Support Vector Machines

Lecture 10

Termeh Shafie

1

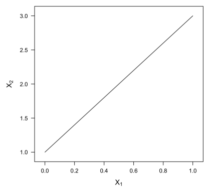
Support Vector Machine (SVM) is a supervised learning algorithm used to learn a **hyperplane** that can solve the **binary classification problem**

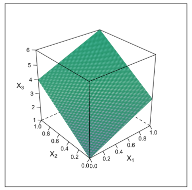
2

Hyperplanes

`“a flat affine subspace”` 😐

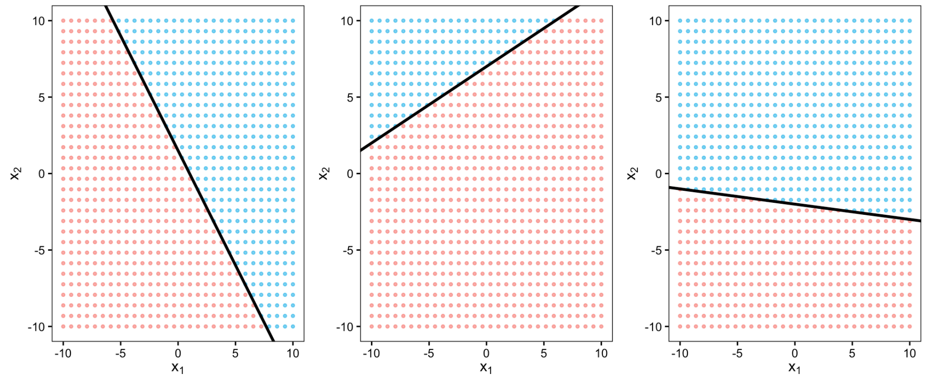
- **Flat:**
  - hyperplane is not curved, it increases/decreases constantly in each direction
- **Affine:**
  - the hyperplane doesn’t need to pass through the origin
  - it can have an “offset” or be shifted (may have intercept)
- **Subspace:**
  - a subset of vectors in a larger vector space
  - in a $d-$dimensional space, a hyperplane has dimension $d-1$
  - in 3D it is a **plane**
  - in 2D it is a **line**
  - in 1D it is a **point**





3

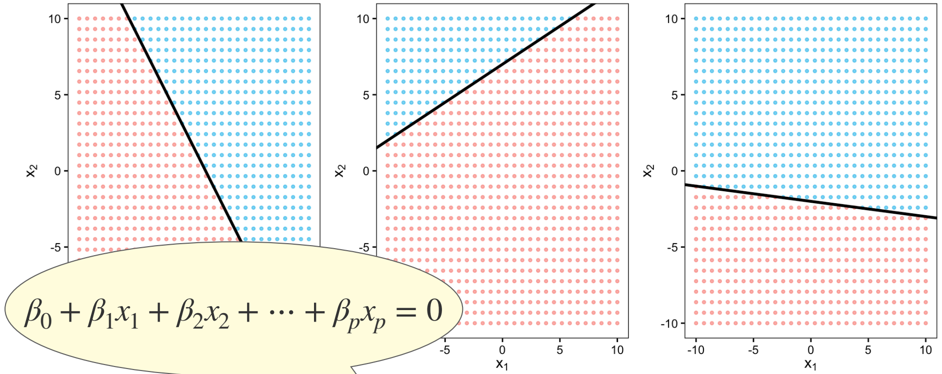
Hyperplanes Divide the Space in Half



$$\beta_0 + \beta^T x = 0$$

4

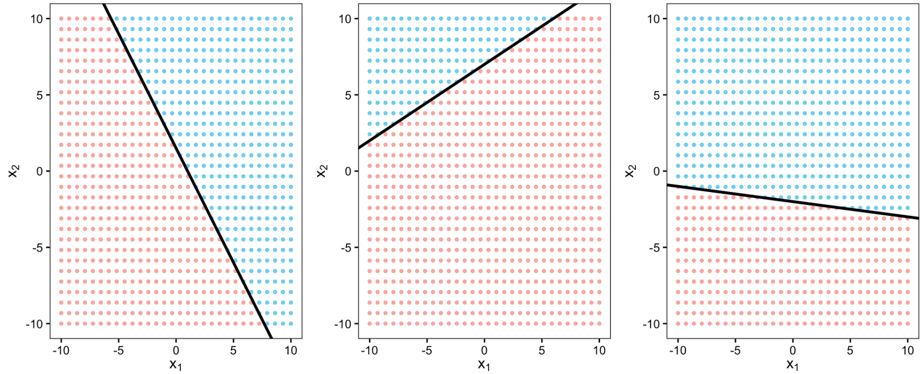
Hyperplanes Divide the Space in Half



$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = 0$

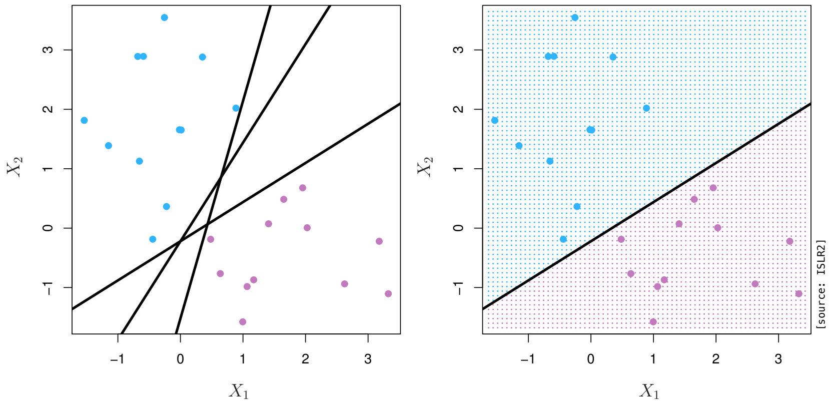
$\beta_0 + \beta^T x = 0$

Hyperplanes Divide Spaces in Half



$$\begin{aligned} \beta_0 + \beta^T x > 0 \text{ if } y_i = 1 \\ \beta_0 + \beta^T x < 0 \text{ if } y_i = -1 \end{aligned} \implies y_i(\beta_0 + \beta^T x) > 0$$

Hyperplanes Divide Spaces in Half

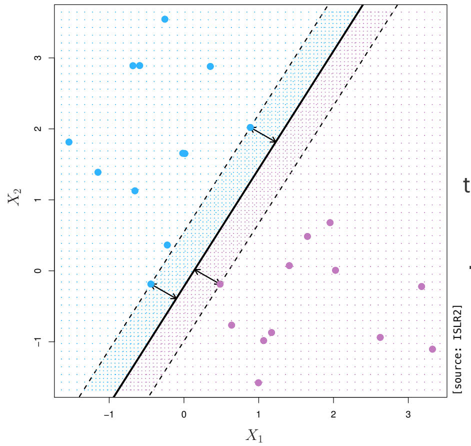


if a separating hyperplane exists, there will be ∞ many of them

how do we choose just one?

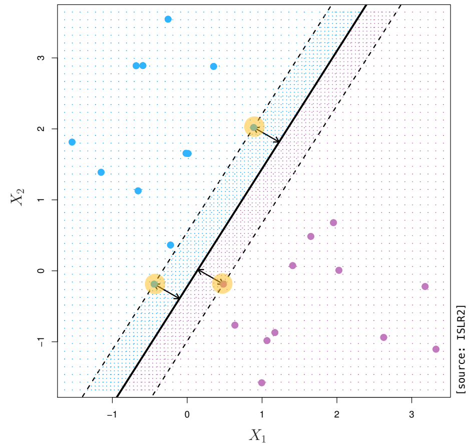
Classification

Maximal Margin Classifier



choose the separating hyperplane that's furthest from the training examples

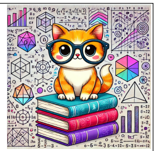
Maximal Margin Classifier



support vectors
the closest points from both classes

the margin
the distance from hyperplane to support vectors

Maximal Margin Classifier: The Math



The maximal margin classifier solves a constrained optimization problem:

$$\max_{\beta_0, \beta_1, \ldots, \beta_p} \quad M$$

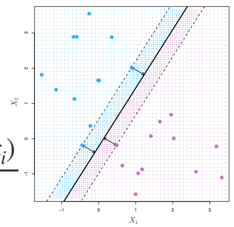
subject to:

$$\|\beta\| = 1$$

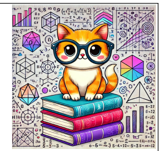
$$y_i(\beta_0 + \beta^T x_i) \geq M, \quad \forall i = 1, \ldots, n$$

ensured each observation is on the correct side of the hyperplane and at least a distance M from the hyperplane, i.e., M is the margin of the hyperplane

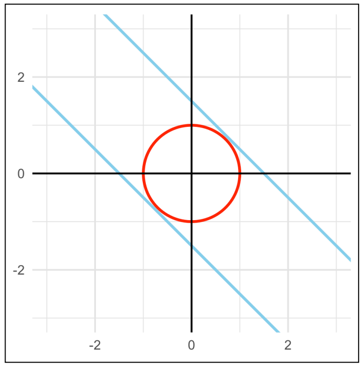
distance between x_i and line where $\|\beta\| = \sqrt{\sum_{j=1}^{p} \beta_j^2}$ is the Euclidean norm of β $\left\{ \frac{y_i(\beta_0 + \beta^T x_i)}{\|\beta\|} \right.$



What is a Constrained Optimization Problem?



Optimize $f(x, y)$ subject to $g(x, y) = k$



$$f(x, y) = 2x + y$$

$$g(x, y) = x^2 + y^2 = 1$$

$$\max_{\beta_0, \beta_1, \ldots, \beta_p} \quad M$$

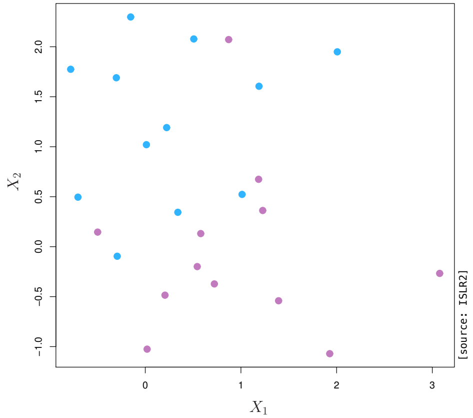
subject to:

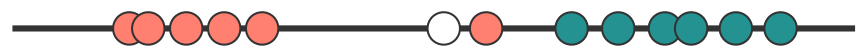
$$\|\beta\| = 1$$

$$y_i(\beta_0 + \beta^T x_i) \geq M$$

The Non-Separable Case

the optimization problem for the maximal margin classifier often has no solution with $M > 0$



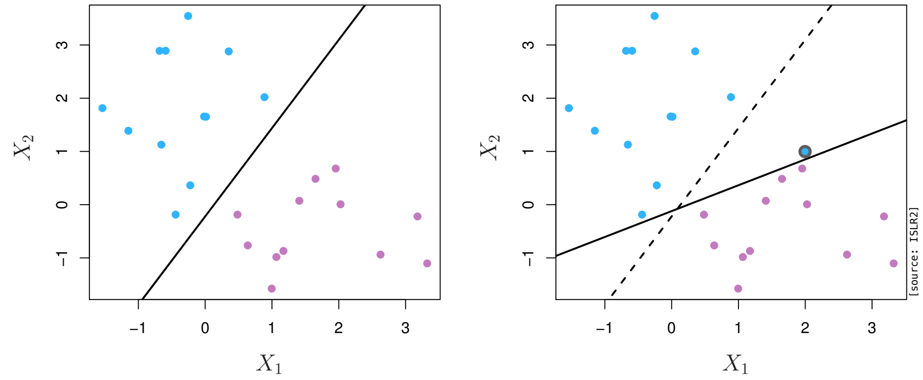


and it's very sensitive to outliers!

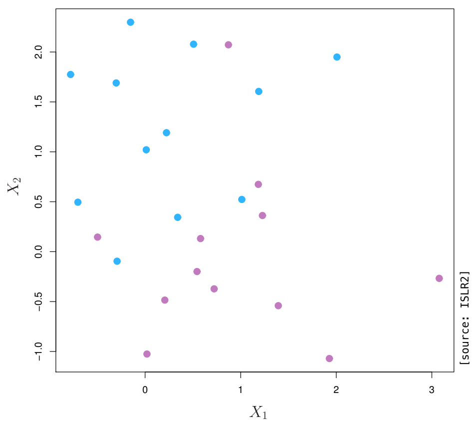
The Non-Separable Case

even if the data are separable, they are sometimes noisy

$\Longrightarrow$ poor solution for the maximum margin classifier



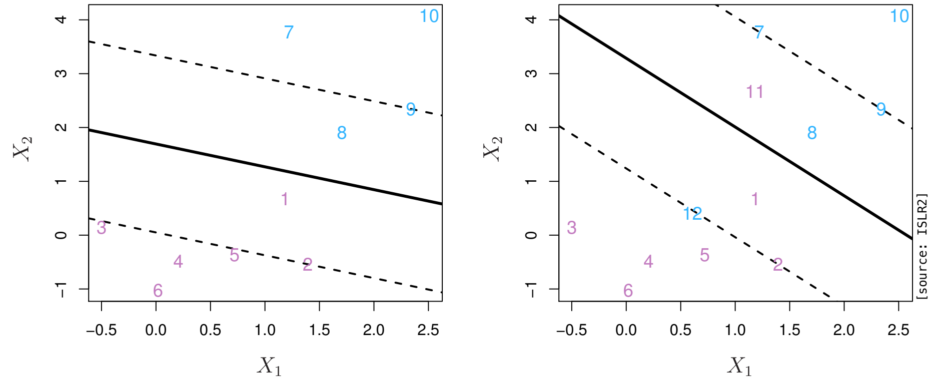
The Non-Separable Case



support vector classifier:
using a so-called **soft margin**
to **almost** separate the classes

Support Vector Classifier

allows us to classify data that is **not linearly separable**



Bias Variance Trade Off

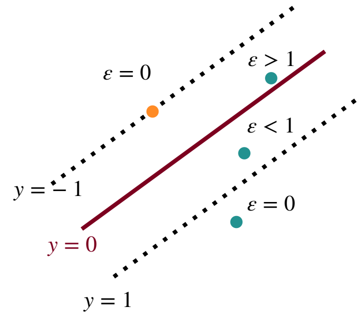
support vector classifier:
how do we choose the soft margin?
→ ***cross validation!***

Support Vector Classifier

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\epsilon_2,\ldots,\epsilon_n} M$$

subject to:

$$\|\beta\| = 1$$
$$y_i(\beta_0 + \beta^T x_i) \geq M(1 - \epsilon_i)$$
$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$



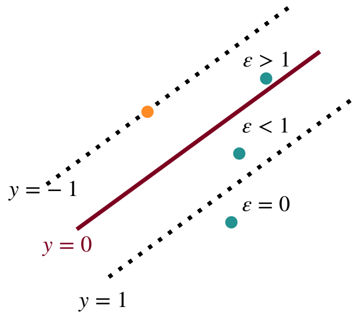
$\epsilon_1, \ldots, \epsilon_n$ are **slack variables** where $\epsilon_i = 0$ means i^{th} observation is on correct side of margin
< 1 means i^{th} observation is on wrong side of margin
> 1 means i^{th} observation is on wrong side of hyperplane

Support Vector Classifier

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\epsilon_2,\ldots,\epsilon_n} M$$

subject to:

$$\|\beta\| = 1$$
$$y_i(\beta_0 + \beta^T x_i) \geq M(1 - \epsilon_i)$$
$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

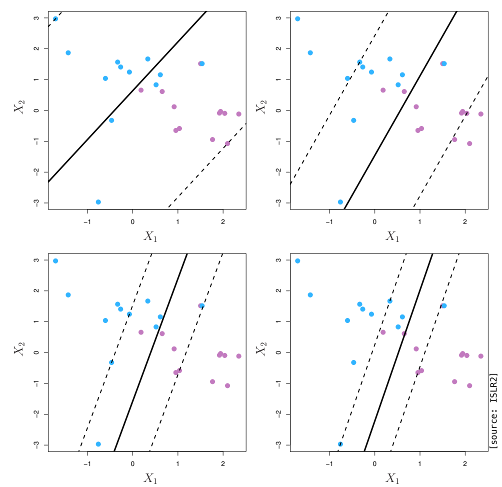


C is the tuning parameter/penalty on error:
$C = 0$ implies maximal margin hyperplane (superposed it exists)
$C > 0$ is the total violations to the margin that we can tolerate
$\implies$ max C observations can be on the wrong side of hyperplane

Support Vector Classifier

C: penalty on error

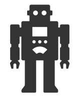
	Regularization	Margins	Bias/Variance
Small C	more	wider	prone to underfitting
Large C	less	narrower	prone to overfitting



But what if data looks like this?

21

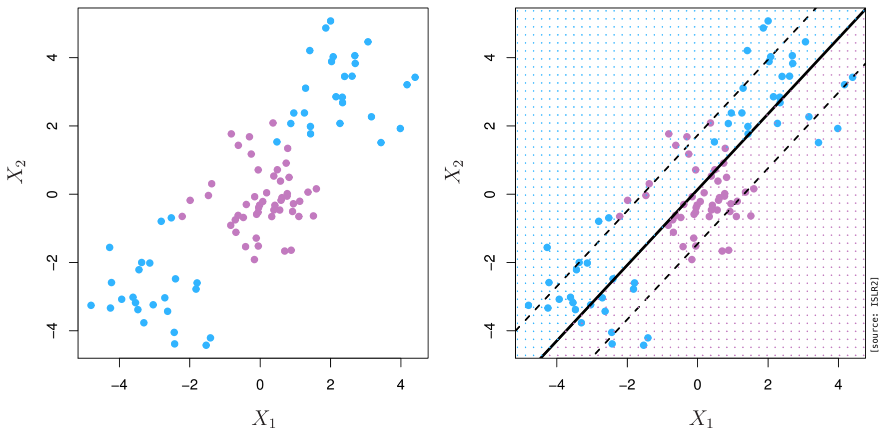
Support Vector Machines



Support Vector Machine (SVM) use **Kernel Functions** to systematically find **Support Vector Classifiers** in higher dimensions

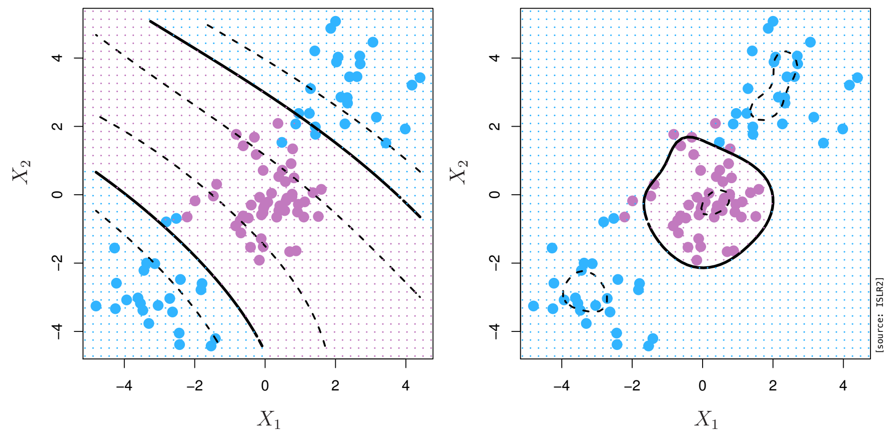
22

Not Linearly Separable Even With Error



23

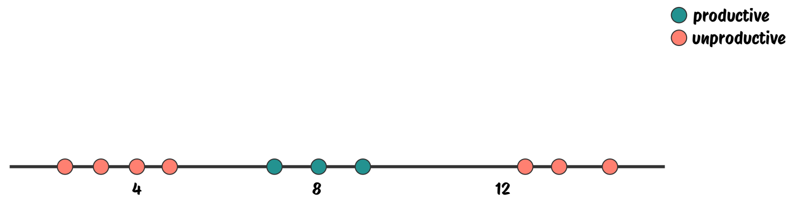
The Kernel Trick



24

The Kernel Trick

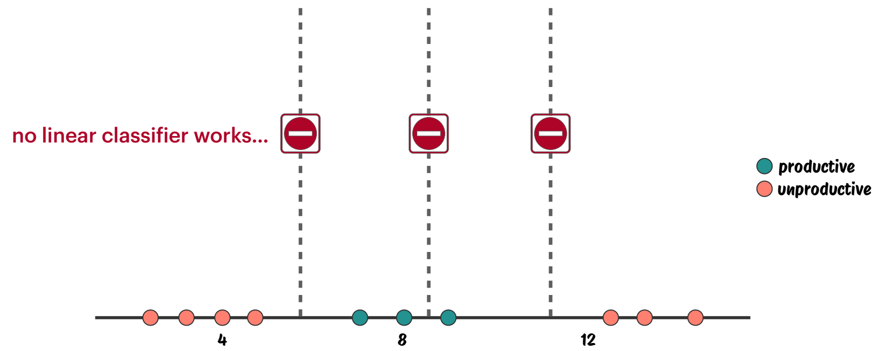
my productivity based on hours of sleep



25

The Kernel Trick

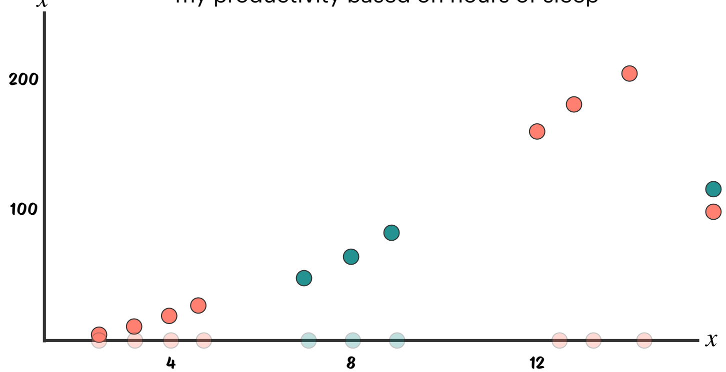
my productivity based on hours of sleep



26

The Kernel Trick

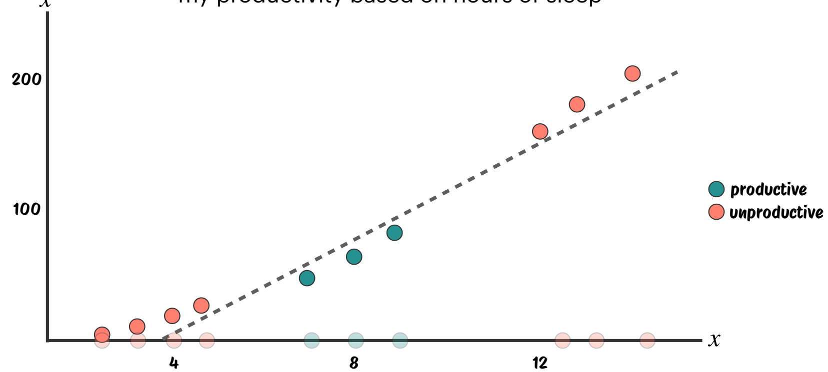
my productivity based on hours of sleep



27

The Kernel Trick

my productivity based on hours of sleep

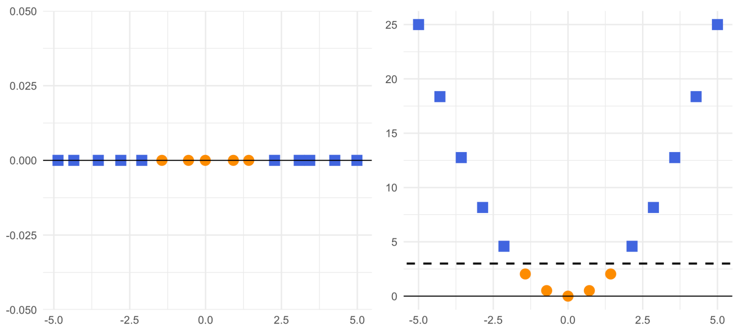


28

The Kernel Trick

what is an SVM Kernel?

A function that computes the relationship between vectors in multiple dimensions
(without actually having to calculate the coordinates for those dimensions)



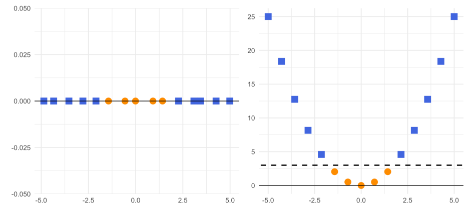
The Polynomial Kernel

The **Polynomial Kernel** in the previous sleep vs. productivity example

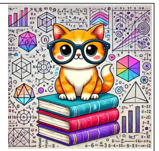
$$\underbrace{K(a,b)}_{\text{Kernel}} = (a \cdot b + r)^d$$

different observations

where r is the coefficient and d the degree of polynomial (determined by **cross validation**)



The Polynomial Kernel



The **Polynomial Kernel** in the previous sleep vs. productivity example

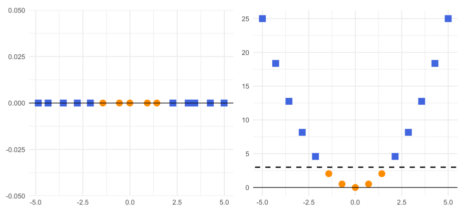
$K(a,b) = (a \cdot b + r)^d$ where r is the coefficients and d the degree

we set $r = \frac{1}{2}$ and $d = 2$:

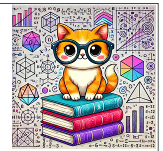
$$\left(a \cdot b + \frac{1}{2}\right)^2 = \left(a \cdot b + \frac{1}{2}\right)\left(a \cdot b + \frac{1}{2}\right)$$

$$+a^2b^2 + \frac{1}{2}ab + \frac{1}{2}ab + \frac{1}{4}$$

$$= \boxed{ab + a^2b^2 + \frac{1}{4}}$$



The Polynomial Kernel



The **Polynomial Kernel** in the previous sleep vs. productivity example

$K(a,b) = (a \cdot b + r)^d$ where r is the coefficients and d the degree

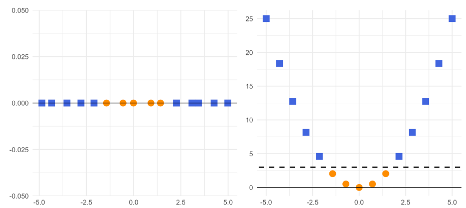
we set $r = \frac{1}{2}$ and $d = 2$:

$$\left(a \cdot b + \frac{1}{2}\right)^2 = \left(a \cdot b + \frac{1}{2}\right)\left(a \cdot b + \frac{1}{2}\right)$$

$$+a^2b^2 + \frac{1}{2}ab + \frac{1}{2}ab + \frac{1}{4}$$

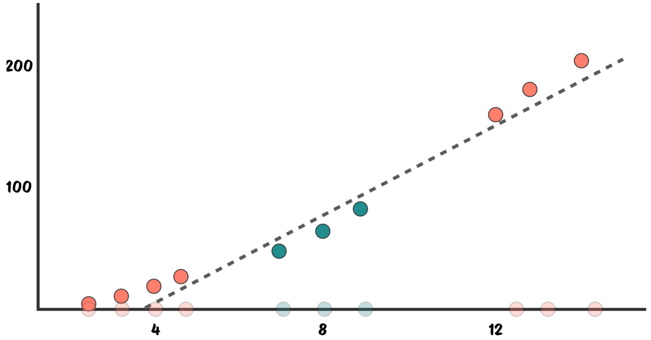
$$= \boxed{ab + a^2b^2 + \frac{1}{4}} = \underbrace{\left(a, a^2, \frac{1}{2}\right) \cdot \left(b, b^2, \frac{1}{2}\right)}_{\text{dot product}}$$

gives us the high dimensional coordinates for the data



The Polynomial Kernel

$$(a, a^2, \frac{1}{2}) \cdot (b, b^2, \frac{1}{2})$$



33

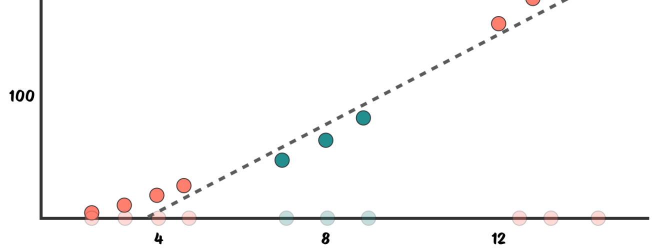
The Polynomial Kernel

A function that computes the relationship between vectors in multiple dimensions
(without actually having to calculate the coordinates for those dimensions)

example: $a = 4,\ b = 8$

plug values into the Kernel to get
the high-dimensional relationship

$$(a \cdot b + \frac{1}{2})^2 = (16 \cdot 64 \cdot \frac{1}{2}) = 512$$



34

The Polynomial Kernel

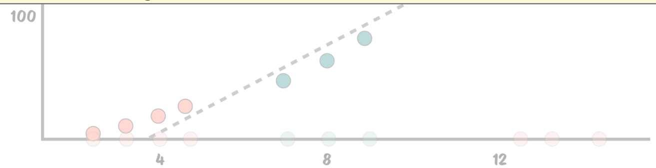
A function that computes the relationship between vectors in multiple dimensions
(without actually having to calculate the coordinates for those dimensions)

example: $a = 4,\ b = 8$

plug values into the Kernel to get
the high-dimensional relationship

$$(a \cdot b + \frac{1}{2})^2 = (16 \cdot 64 \cdot \frac{1}{2}) = 512$$

```
one of the 2-dimensional relationships we need to solve for the SV classifier
(even though we did not transform the data into 2 dimensions)
```



35

The Radial Kernel (RBF)

The **Radial Kernel**

$$K(a, b) = e^{-\gamma(a-b)^2}$$

projects to **infinite dimensional** space
works similar to nearest neighbors classifier

36

The Radial Kernel (RBF)

The **Radial Kernel**

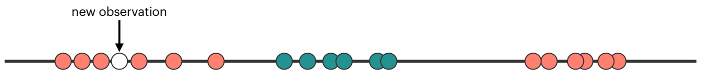
$K(a, b) = e^{-\gamma(a-b)^2}$ projects to **infinite dimensional** space

works similar to nearest neighbors classifier

the amount of influence one observation has on another is a function of the squared distance

$K(a, b) = e^{-\gamma(a-b)^2}$

γ scales the squared distance to determine the strength of influence (determined by **cross validation)**



$K(a, b) = e^{-\gamma(a-b)^2} =$ high dimensional relationship

The Radial Kernel (RBF)

The **Radial Kernel**

$K(a, b) = e^{-\gamma(a-b)^2}$ projects to **infinite dimensional** space

works similar to nearest neighbors classifier

we can use the Polynomial Kernel to get the intuition behind how Radial Kernel works in infinite dimensions

$K(a, b) = (a \cdot b + r)^d$

set $r = 0 \implies (a \cdot b)^d = a^d \cdot b^d$

set $d = 1 \implies ab = (a) \cdot (b)$

The Radial Kernel (RBF)

The **Radial Kernel**

$K(a, b) = e^{-\gamma(a-b)^2}$ projects to **infinite dimensional** space

works similar to nearest neighbors classifier

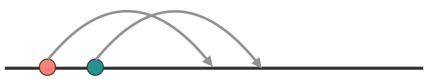
we can use the Polynomial Kernel to get the intuition behind how Radial Kernel works in infinite dimensions

$K(a, b) = (a \cdot b + r)^d$

set $r = 0 \implies (a \cdot b)^d = a^d \cdot b^d$

set $d = 1 \implies ab = (a) \cdot (b)$

set $d = 2 \implies a^2b^2 = (a^2) \cdot (b^2)$



The Radial Kernel (RBF)

The **Radial Kernel**

$K(a, b) = e^{-\gamma(a-b)^2}$ projects to **infinite dimensional** space

works similar to nearest neighbors classifier

we can use the Polynomial Kernel to get the intuition behind how Radial Kernel works in infinite dimensions

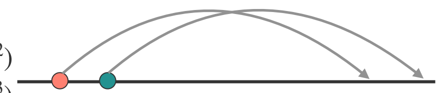
$K(a, b) = (a \cdot b + r)^d$

set $r = 0 \implies (a \cdot b)^d = a^d \cdot b^d$

set $d = 1 \implies ab = (a) \cdot (b)$

set $d = 2 \implies a^2b^2 = (a^2) \cdot (b^2)$

set $d = 3 \implies a^3b^3 = (a^3) \cdot (b^3)$



The Radial Kernel (RBF)

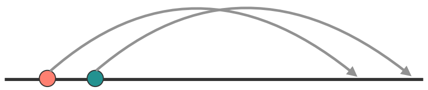
The **Radial Kernel**

$K(a,b) = e^{-\gamma(a-b)^2}$

projects to **infinite dimensional** space
works similar to nearest neighbors classifier

we can use the Polynomial Kernel to get the intuition
behind how Radial Kernel works in infinite dimensions

$K(a,b) = (a \cdot b + r)^d$
set $r = 0 \implies (a \cdot b)^d = a^d \cdot b^d$
set $d = 1 \implies (a) \cdot (b)$
set $d = 2 \implies (a^2) \cdot (b^2)$
set $d = 3 \implies (a^3) \cdot (b^3)$



we stay in same dimension when $r = 0$,
but what if we took these polynomials as a sum?

The Radial Kernel (RBF)

The **Radial Kernel**

$K(a,b) = e^{-\gamma(a-b)^2}$

projects to **infinite dimensional** space
works similar to nearest neighbors classifier

we can use the Polynomial Kernel to get the intuition
behind how Radial Kernel works in infinite dimensions

$K(a,b) = (a \cdot b)^d$
$ab + a^2b^2 = (a, a^2)(b, b^2)$

The Radial Kernel (RBF)

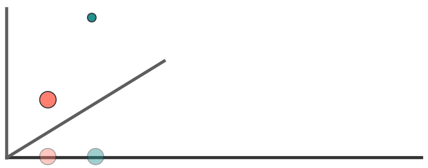
The **Radial Kernel**

$K(a,b) = e^{-\gamma(a-b)^2}$

projects to **infinite dimensional** space
works similar to nearest neighbors classifier

we can use the Polynomial Kernel to get the intuition
behind how Radial Kernel works in infinite dimensions

$K(a,b) = (a \cdot b)^d$
$ab + a^2b^2 + a^3b^3 = (a, a^2, a^3)(b, b^2, b^3)$



The Radial Kernel (RBF)

The **Radial Kernel**

$K(a,b) = e^{-\gamma(a-b)^2}$

projects to **infinite dimensional** space
works similar to nearest neighbors classifier

we can use the Polynomial Kernel to get the intuition
behind how Radial Kernel works in infinite dimensions

$K(a,b) = (a \cdot b)^d$

$$ab + a^2b^2 + a^3b^3 + \cdots + a^\infty b^\infty = (a, a^2, a^3, \ldots, a^\infty)(b, b^2, b^3, \ldots, b^\infty)$$

take sum for infinite terms gives dot product with infinite dimensions!

The Radial Kernel: Taylor Series Expansion

$K(a,b) = e^{-\gamma(a-b)^2} = e^{-\gamma(a^2+b^2-2ab)} = e^{-\gamma(a^2+b^2)}e^{\gamma 2ab}$

set $\gamma = \frac{1}{2} \implies e^{-\frac{1}{2}\gamma(a^2+b^2)}\boxed{e^{ab}}$ Taylor expansion of this term

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(\infty)}(a)}{\infty!}(x-a)^\infty$$

$$e^x = e^a + \frac{e^a}{1!}(x-a) + \frac{e^a}{2!}(x-a)^2 + \frac{e^a}{3!}(x-a)^3 + \cdots + \frac{e^a}{\infty!}(x-a)^\infty$$, around $a = 0$ we get

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{\infty!}x^\infty,$$ so for the point ab we get

$$e^{ab} = 1 + (ab) + \frac{(ab)^2}{2!} + \frac{(ab)^3}{3!} + \cdots + \frac{(ab)^\infty}{\infty!}$$

each term contains Polynomial Kernel with $r = 0$ and d from 0 to $d = \infty$

45

The Radial Kernel: Taylor Series Expansion

$K(a,b) = e^{-\gamma(a-b)^2} = e^{-\gamma(a^2+b^2-2ab)} = e^{-\gamma(a^2+b^2)}e^{\gamma 2ab}$

set $\gamma = \frac{1}{2} \implies e^{-\frac{1}{2}\gamma(a^2+b^2)}\boxed{e^{ab}}$ Taylor expansion of this term

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(\infty)}(a)}{\infty!}(x-a)^\infty$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^\infty}{\infty!}$$

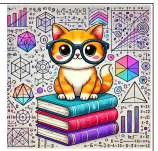
Radial Kernels have coordinates for infinite dimensions!

$$e^{ab} = \boxed{1} + \boxed{(ab)} + \frac{1}{2!}\boxed{(ab)^2} + \frac{1}{3!}(ab)^3 + \cdots + \frac{1}{\infty!}\boxed{(ab)^\infty}$$

$$\boxed{a^0b^0} + \boxed{a^1b^1} + \boxed{a^2b^2} + a^3b^3 + \cdots + \boxed{a^\infty b^\infty} = (1, a, a^2, a^3, \ldots, a^\infty)(1, b, b^2, b^3, \ldots, b^\infty)$$

46

The Radial Kernel: Taylor Series Expansion



$K(a,b) = e^{-\gamma(a-b)^2} = e^{-\gamma(a^2+b^2-2ab)} = e^{-\gamma(a^2+b^2)}e^{\gamma 2ab}$

set $\gamma = \frac{1}{2} \implies e^{-\frac{1}{2}\gamma(a^2+b^2)}e^{ab}$

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(\infty)}(a)}{\infty!}(x-a)^\infty$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^\infty}{\infty!}$$

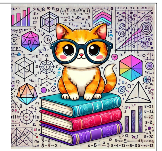
Radial Kernels have coordinates for infinite dimensions!

$$e^{ab} = 1 + (ab) + \frac{1}{2!}(ab)^2 + \frac{1}{3!}(ab)^3 + \cdots + \frac{1}{\infty!}(ab)^\infty$$

$$e^{ab} = \left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^3, \ldots, \sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^3, \ldots, \sqrt{\frac{1}{\infty!}}b^\infty\right)$$

47

The Radial Kernel: Taylor Series Expansion



$K(a,b) = e^{-\gamma(a-b)^2} = e^{-\gamma(a^2+b^2-2ab)} = e^{-\gamma(a^2+b^2)}e^{\gamma 2ab}$

set $\gamma = \frac{1}{2} \implies e^{-\frac{1}{2}\gamma(a^2+b^2)}e^{ab}$

$$e^{-\frac{1}{2}(a-b)^2} = e^{-\frac{1}{2}(a^2+b^2)}\left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \ldots\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \ldots\right)$$

let $s = \sqrt{e^{-\frac{1}{2}(a^2+b^2)}}$

$$e^{-\frac{1}{2}(a-b)^2} = \left(s, s\sqrt{\frac{1}{1!}}a, s\sqrt{\frac{1}{2!}}a^2, \ldots, s\sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(s, s\sqrt{\frac{1}{1!}}b, s\sqrt{\frac{1}{2!}}b^2, \ldots, s\sqrt{\frac{1}{\infty!}}b^\infty\right)$$

$$e^{ab} = \left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^3, \ldots, \sqrt{\frac{1}{\infty!}}a^\infty\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^3, \ldots, \sqrt{\frac{1}{\infty!}}b^\infty\right)$$

48

The Radial Kernel

$$K(a,b) = e^{-\gamma(a-b)^2} = e^{-\gamma(a^2+b^2-2ab)} = e^{-\gamma(a^2+b^2)}e^{\gamma 2ab}$$

$$\text{set } \gamma = \frac{1}{2} \implies e^{-\frac{1}{2}\gamma(a^2+b^2)}e^{ab}$$

$$e^{-\frac{1}{2}(a-b)^2} = e^{-\frac{1}{2}(a^2+b^2)}\left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \ldots\right)\cdot\left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \ldots\right)$$

$$\text{let } s = \sqrt{e^{-\frac{1}{2}(a^2+b^2)}}$$

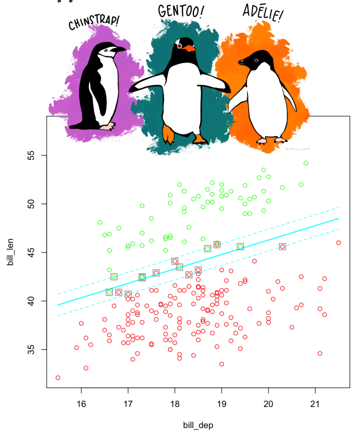
$$e^{-\frac{1}{2}(a-b)^2} = \left(s,\ s\sqrt{\frac{1}{1!}}\,a,\ s\sqrt{\frac{1}{2!}}\,a^2,\ \ldots,\ s\sqrt{\frac{1}{\infty!}}\,a^{\infty}\right)\cdot\left(s,\ s\sqrt{\frac{1}{1!}}\,b,\ s\sqrt{\frac{1}{2!}}\,b^2,\ \ldots,\ s\sqrt{\frac{1}{\infty!}}\,b^{\infty}\right)$$

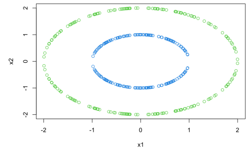
```
the Radial Kernel is equal to a Dot Product that has
coordinates for an infinite number of dimensions.
```

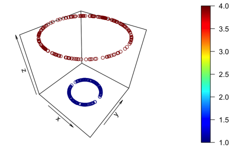
49

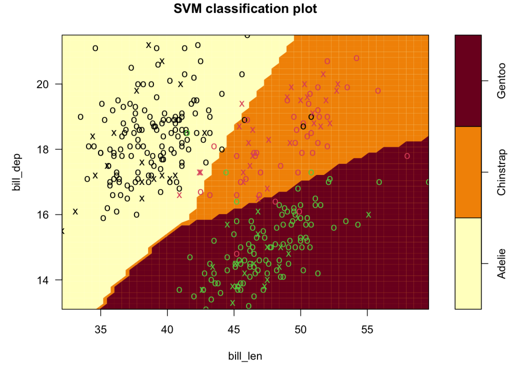
This Week's Practical

Support Vector Machines









50