Machine Learning Algorithms

Optimization Problems
for Machine Learning

Markus Rupp

13.9.2020

Problem Formulation

- Given two sets S1 and S2 with presence and absence of feature A, respectively, we try to find a minimum (maximum)

$$\min_{\underline{\overset{\frown}{w}}} \sum_{\underline{x}_i \in S_1} \left(1 - \underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i\right)^2 + \sum_{\underline{x}_i \in S_2} \left(-1 - \underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i\right)^2$$

- As w is underdetermined, we also like to constrain it
- For example the smallest, which can be achieved by adding an additional constraint

$$\min_{\underline{\overset{\frown}{w}}} \sum_{\underline{x}_i \in S_1} \left(1 - \underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i\right)^2 + \sum_{\underline{x}_i \in S_2} \left(-1 - \underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i\right)^2 + \lambda \left\| \underline{\overset{\frown}{w}} \right\|^2$$

Problem Formulation

- This has an additional advantage, as often saddle points occur when nonlinear formulations σ(x) are included, e.g.,

$$\min_{\underline{\tilde{w}}} \sum_{i \in S_1} \left(1 - \sigma\left(\underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i\right)\right)^2 + \sum_{i \in S_2} \left(-1 - \sigma\left(\underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i\right)\right)^2 + \lambda \left\| \underline{\overset{\frown}{w}} \right\|^2$$

- With the additional term the saddle point is often deformed into a continuous monotone region which can be searched through with a gradient approach.

Interpretation

- The additional term can be interpreted in different ways
- 1) additive weight with $\lambda>0$
- 2) Lagrangian Multiplier
- 3) reformulated problem with side constraints

Interpretation 1: additive weight

- In a linear classification setting, adding a positive term changes the LS solution :

$$R=\sum_{\underline{\hat{x}}_i\in S}\underline{\hat{x}}_i\underline{\hat{x}}_i^H \quad\rightarrow\quad R=\lambda I+\sum_{\underline{\hat{x}}_i\in S}\underline{\hat{x}}_i\underline{\hat{x}}_i^H$$

- Regularisation term to ensure positive definiteness of R.
  A very small $\lambda>0$ will do!

Interpretation 1: additive weight

- Consequence for the gradient approach:

$$\underline{\hat{w}}_k = (1-\lambda)\underline{\hat{w}}_{k-1} + \mu \underline{x}_i (y_i - \underline{\hat{w}}_{k-1}^H \underline{x}_i)$$

- LMS obtains a „leaky“ term,
  forgetting a bit of its previous estimate

Interpretation 2: Lagrangian Multiplier

- In this case λ is interpreted as Lagrangian multiplier and thus the minimum of $\underline{\hat{w}}$ becomes a constraint.
- Under all minimal $\underline{\hat{w}}$ we search for those that minimize the LS approach.

$$\frac{\partial}{\partial \underline{\hat{w}}} \sum_{\underline{\hat{x}}_i \in S_1} \left(1-\underline{\hat{w}}^H \underline{\hat{x}}_i\right)^2 + \sum_{\underline{\hat{x}}_i \in S_2} \left(-1-\underline{\hat{w}}^H \underline{\hat{x}}_i\right)^2 + \lambda \left\|\underline{\hat{w}}\right\|^2 = \underline{0}$$

$$-2\underline{p}+2R\underline{\hat{w}}+2\lambda\underline{\hat{w}}=0$$

$$\left(R+\lambda I\right)\underline{\hat{w}}=\underline{p}$$

Nonlinear in λ
Can be solved numerically

$$\min_{\lambda} \sum_{\underline{\hat{x}}_i \in S_1} \left(1-\underline{p}^H \left(R+\lambda I\right)^{-1} \underline{\hat{x}}_i\right)^2 + \sum_{\underline{\hat{x}}_i \in S_2} \left(-1-\underline{p}^H \left(R+\lambda I\right)^{-1} \underline{\hat{x}}_i\right)^2 + \lambda \underline{p}^H \left(R+\lambda I\right)^{-2} \underline{p}$$

Interpretation 3: Reformulation

$$\min \left\| \underline{\overset{\frown}{w}} \right\|^2$$

subject to

$$\underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i = y_i; \quad i = 1, 2, \ldots$$

$$\min \frac{1}{2} \left\| \underline{\overset{\frown}{w}} \right\|^2 + \sum \lambda_i \left(\underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i - y_i \right)$$

- Which turns out to be an other form of a Lagrangian optimization problem

Interpretation 3: Reformulation

$$\frac{\partial}{\partial \underline{\overset{\frown}{w}}}\left\{\frac{1}{2}\left\|\underline{\overset{\frown}{w}}\right\|^2+\sum \lambda_i\left(\underline{\overset{\frown}{w}}^H\underline{\overset{\frown}{x}}_i-y_i\right)\right\}=\underline{0}$$

$$\underline{\overset{\frown}{w}}+\sum \lambda_i\underline{\overset{\frown}{x}}_i=\underline{0}$$

$$\min\frac{1}{2}\left\|\sum \lambda_i\underline{\overset{\frown}{x}}_i\right\|^2-\left\|\sum \lambda_i\underline{\overset{\frown}{x}}_i\right\|^2-\sum \lambda_i y_i$$

$$=\min\frac{1}{2}\left\|\sum \lambda_i\underline{\overset{\frown}{x}}_i\right\|^2+\sum \lambda_i y_i$$

Interpretation 3: Reformulation

- Which can be explicitly solved for λ_i:
- Consider Gramian R built from $\underline{x}_i$

$$X = \left[\underline{\overset{\frown}{x}}_1, \underline{\overset{\frown}{x}}_2, \ldots, \underline{\overset{\frown}{x}}_N\right] \to R = X^H X$$

$$R\underline{\lambda} = -\underline{y} \to \underline{\lambda} = -R^{-1}\underline{y}$$

$$\frac{1}{2}\left\|\underline{\overset{\frown}{w}}\right\|^2 + \sum \lambda_i \left(\underline{\overset{\frown}{w}}^H \underline{\overset{\frown}{x}}_i - y_i\right)$$

$$= \frac{1}{2}\left\|\underline{\overset{\frown}{w}}\right\|^2 + \underline{\lambda}^H \left(X\underline{\overset{\frown}{w}} - \underline{y}\right)$$

$$= \frac{1}{2}\left\|\underline{\overset{\frown}{w}}\right\|^2 - \underline{y}^H \left[X^H X\right]^{-1} \left(X\underline{\overset{\frown}{w}} - \underline{y}\right)$$

Interpretation 3: Reformulation

- Now solving with respect to $\overset{\frown}{\underline{w}}$ is simply

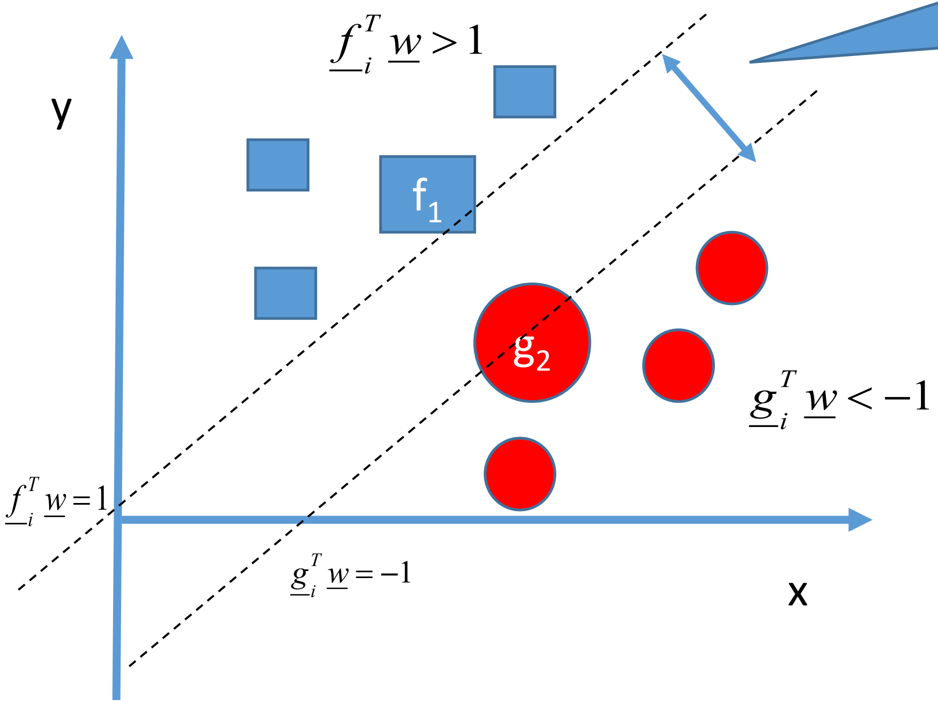
$$\min_{\overset{\frown}{\underline{w}}} \frac{1}{2}\left\|\overset{\frown}{\underline{w}}\right\|^2 - \underline{y}\left[X^H X\right]^{-1}\left(X\overset{\frown}{\underline{w}} - \underline{y}\right)$$

$$\overset{\frown}{\underline{w}} = X^H\left[X^H X\right]^{-1}\underline{y}$$

Note that this is an underdetermined LS solution which would only work perfectly if the observations are less than parameters, which is **not** the case!

We need a better Re-Formulation: Let‘s go deeper (Vapnik)

- We do not want any hyperplane that separates two classes but an optimal line that guarantees that the border members of the classes are at maximum distance.



Let us maximize this distance here

$$\left\langle \underline{w}, \underline{f}_1 - \underline{g}_2 \right\rangle = 1-(-1) = 2$$

$$\left\langle \frac{\underline{w}}{\|\underline{w}\|}, \underline{f}_1 - \underline{g}_2 \right\rangle = \frac{2}{\|\underline{w}\|}$$

which means minimizing $||\underline{w}||$

$\underline{f}_1$ and $\underline{g}_2$ are supporting these separation lines
→Vector support machine

We need a better Re-Formulation: Let‘s go deeper (Vapnik)

- Let us do this again but with more subtleties
- Recall:

$$S_1: \quad \overset{\frown}{\underline{w}}^H \overset{\frown}{\underline{x}}_i \geq 1 \quad \rightarrow y_i = 1$$

$$\underline{w}^H \underline{x}_i + b \geq 1$$

$$y_i\left(\underline{w}^H \underline{x}_i + b\right) \geq 1$$

$$S_2: \quad \overset{\frown}{\underline{w}}^H \overset{\frown}{\underline{x}}_i \leq -1 \quad \rightarrow y_i = -1$$

$$\underline{w}^H \underline{x}_i + b \leq -1$$

$$y_i\left(\underline{w}^H \underline{x}_i + b\right) \geq 1$$

Vector Support Machine

- Let‘s differentiate

Note the difference

and here

$$\min_{\overset{\frown}{\underline{w}}}\left\{\frac{1}{2}\left\|\boxed{\underline{w}}\right\|^2-\sum\lambda_i\left(\boxed{y_i}\left(\overset{\frown}{\underline{w}}^H\overset{\frown}{\underline{x}}_i\right)-1\right)\right\}$$

$$=\min_{\underline{w},b}\left\{\frac{1}{2}\left\|\underline{w}\right\|^2-\sum\lambda_i\left(\left(\underline{w}^H\underline{x}_i+b\right)y_i-1\right)\right\}$$

$$\frac{\partial}{\partial\underline{w},b}\left\{\frac{1}{2}\left\|\underline{w}\right\|^2-\sum\lambda_i\left(\left(\underline{w}^H\underline{x}_i+b\right)y_i-1\right)\right\}=\underline{0}$$

https://youtu.be/_PwhiWxHK8o

Vector Support Machine (Vapnik 1963)

- Let‘s differentiate

$$\frac{\partial}{\partial \underline{w}}\left\{\frac{1}{2}\|\underline{w}\|^2-\sum \lambda_i\left(\left(\underline{w}^H \underline{x}_i+b\right)y_i-1\right)\right\}=\underline{0}$$

$$\underline{w}=\sum \lambda_i y_i \underline{x}_i$$

$$\frac{\partial}{\partial b}\left\{\frac{1}{2}\|\underline{w}\|^2-\sum \lambda_i\left(\left(\underline{w}^H \underline{x}_i+b\right)y_i-1\right)\right\}=0$$

$$0=\sum \lambda_i y_i$$

solution by minimizing this w.r.t. λ_i:

$$-\frac{1}{2}\left\|\sum \lambda_i y_i \underline{x}_i\right\|^2=-\frac{1}{2}\sum_i\sum_j \lambda_i\lambda_j y_i y_j\left\langle \underline{x}_i,\underline{x}_j\right\rangle$$

Quadratic in λ, thus can be minimized.

Note that the minimum of the cost function depends only on the dot product https://www.youtube.com/watch?v=wlgvAar4pQE

Vector Support Machine

- Given the training data $\{\underline{x}_i, y_i\}$

$$\min_{\lambda_j} - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \left\langle \underline{x}_i , \underline{x}_j \right\rangle$$

- By maximizing the dual problem:

$$\max_{\lambda_j} \sum_j \lambda_j - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \left\langle \underline{x}_i , \underline{x}_j \right\rangle$$

Only support vectors have $\lambda_i > 0$

- That satisfies

$$\lambda_i \geq 0 \quad \wedge \quad \sum \lambda_i y_i = 0$$

Vector Support Machine

- With the support vectors compute

$$\underline{w} = \sum \lambda_i y_i \underline{x}_i$$

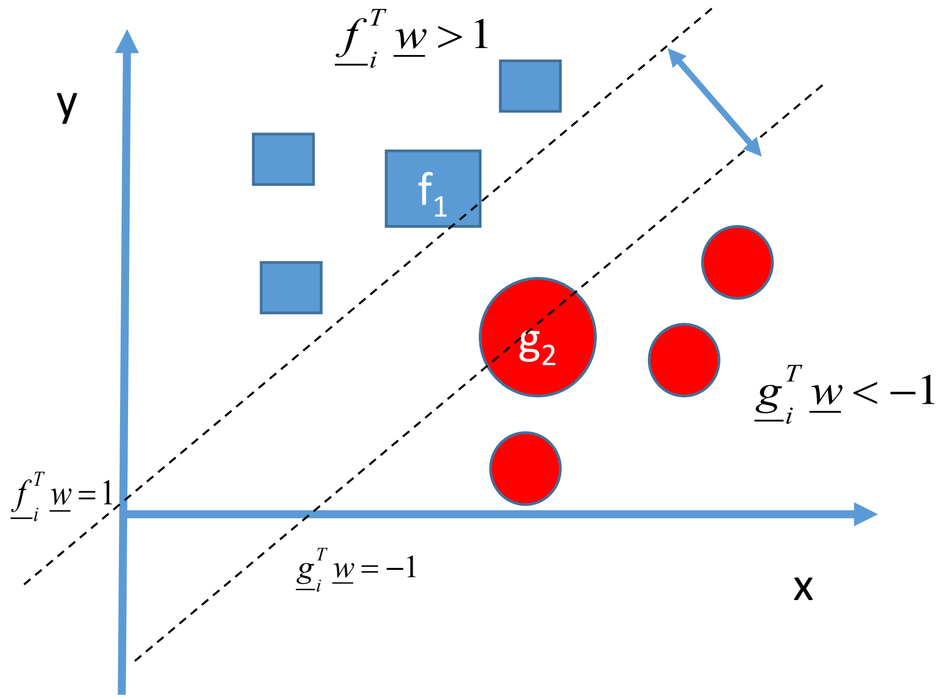
- Select support vector set $\{\underline{f}_1, y_1 = 1\}$

$$\underline{w}^T \underline{f}_1 + b = 1$$

$$b = 1 - \underline{w}^T \underline{f}_1$$

Vector Support Machine Example

- Recall our initial problem



We have two support vectors $\underline{f}_1$ and $\underline{g}_2$ with y=1 and -1, respectively

This results in $\lambda_1=\lambda_2$

We need to

$$\max_{\lambda_j} \sum_j \lambda_j - \frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \left\langle \underline{x}_i , \underline{x}_j \right\rangle$$

$$= \max 2\lambda_1 - \frac{1}{2}\lambda_1^2 \left\| \underline{f}_1 - \underline{g}_2 \right\|_2^2$$

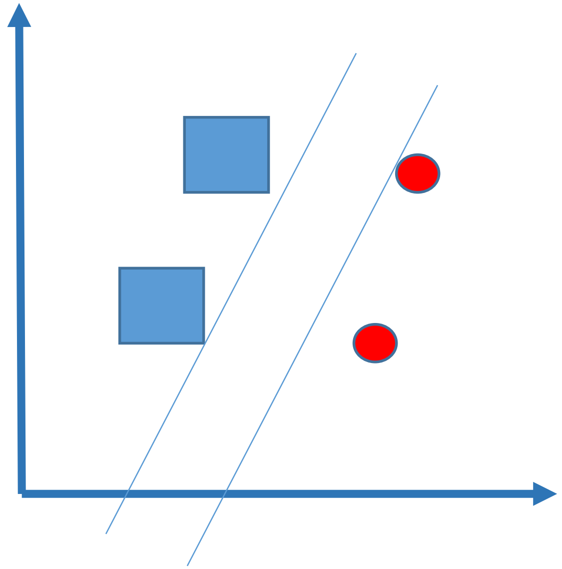
$$\lambda_1 = \lambda_2 = \frac{2}{\left\| \underline{f}_1 - \underline{g}_2 \right\|_2^2}$$

$$\underline{w} = \frac{2}{\left\| \underline{f}_1 - \underline{g}_2 \right\|_2^2} \left(\underline{f}_1 - \underline{g}_2 \right)$$

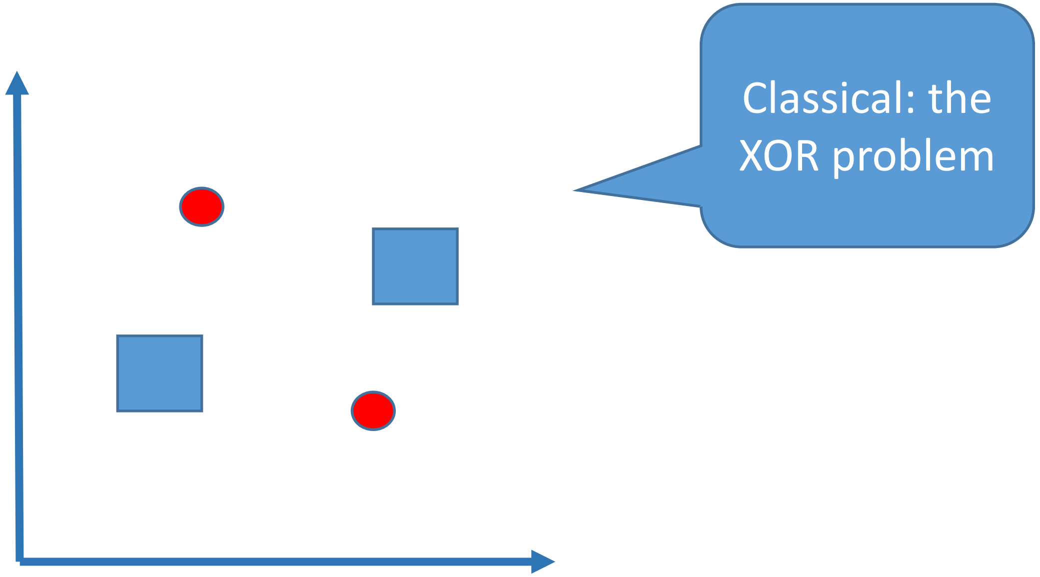
$$b = 1 - \frac{2}{\left\| \underline{f}_1 - \underline{g}_2 \right\|_2^2} \left(\underline{f}_1 - \underline{g}_2 \right)^T \underline{f}_1$$

Recall Linear Separable

- Linearly separable



non-linearly separable

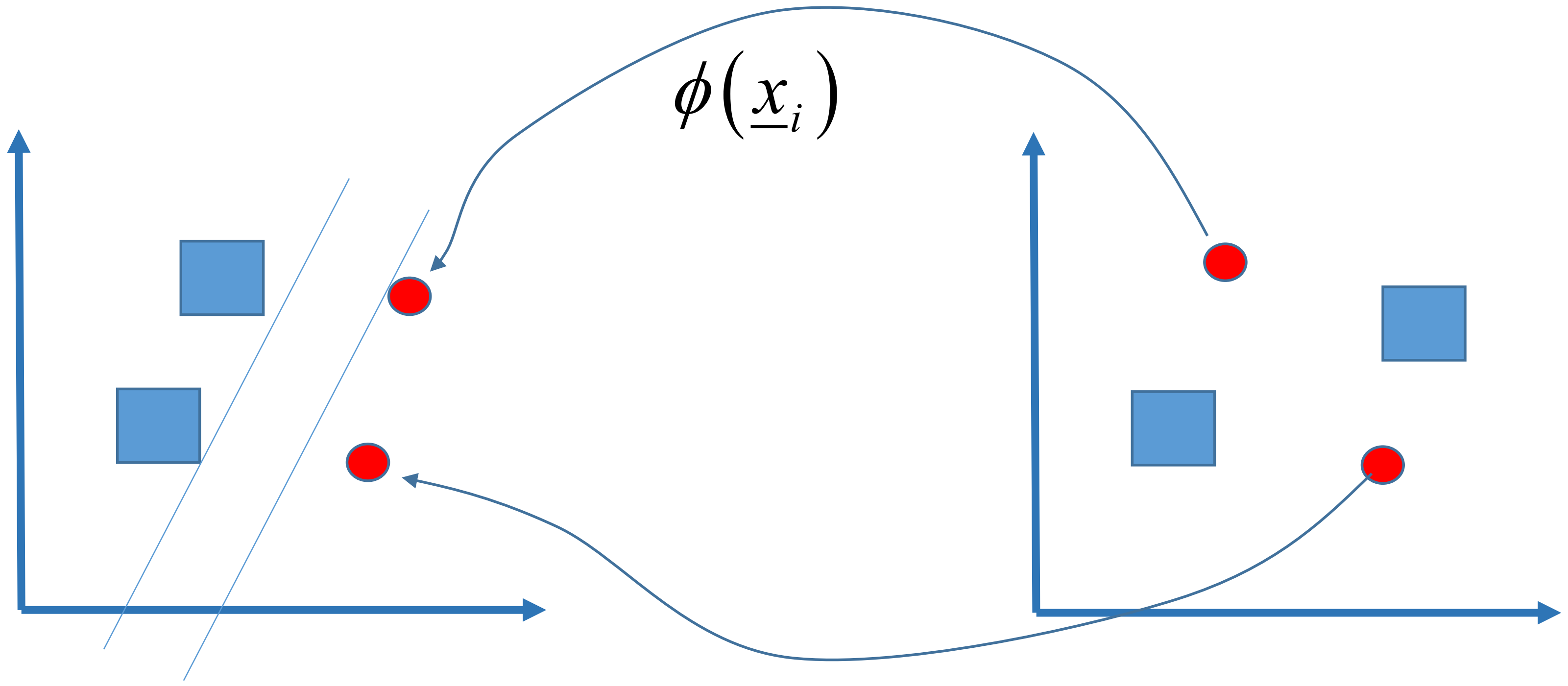


Ways to obtain linearly separable sets

- Nonlinear mapping of input vectors
- Increasing dimensions of the feature space

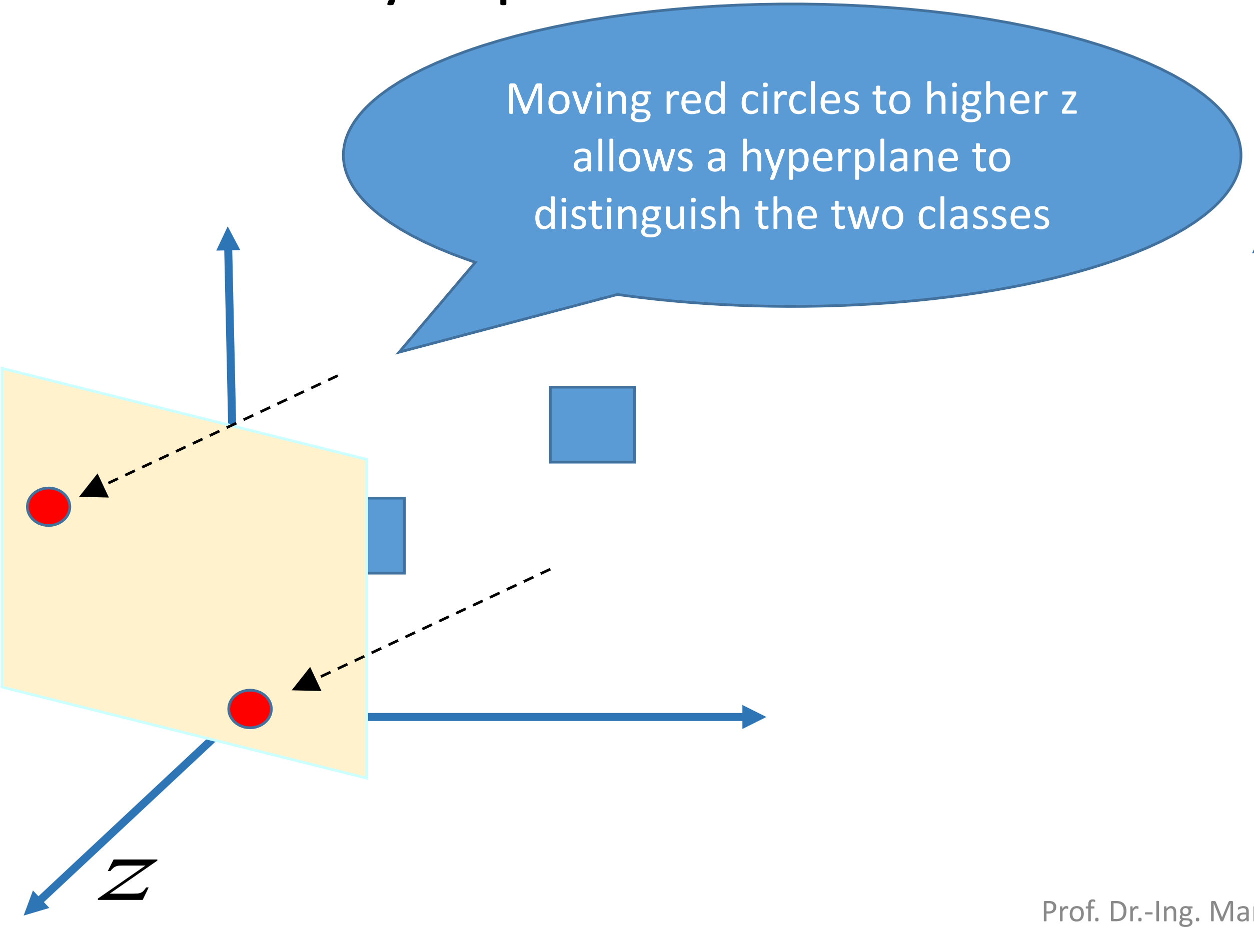
Recall Linear Separable
nonlinear mapping of input vectors

- Linearly separable   non-linearly separable

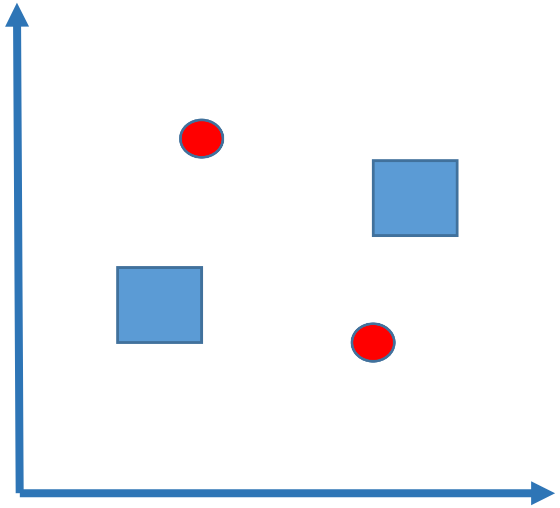


Recall Linear Separable

- Linearly separable



non-linearly separable



Moving into higher dimensions (Vapnik 1992)

- For example use a nonlinear mapping of input values

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \rightarrow \quad \phi[\underline{x}] = \begin{bmatrix} \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$$

$$\left(1 + \left\langle \underline{x}, \underline{y} \right\rangle\right)^2 - 1 = \left\langle \phi[\underline{x}], \phi\left[\underline{y}\right] \right\rangle$$

- and obtain a polynomial kernel

Now back to our original problem

$$L = -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j \left\langle \underline{x}_i , \underline{x}_j \right\rangle$$

$$= -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j K\left(\underline{x}_i , \underline{x}_j \right)$$

- Interpret dot product as a kernel
- Kernel defines the (non)-linear mapping

$$K\left(\underline{x}, \underline{y}\right) = \left\langle \phi(\underline{x}), \phi\left(\underline{y}\right)\right\rangle$$

Kernels

Vladimir Naumovich Vapnik (Russian: Владимир Наумович Вапник;
born 6 December 1936)
~1990 at AT&T Bell Labs

- $K_0(\underline{x},\underline{y}) = \tanh\left(\beta^2 \underline{x}^H \underline{y} + \alpha\right)$ Neuronal Network
- $K_1(\underline{x},\underline{y}) = \left(\underline{x}^H \underline{y} + 1\right)^n - 1$ Polynomial Kernel of degree n
- $K_2(\underline{x},\underline{y}) = e^{-\beta\|\underline{x}-\underline{y}\|^2}$ Radial Basis Function (RBF)
- $K_3(\underline{x},\underline{y}) = \prod_{n=1}^{N} \frac{\sin\left((2D+1)\pi(x_n - y_n)\right)}{\sin\left(\pi(x_n - y_n)\right)} - 1$ Fourier