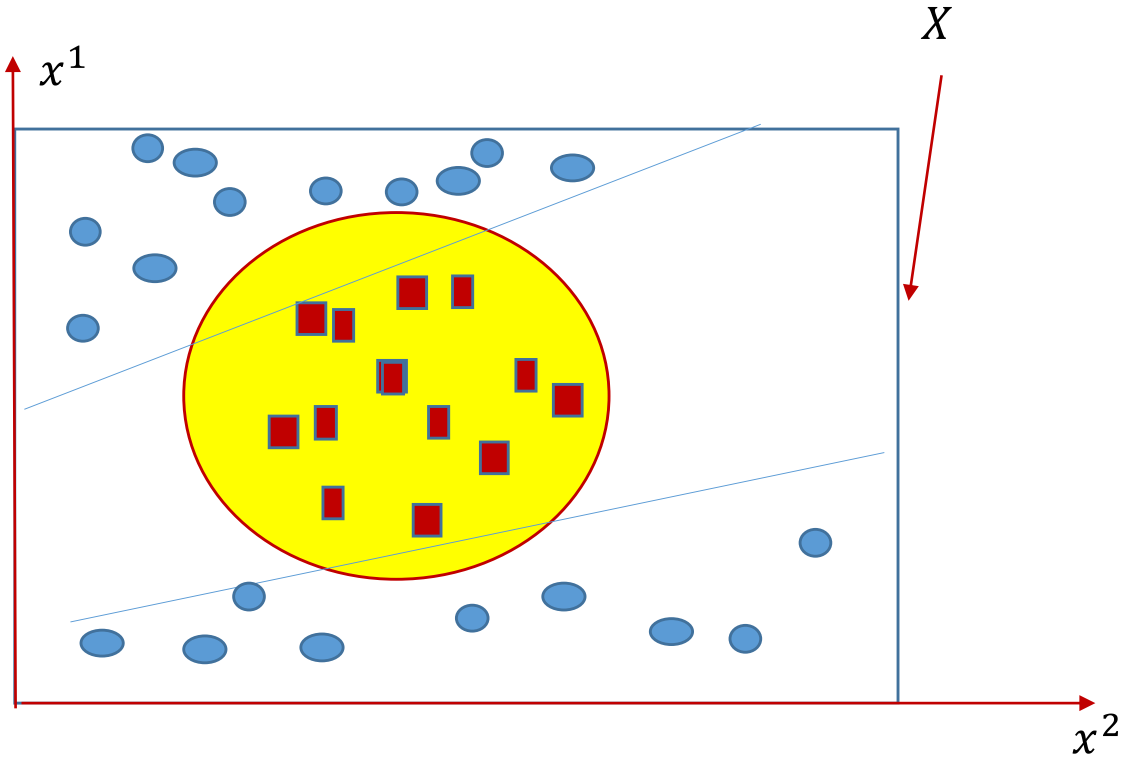
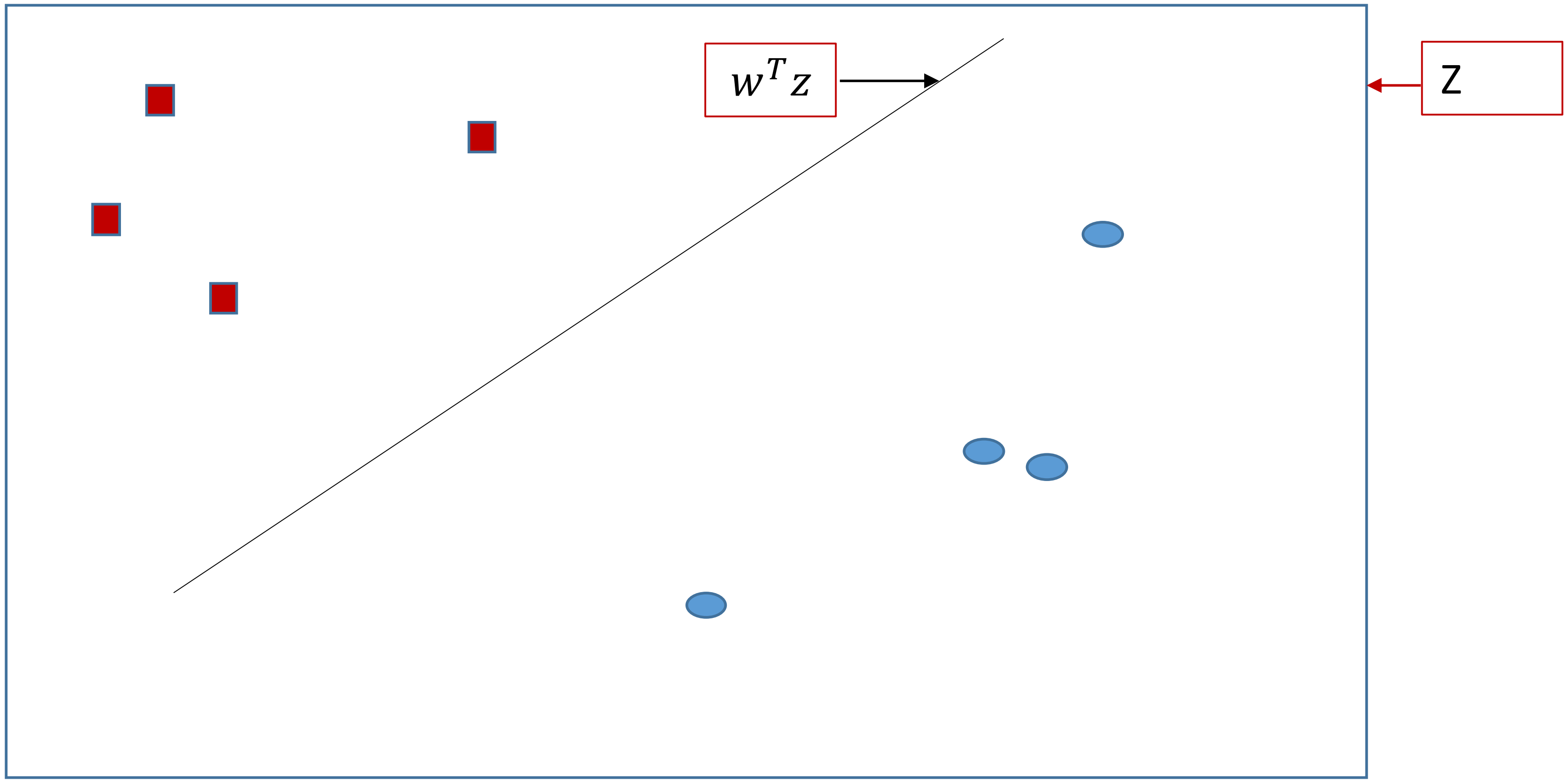
Classification
Non-linear transformations

Nonlinear Transformation from space X to space Z

- $\{x_i, y_i\}_{i=1}^{m}\ x_i = \left(x_i^1, x_i^2\right), y_i \in \{red, bleu\}$
- $h_{c,r}(x_i) = -\mathrm{r} + \left(\boldsymbol{x_i^1} - c\right)^2 + \left(\boldsymbol{x_i^2} - \boldsymbol{c}\right)^2 = \boldsymbol{z_0} + z_1 + z_2$
- $\boldsymbol{h}(x_i) = \boldsymbol{sign}\left(h_{c,r}(x_i)\right)$
- $\varphi: X \to Z$ such that $\boldsymbol{\varphi}(x) = \boldsymbol{z}$
- φ is an operator of the transformation from X to Z
- $\boldsymbol{h}(\boldsymbol{z}) = \boldsymbol{sign}\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\varphi}(x)\right) = \boldsymbol{sign}\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{z}\right)$



Nonlinear Transformation from space X to space Z



Nonlinear transformations: Space X to Space Z

Consider non-linearly separable data: $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}) \in \boldsymbol{X} \subset \mathbb{R}^2$

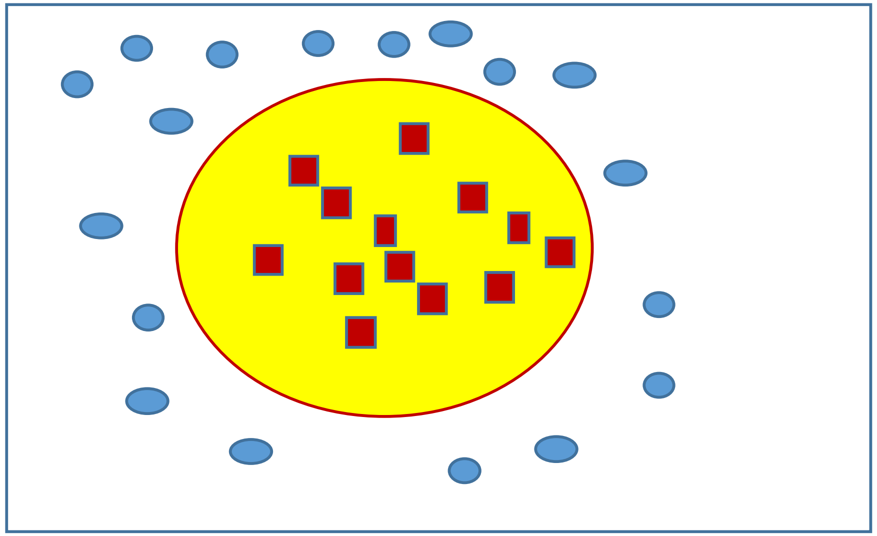
The circle presents the following equation: $C_{c,r}(x) = \boldsymbol{x_1^2} + \boldsymbol{x_2^2} - \boldsymbol{0.6} = \boldsymbol{0}$

- $\Longrightarrow \boldsymbol{c} = \boldsymbol{0}, \boldsymbol{r} = \sqrt{\boldsymbol{0,6}} \Longrightarrow$ the hypotheses set is $\boldsymbol{H} = \{C_{c,r}: \boldsymbol{c}, \boldsymbol{r} \in \mathbb{R}^2\}$
- Thus the activation function is :

$$\boldsymbol{h_{c,r}(x) = sign(-0.6 + x_1^2 + x_2^2)}$$

After applying a non-linear transformation φ to the x_i.

- In particular, consider $z_0 = 1$, $z_1 = x_1^2$ and $z_2 = x_2^2$: $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}) \rightarrow \boldsymbol{z} = (\boldsymbol{z_0}, \boldsymbol{z_1}, \boldsymbol{z_2})$



$$\boldsymbol{h(x) = sign\left(w_0.\underbrace{x_0}_{z_0} + w_1.\underbrace{(x_1)^2}_{z_1} + w_2.\underbrace{(x_2)^2}_{z_2}\right)}$$

$$\Longrightarrow \boldsymbol{h(\varphi(x)) = h(z) = sign(\tilde{w}_0 z_0 + \tilde{w}_1 z_1 + \tilde{w}_2 z_2)}$$

Hyperplan: $\boldsymbol{\tilde{w}_0 z_0 + \tilde{w}_1 z_1 + \tilde{w}_2 z_2 \rightarrow d_{CV} = 4}$

Nonlinear transformations: Space X to Space Z

- $\boldsymbol{x = (x_1, x_2)}$ ***if we have the linear separtor*** then $\boldsymbol{d_{CV} = 3 \rightarrow PLA}$

- ***Else:***
  - $\boldsymbol{h_{c,r}(x) = sign(-0.6 + x_1^2 + x_2^2)\ w_0 = -0.6, w_1 = 1, w_2 = 1}$
  - $\boldsymbol{h_{c,r}(x) = \begin{cases} 0 \ \ if - 0.6 + x_1^2 + x_2^2 \leq 0 \\ 1 if - 0.6 + x_1^2 + x_2^2 > 0 \end{cases}}$
  - $\boldsymbol{h_{c,r}(x) = \begin{cases} 0 \ \ if\ interior \\ 1\ if\ exterior \end{cases}}$
  - $\boldsymbol{x_0 = 1}$
  - $\boldsymbol{h_{c,r}(x) = h(x) = sign\left(w_0 . \underbrace{x_0}_{z_0} + w_1 . \underbrace{(x_1)^2}_{z_1} + w_2 . \underbrace{(x_2)^2}_{z_2}\right)}$
  - $\boldsymbol{\Longrightarrow h(z) = sign(\tilde{w}_0 z_0 + \tilde{w}_1 z_1 + \tilde{w}_2 z_2) \rightarrow hyperplan: PLA \rightarrow d_{CV} = 4}$

Polynomial transformation

$$\phi_Q: \quad X \rightarrow Z$$
$$x = (x_1, \dots, x_d) \rightarrow \phi_Q(x) = z$$

such that

$$\phi_Q(x) = z = \begin{pmatrix} 1, \\ x_1, \dots, x_d, \\ x_1^2, x_1x_2, \dots, x_d^2, \\ \dots \\ x_1^Q, x_1^{Q-1}x_2, \dots, x_d^Q \end{pmatrix}$$

- $\boldsymbol{Q = 2}$
- $\boldsymbol{x = (x_1, x_2) \in X}$
- $\boldsymbol{\phi_2(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2) = (z_0, z_1, z_2, z_3, z_4, z_5) = z \in Z}$

Polynomial transformation

- $\boldsymbol{Q = 3}$
- $\boldsymbol{x = (x_1, x_2, x_3) \in X}$
- $\boldsymbol{\phi_2(x) = \left(1, x_1, x_2, x_3, x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, , x_3^2\right) = z \in Z}$
- $\boldsymbol{z = (z_0, z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8, z_9)}$
- $\boldsymbol{\rightarrow hyperplan: PLA \rightarrow d_{CV} = 10}$

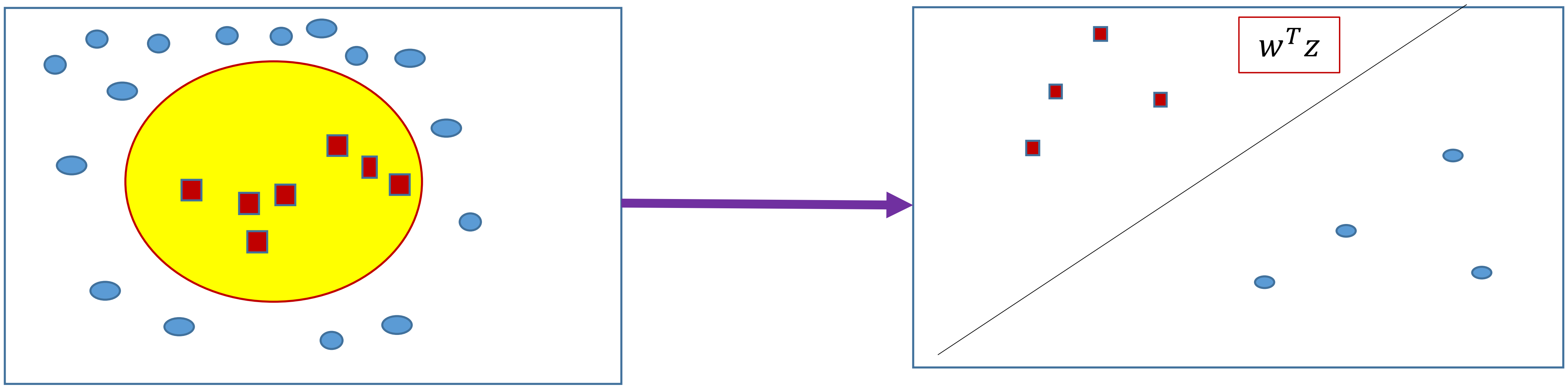
- $x \in X \subseteq \mathbb{R}^d \rightarrow x = (x_1, \cdots, x_d)$
- $z \in Z \subseteq \mathbb{R}^r \ \ avec\, r \geq d,\ \mathrm{z} = (z_1, \cdots, z_r)$ $\color{red}{\phi_Q(X) = Z}$

- $d = 2, Q = 2 \Longrightarrow x = (1, x_1, x_2), \phi_2(x) = \begin{pmatrix} 1 \\ x_1, x_2 \\ x_1^2, x_1\, x_2, x_2^2 \end{pmatrix} \Longrightarrow r = 6$

Nonlinear Transformation from space X to space Z

$$h(x) = sign(\tilde{w}^T z)$$

$\tilde{w} = (\tilde{w}_0, \ldots, \tilde{w}_{\tilde{d}})$is the weight in the space Z, and $\tilde{d}$ is its dimension. Thus, the data can be presented in terms of z instead of x

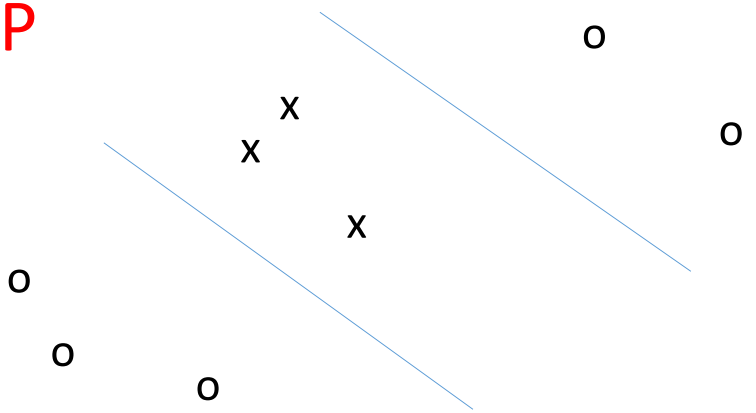


- The space Z containing the vectors z, is called the space of Features.
- The transform Φ that binds X to Z is called "the transform of Features ": $Z = \Phi(X)$

Perceptron Learning Algorithm

- If the data is linearly separable then
  - PLA converges
  - $d_{CV} = n + 1 < \infty \Longrightarrow$ we have the learning: (PAC, APAC, CU)

- Else MLP



Nonlinear Transformation from space X to space Z

The general form of the polynomial transform of features is:

$$\phi_Q(x) = \begin{pmatrix} 1, \\ x_1, \dots, x_d, \\ x_1^2, x_1 x_2, \dots, x_d^2, \\ \dots \\ x_1^Q, x_1^{Q-1} x_2, \dots, x_d^Q \end{pmatrix}$$

Q is the order of the transform.

For the case of the circle hypothesis in X: $x = (x_1, x_2) \to \mathrm{z} = \boldsymbol{\Phi}(\boldsymbol{x}) = (\mathbf{1}, \boldsymbol{x}_\mathbf{1}^\mathbf{2}, \boldsymbol{x}_\mathbf{2}^\mathbf{2})$

All nonlinear assumptions h (cercle) in space X can be presented by a linear hypothesis (hyperplan) in space Z :

$$h(x) = \tilde{h}(\boldsymbol{z}) = \tilde{h}(\boldsymbol{\Phi}(\boldsymbol{x}))$$

Nonlinear Transformation from space X to space Z

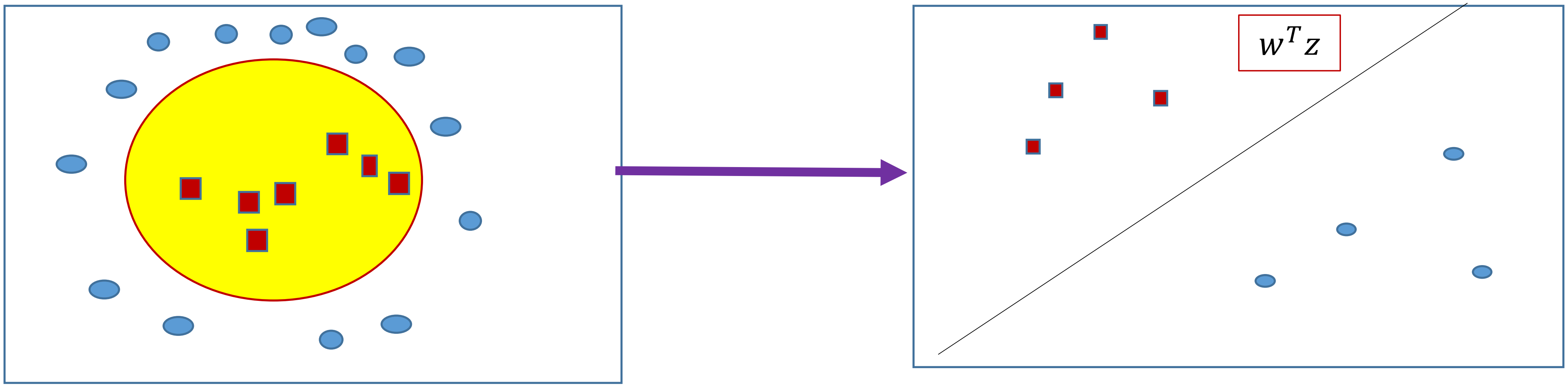
- All transformed data $(z_1, y_1), (z_2, y_2), \dots, (z_m, y_m)$ are linearly separable in Z.

- Then, one can apply PLA on transformed data to get $\widetilde{w}_{PLA}$:

$$g(z) = sign(\widetilde{w}_{PLA}^T z)$$

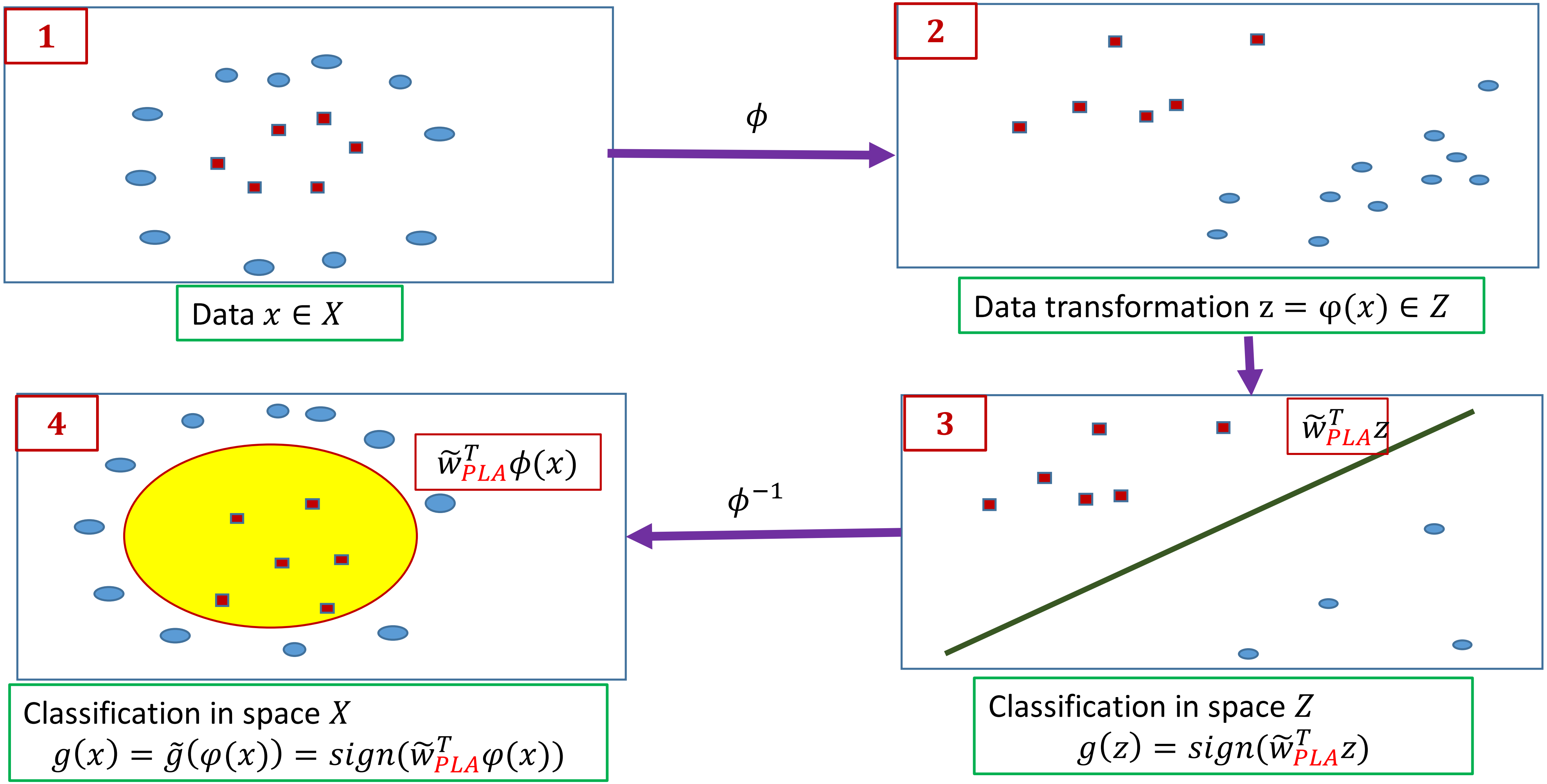
- Empirical error in space X is the same as that in the Features the space Z, so that :

$$L_S(g) = 0$$



Nonlinear Transformation from space X to space Z

The process of transforming characteristics for linear classification:



Nonlinear Transformation from space X to space Z

- $\mathrm{z} = \boldsymbol{\Phi}'_2(\boldsymbol{x}) = \left(\mathbf{1}, \boldsymbol{x}_{\mathbf{1}}^{\mathbf{2}}, \boldsymbol{x}_{\mathbf{2}}^{\mathbf{2}}\right) \rightarrow \boldsymbol{d}_{\boldsymbol{CV}} = \mathbf{3}$
- $\mathrm{z} = \boldsymbol{\Phi}_2(\boldsymbol{x}) = (\mathbf{1}, \boldsymbol{x}_{\mathbf{1}}^{\mathbf{2}}, \boldsymbol{x}_{\mathbf{1}}\boldsymbol{x}_{\mathbf{2}}, \boldsymbol{x}_{\mathbf{2}}^{\mathbf{2}}) \rightarrow \boldsymbol{d}_{\boldsymbol{CV}} = \mathbf{4}$
- **If we have overfitting we reduce the degree of polynomial**

- $x \in X \subseteq \mathbb{R}^d \rightarrow x = (x_1, \cdots, x_d)$
- $z \in Z \subseteq \mathbb{R}^r \;\; avec\; r = f(d)$

Nonlinear Transformation from space X to space Z

Let's go back to the example of which: $\mathrm{z} = \boldsymbol{\phi}(\boldsymbol{x}) = \left(\mathbf{1}, \boldsymbol{x}_{\mathbf{1}}^{\mathbf{2}}, \boldsymbol{x}_{\mathbf{2}}^{\mathbf{2}}\right) = (\mathbf{1}, \boldsymbol{x}_{\mathbf{1}}^{\mathbf{2}}, \boldsymbol{x}_{\mathbf{1}}\boldsymbol{x}_{\mathbf{2}}, \boldsymbol{x}_{\mathbf{2}}^{\mathbf{2}})$

We have: $Z = \{1\} \times \mathbb{R}^2$

Since H_{Φ} is the set of assumptions of a perceptron in Z, therefore:

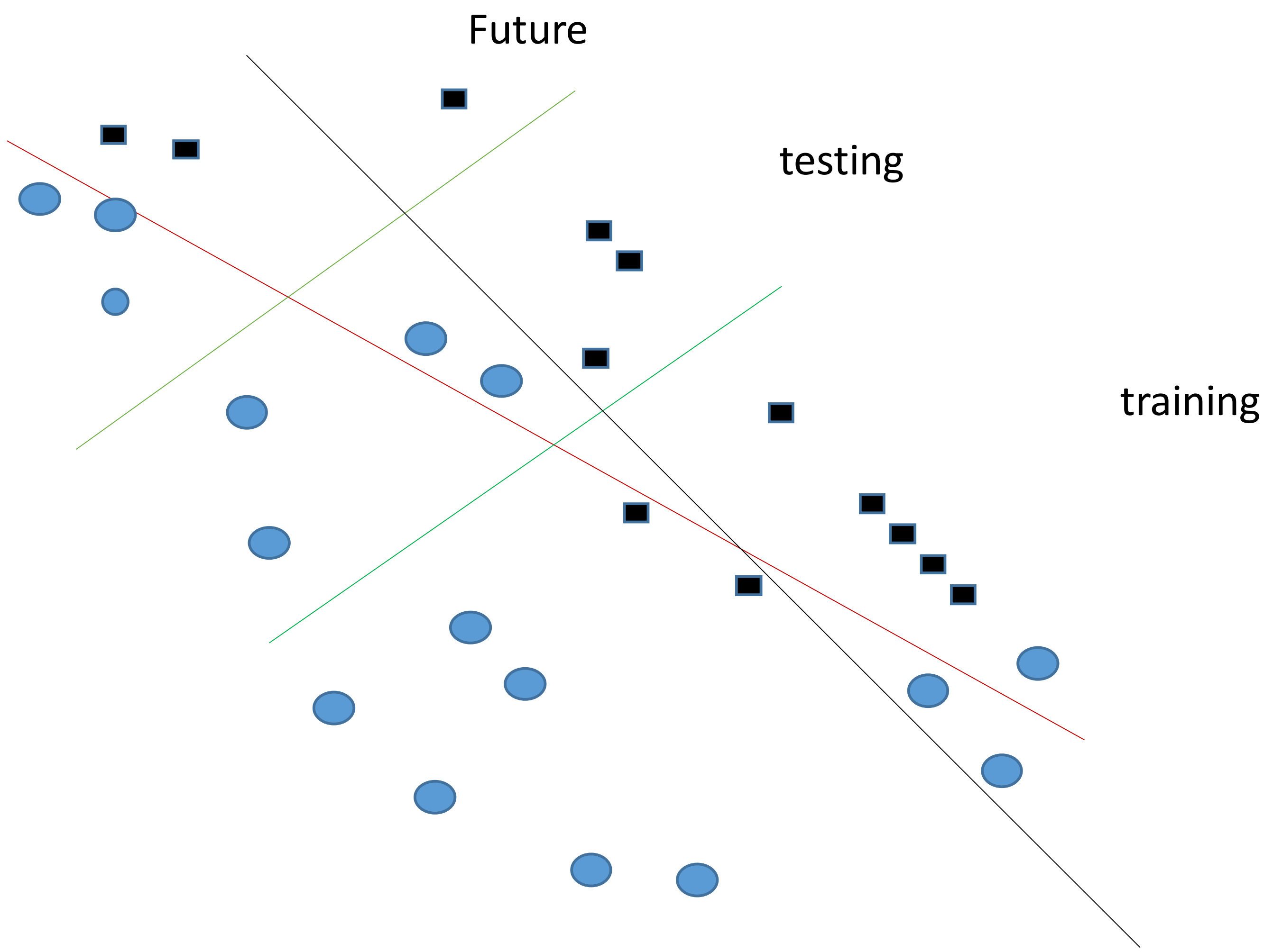
$$\boldsymbol{d_{VC}(H_{\phi}) \leq 3} \textbf{ and not } \boldsymbol{d_{VC}\left(H_{\phi}\right) = 3}$$

? Because there are $\boldsymbol{x \in X}$ that don't have correct transforms in $\boldsymbol{Z}$.

Remark:

The transformation of the Features must be chosen with care:

- You have to choose ϕ before seeing the data or running the algorithm, so as not to fall into overfitting.
- We must not insist on linear separation and then use a very complex hypothesis.

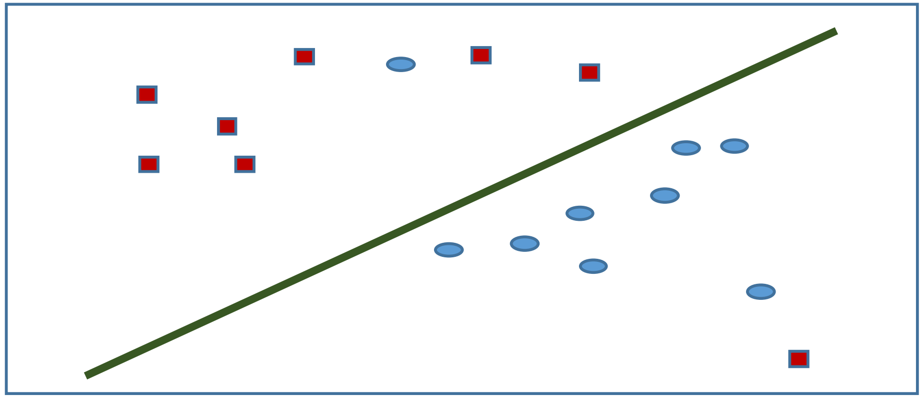

Future
testing
training

Nonlinear Transformation from space X to space Z

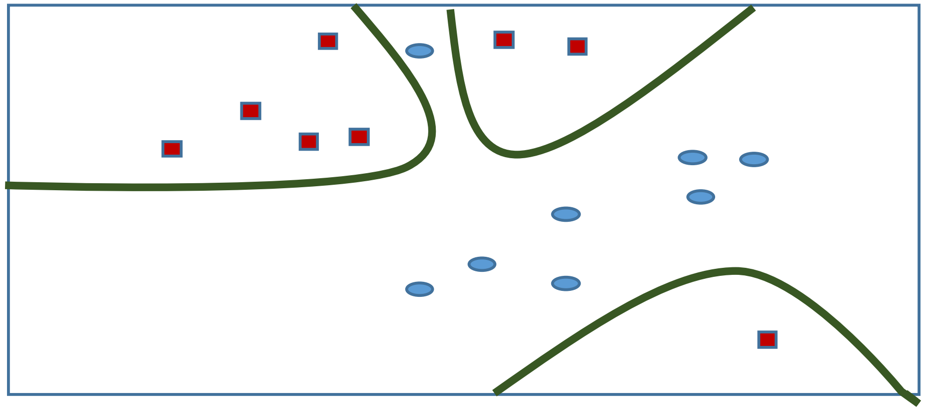
- There is no line that can perfectly separate the training data, neither quadratic curve nor polynomial curve of order three.
- For this, it is necessary to use a polynomial transform of order $Q = 4$:

$$\phi_4(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, x_1^4, x_1^3x_2, x_1^2x_2^2, x_1x_2^3, x_2^4)$$

In that case $\tilde{d} = 14$



Linear Hypothesis

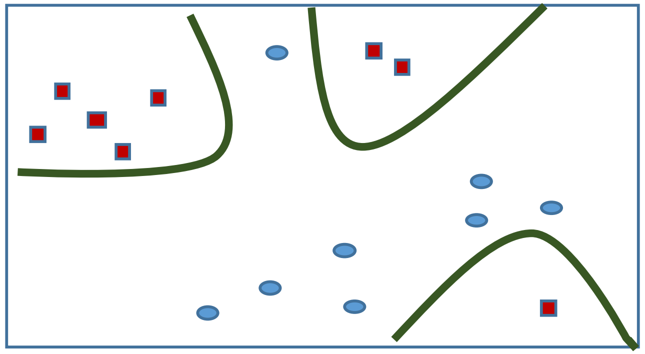


Polynomial Hypothesis of order 4

Nonlinear Transformation from space X to space Z

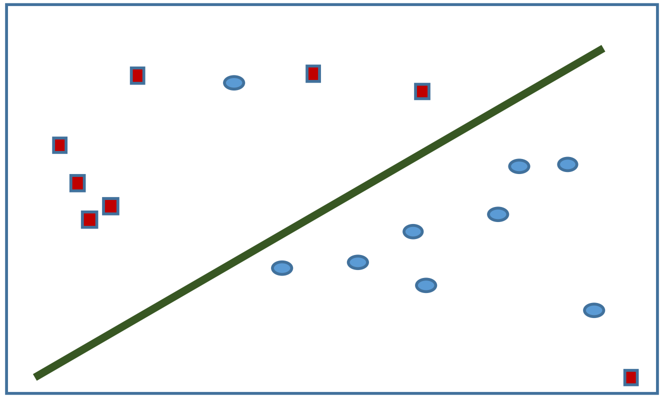
- This figure shows that the data have been overestimated:
  - The capacity for generalization(error ↑) ↓
  - The ability to approximate(error ↓) ↑

$$L_S = 0$$



- The best solution is to ignore the two poorly ranked points:
  - The capacity for generalization ↑
  - The ability to approximate ↓

$$L_S \neq 0$$



Remark: Empirical error must be tolerated while choosing a very simple hypothesis.

Limits: The use of a polynomial transform of very large order Q, gives us a lot of flexibility in terms of the form of decisions in X. But, there is a price to pay. This price is that of:

- Misgeneralization.
- Computational complexity.

Computational complexity

- Calculation is a big problem.

- The transform Φ_Q transforms: $\Phi_Q: X \rightarrow Z$ such that $x = (x_1, \dots, x_d) \in X$

A two-dimensional vector x ($\boldsymbol{d_X = 2}$) into a dimension vector:

$$\boldsymbol{z \in Z, d_Z = \tilde{d} = \frac{Q(Q+3)}{2} \; if \; Q = 2 \rightarrow \tilde{d} = \frac{2(2+3)}{2} = 5?}$$

- This increases the complexity of the computing memory.

- Things can get worse if the dimension of x is very large

Misgeneralization

- If Φ_Q is the transform of a two-dimensional input space, it will exist $\tilde{d} = \frac{Q(Q+3)}{2}$ dimensions in Z:

$$\boldsymbol{d_{VC}(H_Q)} \textbf{ will be close to } \quad \boldsymbol{\tilde{d} + 1 = \frac{Q(Q+3)}{2} + 1}$$

- This means that the second term of the generalization limit can increase significantly:

$$|L_D(h) - L_S(h)| \leq \frac{4 + \sqrt{\log\left(\Pi_H(2m)\right)}}{\delta\sqrt{2m}}$$

Example:

If$\Phi = \Phi_{50}$ is used, the dimension $d_{VC}(H_Q)$ will be approximately equal to :

$$\tilde{d} + 1 = \frac{50(50+3)}{2} + 1 = 1326$$

Misgeneralization

According to the fundamental theorem of learning, we have:

$$C_1 \frac{d_{VC} + \log(\frac{1}{\delta})}{\varepsilon^2} \leq m_H^{APAC} \ (\varepsilon, \delta) \leq C_2 \frac{d_{VC} + \log(\frac{1}{\delta})}{\varepsilon^2}$$

Since $d_{VC}(H_Q)$ is very large, so we will need thousands of data compared to the case where we do not use transformation.

When choosing the dimension of the transform of the characteristics, one cannot avoid the compromise approximation /generalization:

- **Approximation :**

  A very large $\tilde{d}$ $(\mathbf{Q}\uparrow) \rightarrow (L_S)\downarrow$ and $(d_{VC})\uparrow$

- **Generalization :**

  A very small $\tilde{d}(\mathbf{Q}\downarrow) \rightarrow (L_S)\uparrow$ and $(d_{VC})\downarrow$?

Example

A line hardly separates the number 1 from the other digits, but a curve can do better.