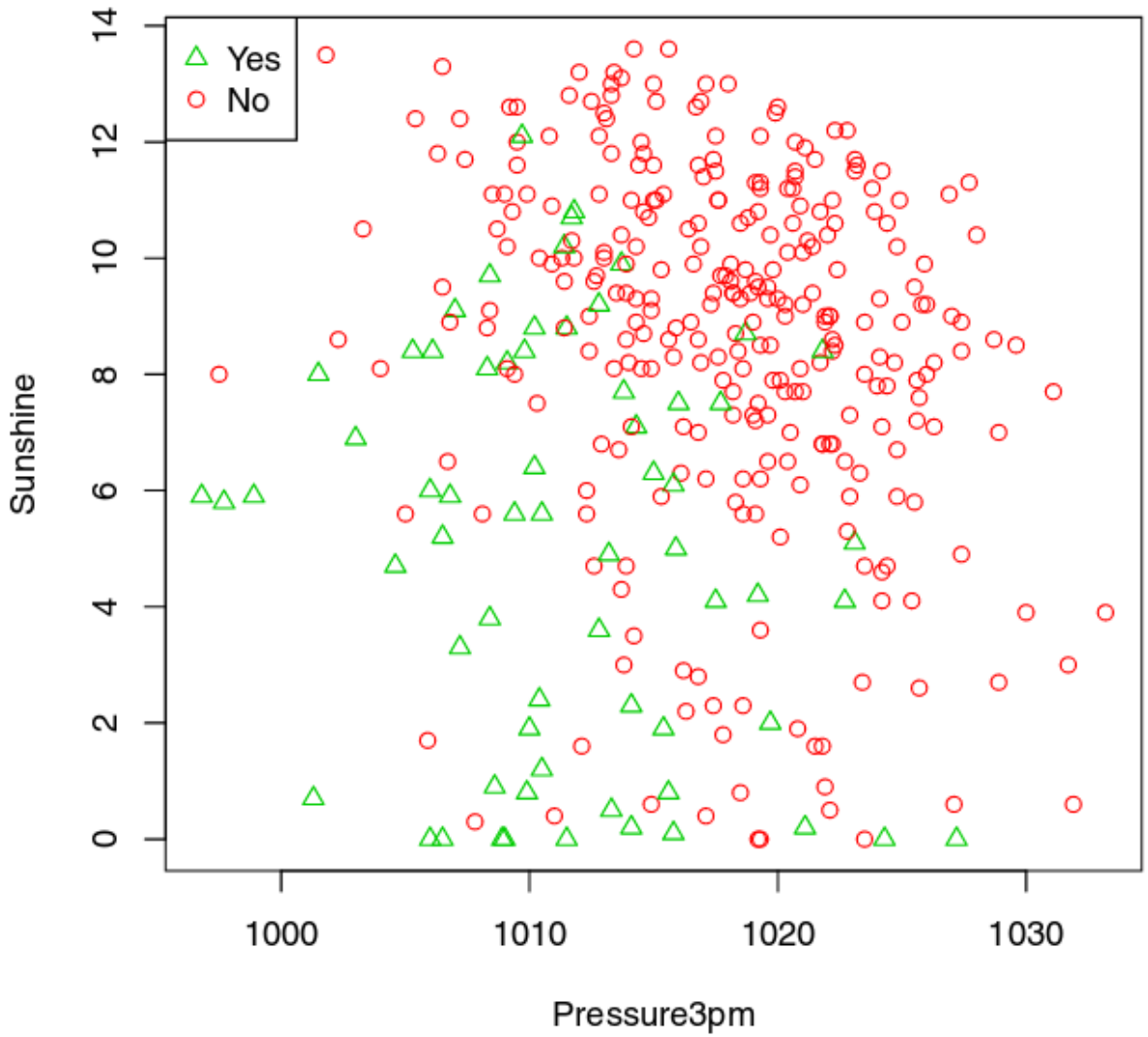


Figure: A nonlinearly separable collection of observations

SVM: Case Study

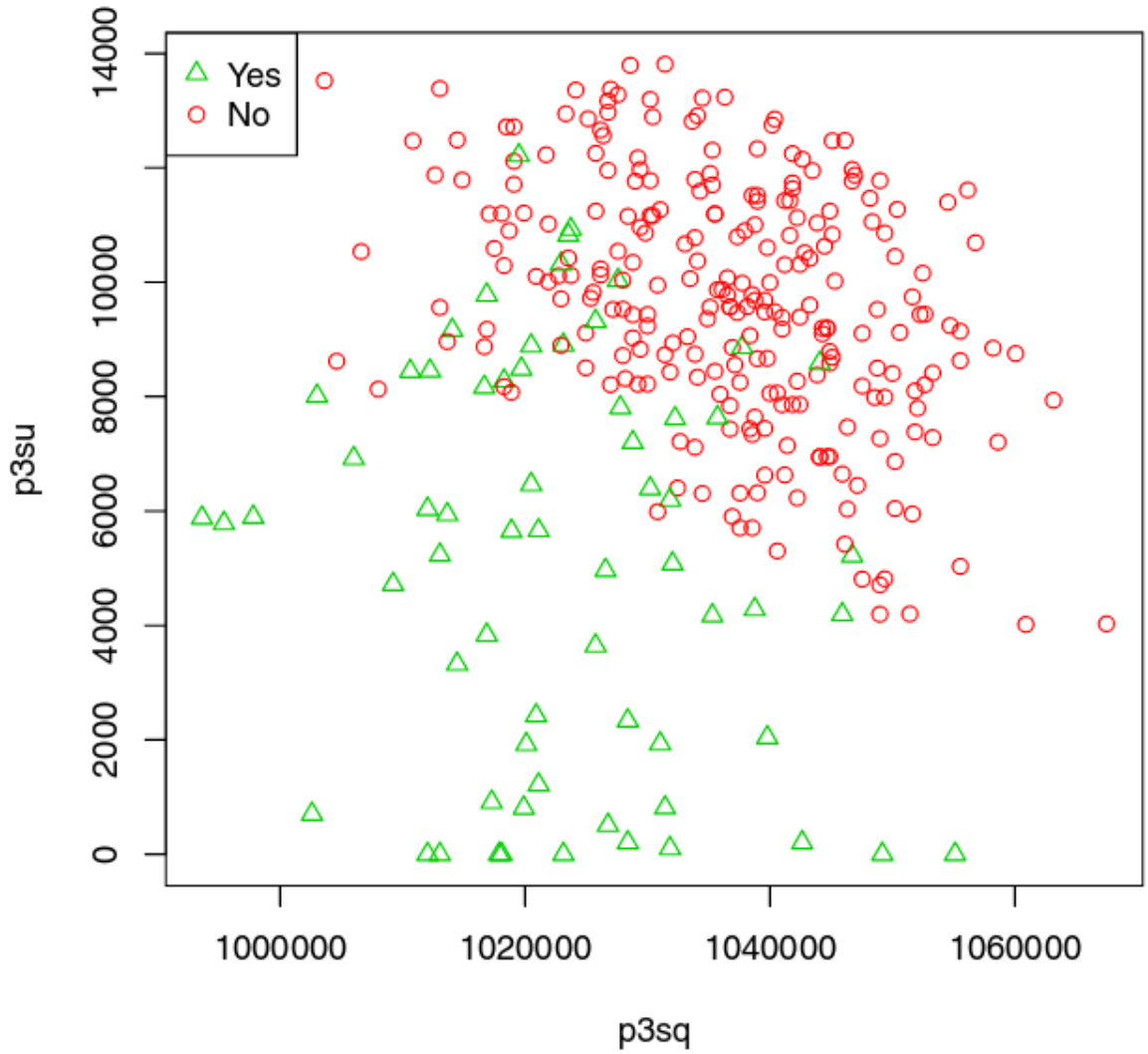


Figure: Nonlinearly transformed observations showing Pressure3pm squared (x-axis) against Pressure3pm multiplied by Sunshine, artificially enhanced

SVM: Overview

SVM: Advantages

- Performs well on problems that are nonlinear, sparse, and high-dimensional.
- Modeling only deals with support vectors instead of whole training data set(training-set size independent).
- The model is less affected by outliers.

SVM: Overview

SVM: Disadvantages

- Sensitive to the choice of tuning option (e.g., the type of transformations to perform), making it harder to use and time-consuming to identify the best model.
- Transformations performed can be computationally expensive and are performed both whilst building the model and when scoring new data.

References

1. Data Mining with Rattle and R: Graham Williams.
2. Tutorial learning-with-kernels: Schölkopf.
3. Support Vector Machine Tutorial: Jason Weston: NEC Labs America,4 Independence Way, Princeton, USA.
4. http://www.svm-tutorial.com

Thank you for your attention.