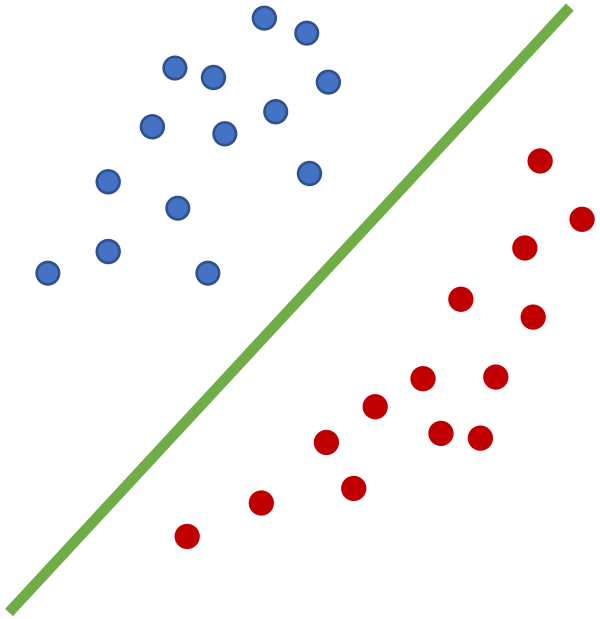
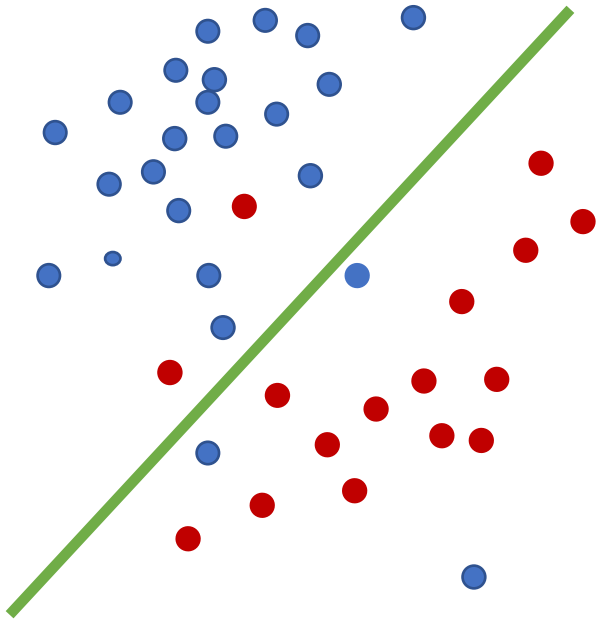
Nonlinear SVMs: kernels

Gal Chechik, Uri Shaham

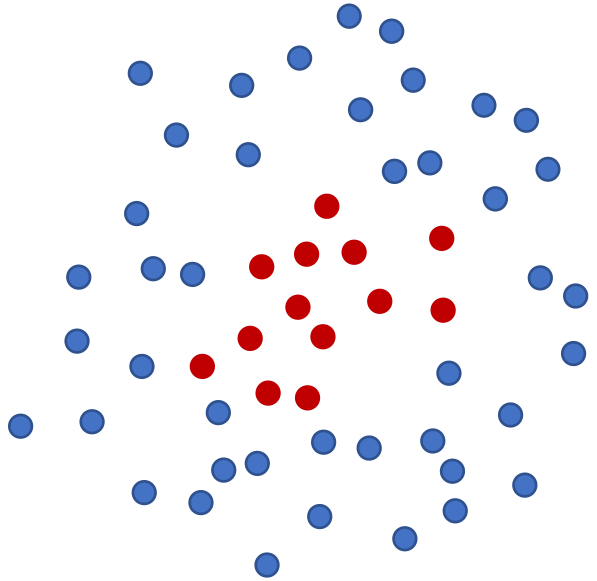
Linear separability



Linearly separable data



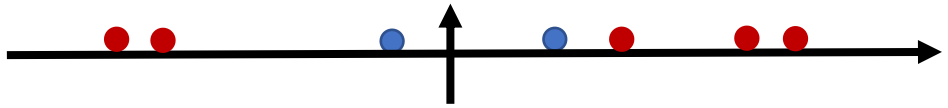
Measurement & label noise

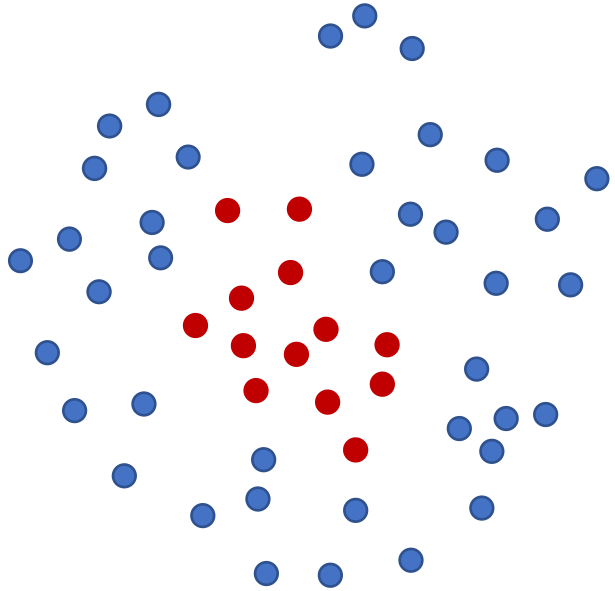


Non-linear

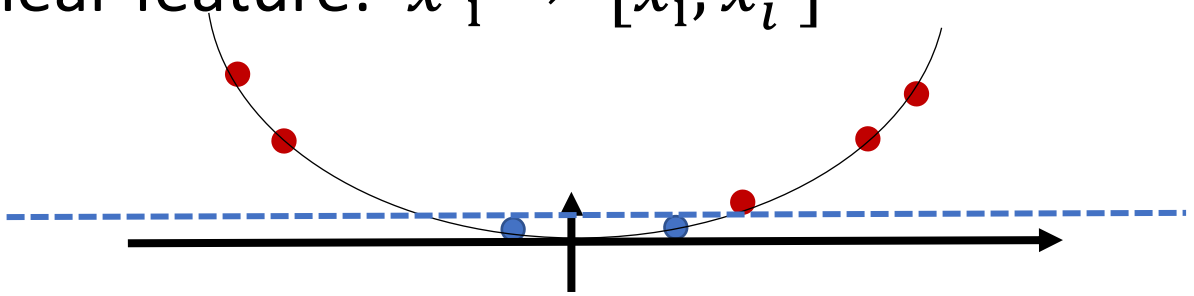
Non-linear separation and kernels

Consider $x \in R$





Add a non-linear feature: $x'_{\mathrm{i}} \rightarrow [x_{\mathrm{i}}, x_i^2]$



It is linearly separable now in R^2!

Non-linear SVM

The idea: Map x into a new feature space $x \rightarrow \phi(x)$. $x \in \mathbb{R}^d$,

$\phi(x) \in \mathbb{R}^D$, and $D \gg d$ (we usually want many possible features).
Then find a linear classifier in feature space.

Large D has two problems: **Computation**, **Generalization**
We take care of generalization through regularization.
What about efficient computation?

We now show that for some feature spaces, computation complexity of learning does not depend on D! (only on d). D can even be infinite!

Linear classifiers over non-linear feature space

A perceptron-style algorithm with $x \rightarrow \phi(x)$

Training:

Initialize w

Repeat until convergence

For $i = 1 \ldots n$

if $y_i w^{\mathrm{T}} \phi(\mathbf{x}_i) < 0$

$w^{\mathrm{t+1}} = w^{\mathrm{t}} + y_i \phi(x_i)$

Inference:

$$\hat{y}_i = Sign(\mathbf{w}^{\mathrm{T}} \phi(x_i))$$

Recall: in SVM $\mathbf{w} = \sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j$.

Training:

Initialize $\alpha_1, \ldots, \alpha_n$

Repeat until convergence

For $i = 1 \ldots n$

if $\sum_{j=1}^{n} \left(\alpha_j y_j \phi^T(x_j)\right) \phi(x_i) y_i < 0$

$\alpha_i = \alpha_i + 1$

Inference:

$$\hat{y}_i = Sign\left(\sum_{j=1}^{n} \alpha_j y_j \phi^T(x_j)\, \phi(x_i)\right)$$

Because we add to the term $\alpha_i y_i \phi(\mathbf{x}_i)$ in the sum,

$\alpha_i y_i \phi(\mathbf{x}_i)$

$= \alpha_i y_i \phi(\mathbf{x}_i) + y_i \phi(\mathbf{x}_i)$

Note that $\phi(x_i)$ appears **only** within inner products!

The SVM Dual Problem

A similar case holds also in the solution of the SVM problem:

It can be shown (e.g., Cortes and Vapnik, 1995) that the Lagrangian dual of the SVM problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2$$

$$\text{s.t. } y_i w^T x_i \geq 1 \; \forall i$$

is

$$\mathcal{L}(x, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_i x_i^T x_j$$

Note that here as well, the x's appear only within inner products. So replacing x with $\phi(x)$, we can apply nonlinearity to the features if we have a way to compute inner products in some feature space $\phi(\cdot)$, without actually transforming the data to that space. **This is what kernels are for.**

The kernel trick: $k(i,j) = \phi^T(x_i)\phi(x_j)$

We never need to compute any individual $\phi(x_i)$, only $\phi^T(x_j)\phi(x_i)$.

For some types of $\phi^T(x)$, that can be done efficiently. For example:

$$\phi(x) = \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ x_1 \\ x_2 \\ 1 \end{bmatrix}$$

$$\phi(x)\cdot\phi(z) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ x_1 \\ x_2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2}z_1z_2 \\ z_1 \\ z_2 \\ 1 \end{bmatrix} = x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 + x_1z_1 + x_2z_2 + 1 = (x_1z_1 + x_2z_2 + 1)^2 =$$

$$= (x\cdot z + 1)^2$$

$$= k(x,z)$$

Computing $k(i,j)$ can be done in the original input space, not feature space!

It can be shown (Moore- Aronszajn theorem) that every PSD function defines a valid kernel.

For example, the RBF kernel: $K(x_i, x_j) = e^{-\|x_i - x_j\|^2/\sigma^2}$.

This simple fact allows us to compute inner products in feature spaces (even infinite ones) without ever transforming the data to these spaces!

More on the mathematics of kernels in “Mathematical methods in Data Science”, 895222-01.

Example of a decision boundary of kernel SVM

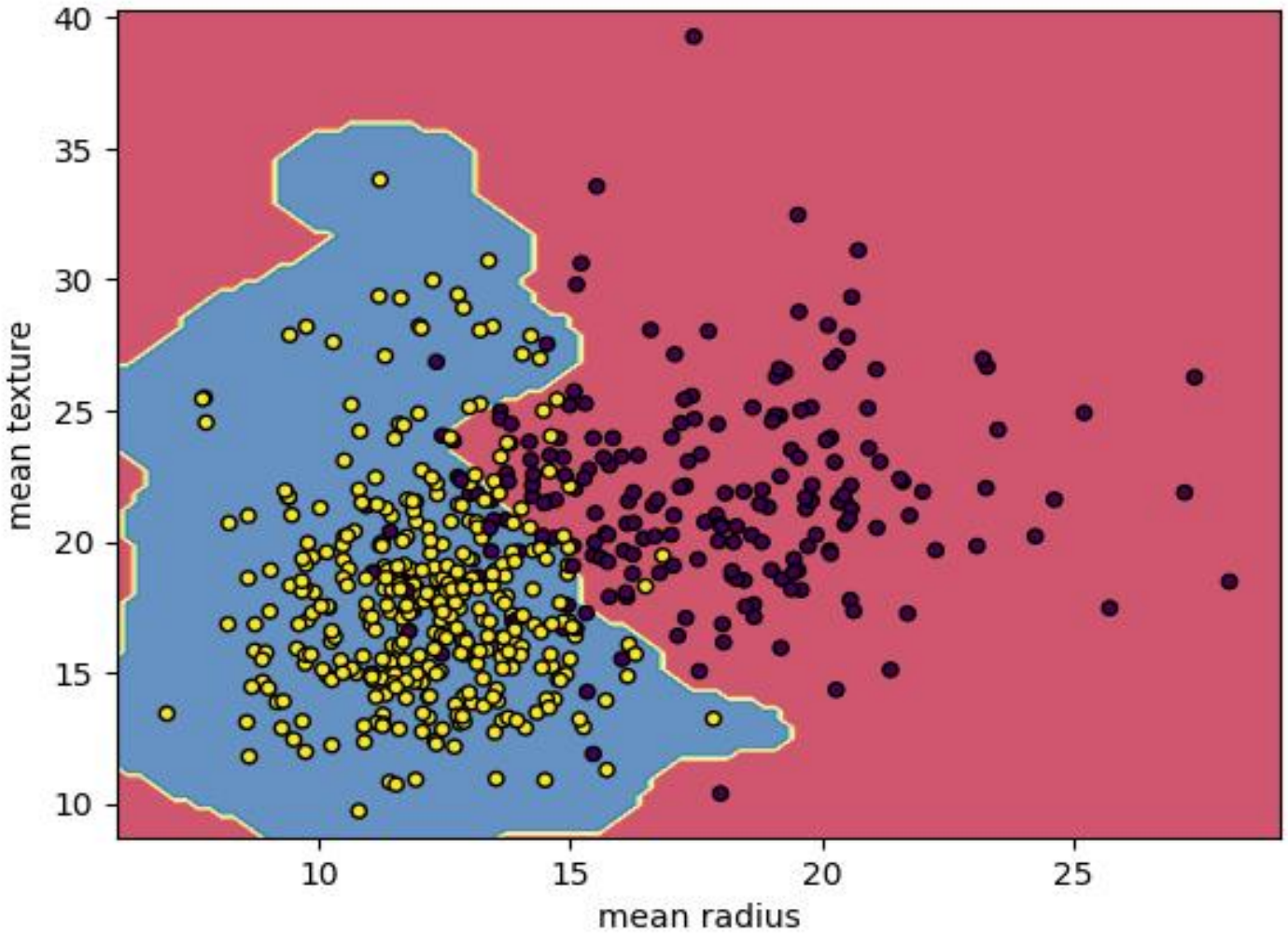
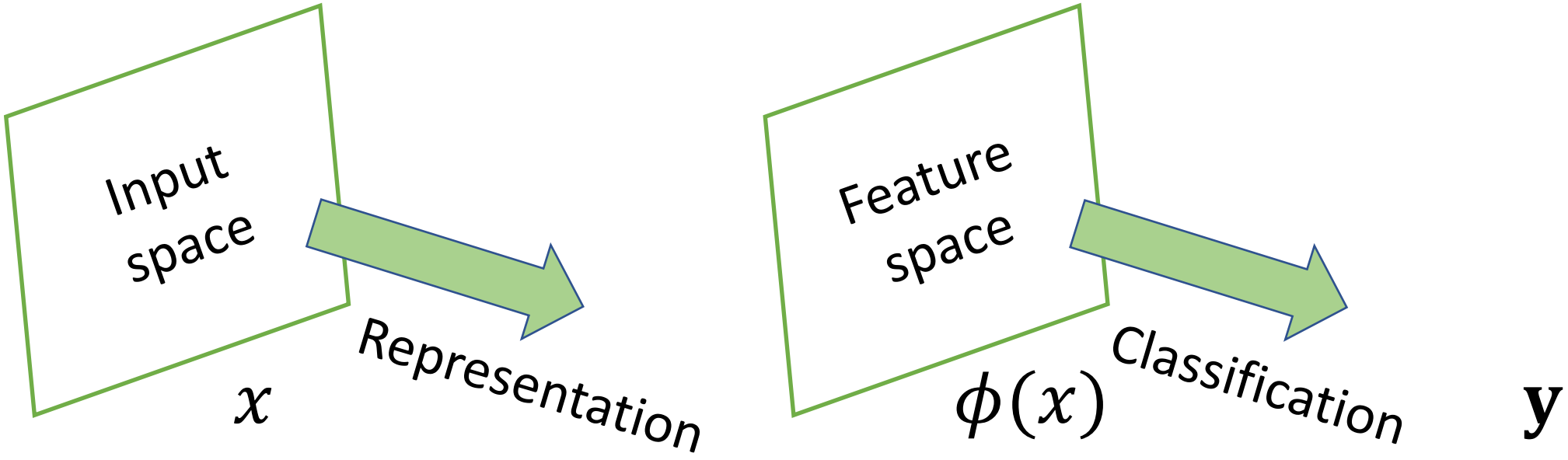


Image: https://www.geeksforgeeks.org/support-vector-machine-algorithm/



Using kernels is one way to learn non-linear features:

Cons: (1) The Kernel matrix has size $n \times n$

(2) Predefined, not data-dependant (not exactly).

Further Reading

- $Shalev - Shwartz\ and\ Ben\ David's\ book$, chapter 16
- $Bishop's\ book$, chapter 6
- ESL, chapter 12.3