Kernels

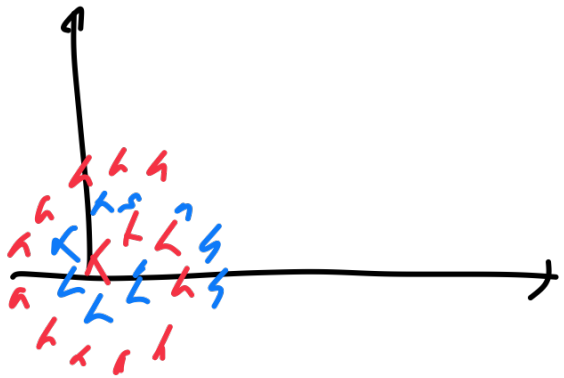
Let us consider a long feature vector

$$\tilde{\phi}(\tilde{x}^{(i)}) = (1, x^{(i)}, (x^{(i)})^2, (x^{(i)})^3, \dots, (x^{(i)})^d)$$

We want to minimize the loss

$$l(\beta) = \frac{1}{N} \sum_{i=1}^{N} (t^{(i)} - \beta^T \tilde{\phi}(x^{(i)}))^2$$

$$\beta^{(k+1)} = \beta^{(k)} + \eta \frac{1}{N} \sum_{i=1}^{N} (t^{(i)} - \beta^T \tilde{\phi}(x^{(i)})) \cdot \tilde{\phi}(x^{(i)})$$



$$X = (x_1, \dots, x_D)$$

$D = 1000$

$$\tilde{\phi}(x) = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \\ x_1^2 \\ \vdots \\ x_D^2 \\ x_1^3 \\ x_1 x_2 x_3 \\ \vdots \end{bmatrix} \longrightarrow \underbrace{1 + D + D^2 + D^3}$$

$\tilde{\phi} = 10^9$

Idea: Write $\beta^{(k)}$ as $\beta^{(k)} = \sum_{j=1}^{N} \phi(x^{(j)}) \lambda_j$

Start $\beta^{(0)} = \vec{0}$

$$\beta^{(k+1)} = \beta^{(k)} + \eta/N \sum_{i=1}^{N} \left(t^{(i)} - \beta^T \tilde{\phi}(x^{(i)}) \right) \cdot \tilde{\phi}^{(i)}$$

$$= \sum_{j=1}^{N} \phi(x^{(j)}) \lambda_j + \frac{\eta}{N} \sum_{i=1}^{N} \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j \phi(x^{(j)})^T \phi(x^{(i)}) \right) \tilde{\phi}^{(i)}$$

$$= \sum_{j=1}^{N} \phi(x^{(j)}) \left(\lambda_j + \frac{\eta}{N} \left(t^{(j)} - \sum_{i=1}^{N} \lambda_i \phi(x^{(i)})^T \phi(x^{(j)}) \right) \right)$$

$$\lambda_j^{(k+1)} = \lambda_j^{(k)} + \frac{\eta}{N} \left(t^{(j)} - \sum_{i=1}^{N} \lambda_i \phi(x^{(i)})^T \phi(x^{(j)}) \right)$$

Kernels

$y(x) = \beta^T \phi(x)$ $\longleftarrow$ $\sum \lambda_i^{(k)} \phi(x^{(i)})$

Let $\phi(x^{(i)})$ to denote a feature vector (e.g polynomial features)

$$\beta^{(k+1)} \leftarrow \beta^{(k)} + \eta \sum_{i=1}^{N} (t^{(i)} - \beta^T \phi(x^{(i)})) \cdot \phi(x^{(i)})$$

$\phi(x^{(i)})$ depends on D (dimension) $\rightarrow$ can be expensive to store when D is large

How about storing β as $\sum_{i=1}^{N} \lambda_i \phi(x^{(i)})$ $\quad \sum_{i=1}^{N} \lambda_i^{(0)} \phi(x^{(i)})$

Always possible to use such a decomposition for $\beta^{(0)} = 0$

$$\beta^{(k+1)} = \sum_{i=1}^{N} \lambda_i \phi(x^{(i)}) + \eta \sum_{i=1}^{N} \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j \phi(x^{(j)})^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$= \sum_{i=1}^{N} \phi(x^{(i)}) \Big(\underbrace{\lambda_i + \eta \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j \phi(x^{(j)})^T \phi(x^{(i)})\right)}_{\lambda^{(k+1)}} \Big)$$

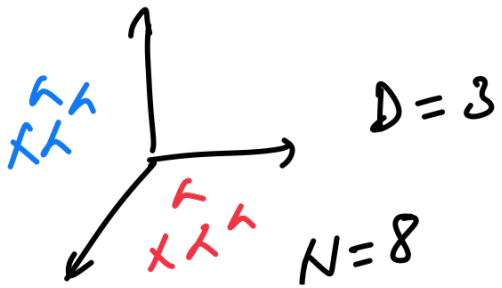
$$\beta^{(k)} = \sum_{i=1}^{N} \lambda_i^{(k)} \phi(x^{(i)}) \rightarrow$$

$$\beta^{(k+1)} = \sum_{i=1}^{N} \lambda_i^{(k+1)} \phi(x^{(i)})$$

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} + \eta \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j^{(k)} \underbrace{\phi(x^{(j)})^T \phi(x^{(i)})}_{\uparrow}\right) \quad (*)$$

General idea: iterate over the $\lambda_j^{(k)}$ instead $\beta_j^{(k)}$

→ storage goes from $O(D)$ to N



$$K(x^{(i)}, x^{(j)}) = K(i,j) = \phi(x^{(i)})^T \phi(x^{(j)})$$

→ K is of size N^2

Why is it interesting to have $\langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle$

$$= \phi(x^{(i)})^T \phi(x^{(j)})$$

Consider polynomial features of max degree 3

$$\phi(x) = [1, x_1, x_2, x_3, \ldots \; x_1^2 \; x_2^2 \; \ldots \; x_1^3 \; x_2^3 \ldots \; x_D^3]$$

$$\phi(z) = [1, z_1, z_2, z_3 \ldots \; z_1^2 \; z_2^2 \; \ldots \; z_1^3 \; z_2^3 \ldots \; z_D^3]$$

$$\phi(x)^T \phi(z) = 1 + \sum x_i z_i + \sum x_i x_j z_j z_j + \sum x_i x_j x_k z_i z_j z_k$$

$$= 1 + (x^T z) + (x^T z)^2 + (x^T z)^3$$

$$\lambda_i^{(k+1)} \leftarrow \lambda_i^{(k)} + \eta \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j^{(k)} \phi(x^{(i)})^T \phi(x^{(j)}) \right)$$

$$\leftarrow \lambda_i^{(k)} + \eta \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j^{(k)} K(i,j) \right)$$

If we want to focus on K instead of ϕ

What are the "good" K matrices that we could use to learn the properties of our data!

$K \rightarrow$ Square Symmetric

$\rightarrow$ positive semidefinite (psd)

To see why K has to be psd

take $z \in \mathbb{R}^N$

if $\exists\ \phi(x^{(i)}), \phi(x^{(j)})$ s.t. $K(i,j) = \phi(x^{(i)})^T \phi(x^{(j)})$

then

$$z^T K z = \sum_{i,j} z_i k_{ij} z_j$$

$$= \sum_{i,j} z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j$$

$$= \sum_{i,j,k} z_i \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j$$

$$= \sum_k \left(\sum_i z_i \phi_k(x^{(i)}) \right) \left(\sum_j z_j \phi_k(x^{(j)}) \right)$$

$$= \sum_k \left(\sum_i z_i \phi_k(x^{(i)}) \right)^2$$

$\Rightarrow z^T K z \geq 0 \iff K$ positive semidefinite matrix

---

$\Rightarrow$ Theorem (Mercer) For K to be a valid kernel (Mercer kernel) it is necessary and sufficient to have K symmetric and positive semidefinite

$$K(i,j) = \frac{1}{\sigma} \exp\left(\frac{-\|x^{(i)} - x^{(j)}\|^2}{\sigma^2}\right)$$

λ_i: we get after our iterations

How do we recover $y(x)$ from λ_i^* ?

$$y(x) = \beta^T \phi(x) = \left(\sum_{i=1}^{N} \lambda_i^* \phi(x^{(i)})\right)^T \phi(x)$$

$$= \sum_{i=1}^{N} \lambda_i^* \phi(x^{(i)})^T \phi(x)$$

$$= \lambda^T \underbrace{K(x^{(i)}, x)}$$

as an example if $K(i,j)$ can be generate from a function $K(x,y)$ of the original data,

We can get the expression of our model $y(x)$ directly from that function $K(x, y)$

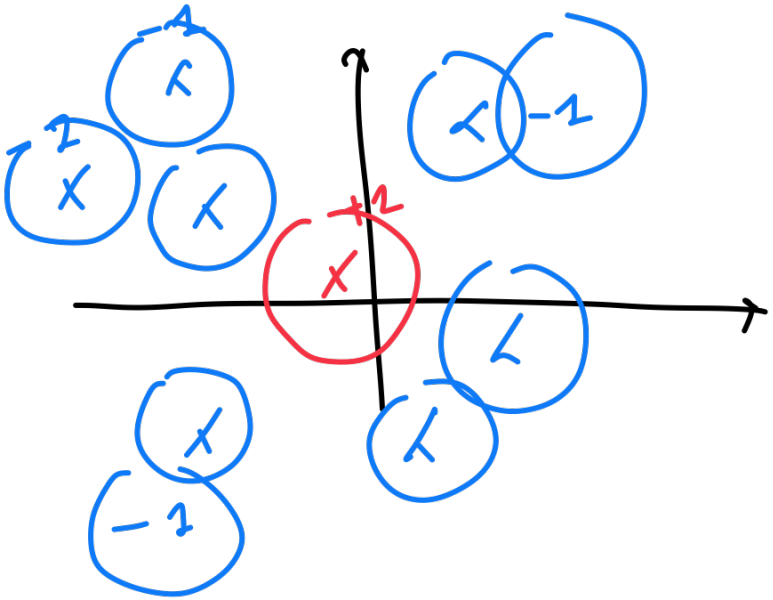
e.g take $K(\underline{x}, \underline{y}) = \frac{1}{\sigma} \exp\left(-\frac{\|\bar{x} - \bar{y}\|^2}{\sigma^2}\right)$

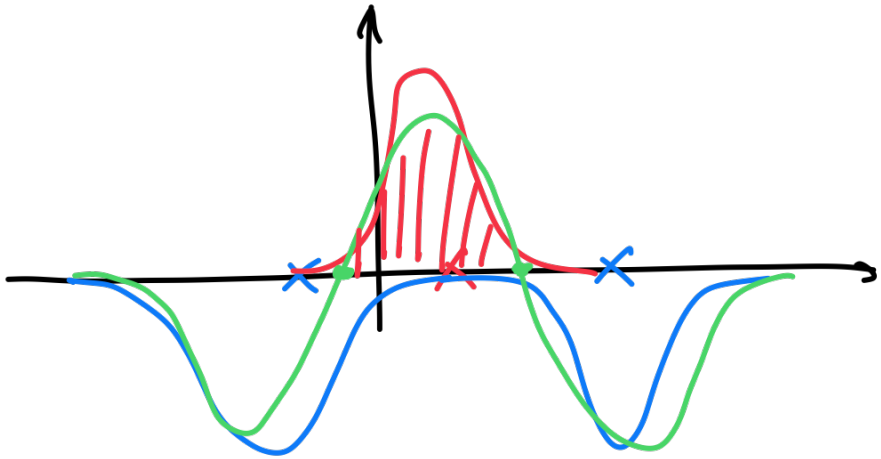
To learn $y(x)$ we first need the N^2

$K(i,j) = \frac{1}{\sigma} \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{\sigma^2}\right)$ for every $x^{(i)}, x^{(j)}$

then learn the $\lambda_i^{(k)}$

Finally you get your model as $y(x) = \sum_{i=1}^{N} \lambda_i^* \exp\left(-\frac{\|x^{(i)} - x\|^2}{\sigma^2}\right)$


-1
-1
+1
-1
-1
-1
-1



Kernels

General idea → Switch from an approach based on feature vectors (hence on the dimension D) to an approach depending on the number of training points N

$$\rightarrow \quad L(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \beta^T \phi(x^{(i)}) \right)^2$$

$$\beta^{(k+1)} \leftarrow \beta^{(k)} + \eta \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \beta^T \phi(x^{(i)}) \right) \phi(x^{(i)})$$

general idea $\beta^{(k)} = \sum_i \lambda_i^{(k)} \phi(x^{(i)})$

$$\beta^{(k+1)} \leftarrow \sum_{i=1}^{N} \lambda_i^{(k)} \phi(x^{(i)}) + \frac{\eta}{N} \sum_{i=1}^{N} \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j^{(k)} \phi(x^{(j)})^T \phi(x^{(i)})\right) \phi(x^{(i)})$$

$$\leftarrow \sum_{i=1}^{N} \left(\lambda_i^{(k)} + \frac{\eta}{N} \sum_{i=1}^{N} \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j^{(k)} \phi(x^{(j)})^T \phi(x^{(i)})\right)\right) \phi(x^{(i)})$$

$$\lambda_i^{(k+1)} \leftarrow \lambda_i^{(k)} + \frac{\eta}{N} \left(t^{(i)} - \sum_{j=1}^{N} \lambda_j^{(k)} \underbrace{\phi(x^{(j)})^T \phi(x^{(i)})}_{K(i,j)}\right) \lambda_j$$

We then introduce $K(x^{(i)}, x^{(j)})$ to store the scalar products between the feature vectors.

We can then study the class of K matrices for which there exist a decomposition in feature vectors

$\longrightarrow$ Mercer's Kernel : Positive semi definite K
(Symmetric)

e.g $K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma}\right)$

Claim $K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma}\right)$ is a valid kernel (meaning $\exists\ \phi(x^{(i)})\ \phi(x^{(j)})$

Such that

$$K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$$

Proof

$$\exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{\sigma}\right) = e^{-\frac{1}{\sigma}\|x^{(i)}\|^2} e^{-\frac{1}{\sigma}\|x^{(j)}\|^2} e^{\frac{2}{\sigma} x^{(i)T} x^{(j)}}$$

$$= e^{-\frac{1}{\sigma}\|x^{(i)}\|^2} e^{-\frac{1}{\sigma}\|x^{(j)}\|^2}$$

Taylor

$$\left(1 + \frac{2}{\sigma}\sum_{k=1}^{n} x_k^{(i)} x_k^{(j)} + \left(\frac{2}{\sigma}\right)^2 \frac{\left(\sum_k x_k^{(i)} x_k^{(j)}\right)^2}{2!} + \frac{1}{3!}\left(\frac{2}{\sigma}\right)^3 \left(\sum_k x_k^{(i)} x_k^{(j)}\right)^3 + \ldots\right.$$

$$\exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{\sigma}\right) = e^{-\frac{1}{\sigma}\|x^{(i)}\|^2} e^{-\frac{1}{\sigma}\|x^{(j)}\|^2}$$

$$\bigcirc = \phi(x^{(i)})^T \phi(x^{(j)})$$

$$* \left(1, \sqrt{\frac{2}{\sigma}}\, x^{(i)}, \sqrt{\left(\frac{2}{\sigma}\right)^2 \frac{1}{2!}}\, (x_k^{(i)} x_l^{(i)})_{k,l}, \sqrt{\left(\frac{2}{\sigma}\right)^3 \frac{1}{3!}}\, (x_k^{(i)} x_l^{(i)} x_m^{(i)})_{k,l,m,\ldots}\right)^T$$

$$\left(1, \sqrt{\frac{2}{\sigma}}\, x^{(j)}, \sqrt{\left(\frac{2}{\sigma}\right)^3 \frac{1}{2!}}\, (x_k^{(j)} x_l^{(j)})_{k,l}, \ldots\right)$$

$$\left(\sum_{k=1}^{D} x_k^{(i)} x_k^{(j)}\right)^2$$

$$= \sum_{k=1}^{D} \sum_{l=1}^{D} x_k^{(i)} x_l^{(i)} x_k^{(j)} x_l^{(j)} =$$

Formal statement of Mercer:

$\mathcal{X}$ measure space, we will $k: L^2(\mathcal{X}\times\mathcal{X}) \to \mathbb{R}$ is a valid kernel iff there exist some feature map $\varphi: \mathcal{X} \to H$ H separable Hilbert space

$$k(x,x') = \langle \varphi(x), \varphi(x') \rangle_H$$

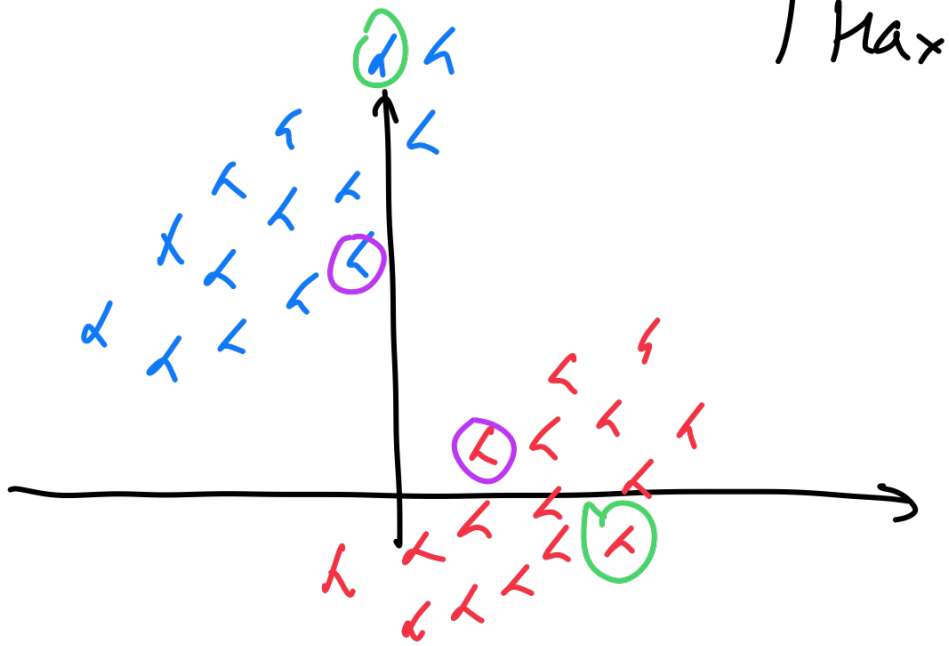
Given a kernel k we say that k satisfies Mercer's condition if

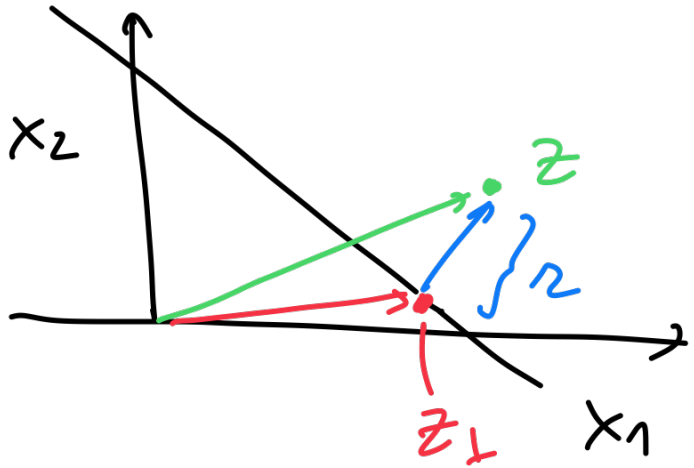
$$\int_{\mathcal{X}\times\mathcal{X}} k(x,x') f(x) f(x') \, dx \, dx' \geq 0 \quad \text{for all } f \in L^2(\mathcal{X})$$

Mercer theorem

k valid kernel iff k is Mercer

Support Vector Machines / Kernel Vector Machines
/ Sparse Vector Machines
/ Maximum Margin classifier

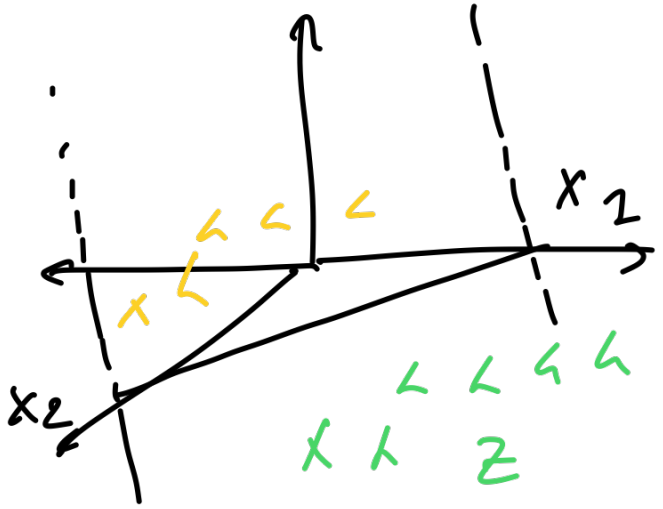




$$z = z_\perp + \frac{(\beta_1, \beta_2)}{\|(\beta_1, \beta_2)\|} \cdot r \quad !$$

plane $\beta_1 x_1 + \beta_2 x_2 + \beta_0 = 0$

normal vector $= (\beta_1, \beta_2)$



$$x_3 \equiv \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

$$\ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \right)^2$$

$$x_2 = -\frac{\beta_1}{\beta_2} x_1 - \frac{\beta_0}{\beta_2}$$

$$z = z_1 + \frac{(\beta_1, \beta_2)}{\|(\beta_1, \beta_2)\|} \cdot r = (z_1)_2, (z_1)_2) + \frac{(\beta_1, \beta_2)}{\|(\beta_1, \beta_2)\|} r$$

$\| \\ (z_1, z_2)$

$$\beta_0 + \beta_1 (z_1)_1 + \beta_2 (z_1)_2 = 0$$

$$\underbrace{\beta_0 + \beta_1 z_1 + \beta_2 z_2}_{y(z)} = \overbrace{\beta_0 + \beta_1 (z_1)_1 + \beta_2 (z_1)_2}^{= 0}$$

$$+ \frac{\beta_1 \beta_1}{\|(\beta_1, \beta_2)\|} r + \frac{\beta_2 \beta_2}{\|(\beta_1, \beta_2)\|} r$$

$$= r \frac{\|(\beta_1, \beta_2)\|^2}{\|(\beta_1, \beta_2)\|} = r(\|\beta_1, \beta_2\|)$$

For any point z, the distance r of z to the plane is given by

$$r = \frac{y(z)}{\|(\beta_1, \beta_2)\|}$$

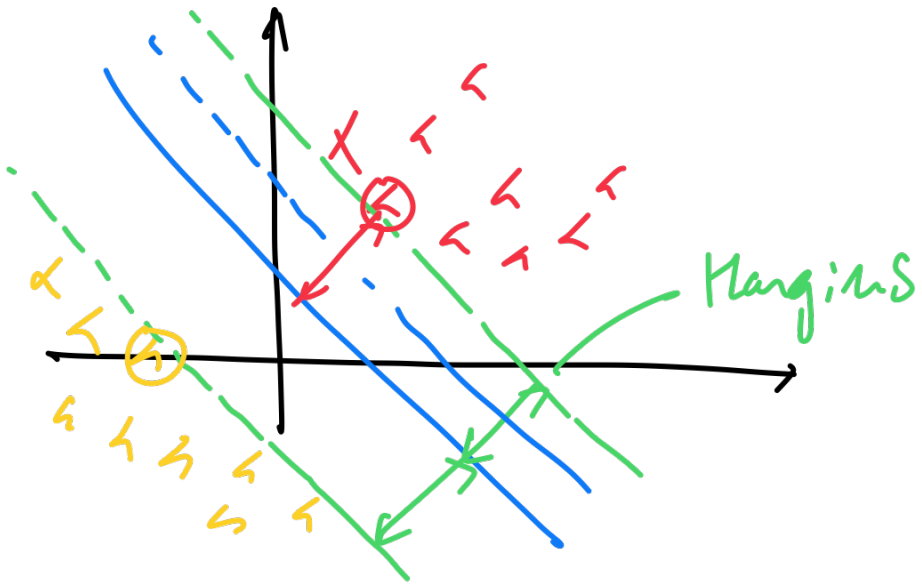
From r we can get $|r|$ by multiplying by $t^{(i)}$ (target of $z^{(i)}$) provided $t^{(i)} = +1$ if $z^{(i)}$ is above -1 if $z^{(i)}$ is below

$$\text{dist} = r\, t(z) = \frac{y(z) \cdot t(z)}{\|(\beta_1, \beta_2)\|}$$

We can then look for the plane that maximizes its distance with respect to the closest $z^{(i)}$

$$\max_{(\beta_0,\beta_1,\beta_2)} \min_{z^{(i)}} \frac{t(z^{(i)})\, y(z^{(i)})}{\|(\beta_1,\beta_2)\|} = \frac{t(z^{(i)})\,(\beta_0+\beta_1 z_1^{(i)}+\beta_2 z_2^{(i)})}{\|(\beta_1,\beta_2)\|}$$

$t(z^{(i)})\, y(z^{(i)}) = \pm 1$



ratio is invariant by rescaling of β

$\beta \leftarrow \alpha\beta$

$$\frac{t(z^{(i)})(\cancel{\alpha}\beta_0 + \cancel{\beta_1} z_1^{(i)} + \cancel{\beta_2} z_2^{(i)})}{\cancel{\alpha}\,\|(\beta_1, \beta_2)\|} \rightarrow$$

→ let us choose β such that for the closest $z^{(i)}$ we have

$$t(z^{(i)})(\beta_0 + \beta_1 z_1^{(i)} + \beta_2 z_2^{(i)}) = 1$$

⇒ For the closest point the distance is $\frac{1}{\|(\beta_1, \beta_2)\|}$

⇒ For all the other points, necessarily we have

$$\text{distance} \geq \frac{1}{\|(\beta_1, \beta_2)\|}$$

For all the other points $t^{(j)}(\beta_0 + \beta_1 z_1^{(j)} + \beta_2 z_2^{(j)}) \geq 1$

the problem then turns into

$$\max_{(\beta_0, \beta_1, \beta_2)} \frac{1}{\|(\beta_1, \beta_2)\|}$$

under the constraints

$$t^{(i)}(\beta_0 + \beta_1 z_1^{(i)} + \beta_2 z_2^{(i)}) \geq 1$$

for every i

We can further simplify the formulation into

$$\min_{\beta} \; \|(\beta_1, \beta_2)\|^2$$

$$t^{(i)}\left(\beta_0 + \beta_1 z_1^{(i)} + \beta_2 z_2^{(i)}\right) \geq 1$$