SVM

1. THE DUAL

- **Primal**

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

data points ↓

constraints $y_i(w^T x_i - b) \geq 1 \quad \forall i = 1, \dots, N$

- **Lagrangian**
  - ↳ α_i are lagrangian multipliers of each data point

↳ $$\mathcal{L}(w,b,\alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i \left((w^T x_i - b) - 1 \right)$$

- **Dual**

$$\max_{\alpha_i \geq 0} \left[\min_{w,b} \mathcal{L}(w,b,\alpha) \right]$$

↑ fixed

• Dual

$$\max_{\alpha_i \geq 0} \left[\min_{w,b} \mathcal{L}(w, b, \alpha) \right]$$

↳ it means solving the min problem first to find w, b with a fixed α_i (fixed) then find an α that maximize everything ? Not sure

↳ then on stationarity constraints with partial derivatives of w and $b = 0$

$$1 \rightarrow \frac{\partial f}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

⇝ But what about α_i?

The only vectors contributing to the margin are support vectors and they are the only vectors that can contribute to the definition of w.

This means that for NON-supporting vectors x_i $\alpha_i = 0$. if not they would contribute to w and that's not true.

$$2 \rightarrow \frac{\partial f}{\partial b} = 0 \Rightarrow 0 = \sum_{i=1}^{N} \alpha_i y_i$$

$$0 = \alpha^T y$$ ← vector dot-product formulation of the previous sum

- We want to define the dual problem only in terms of $\left[\max_{\alpha_i \geq 0}\right]$ so we substitute w and b from stationary constraints to get:

$$\max_{\alpha_i \geq 0}\left[\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j\right]$$

After getting α, we can obtain w and b using stationary constraints

This is simple because NON-SVs have $\alpha_i = 0$ and it means we need to sum only SV because all the others are zero in the sum.

The simplified formula is (change in the sign to get a MIN problem)

$$\min_{\alpha_i \geq 0}\left[\frac{1}{2}\sum_{i}^{\text{SVs}}\sum_{j}^{\text{SVs}} \alpha_i \alpha y_i y x_i^T x_j - \sum_{i}^{\text{SVs}} \alpha_i\right]$$

why D-o-l ?

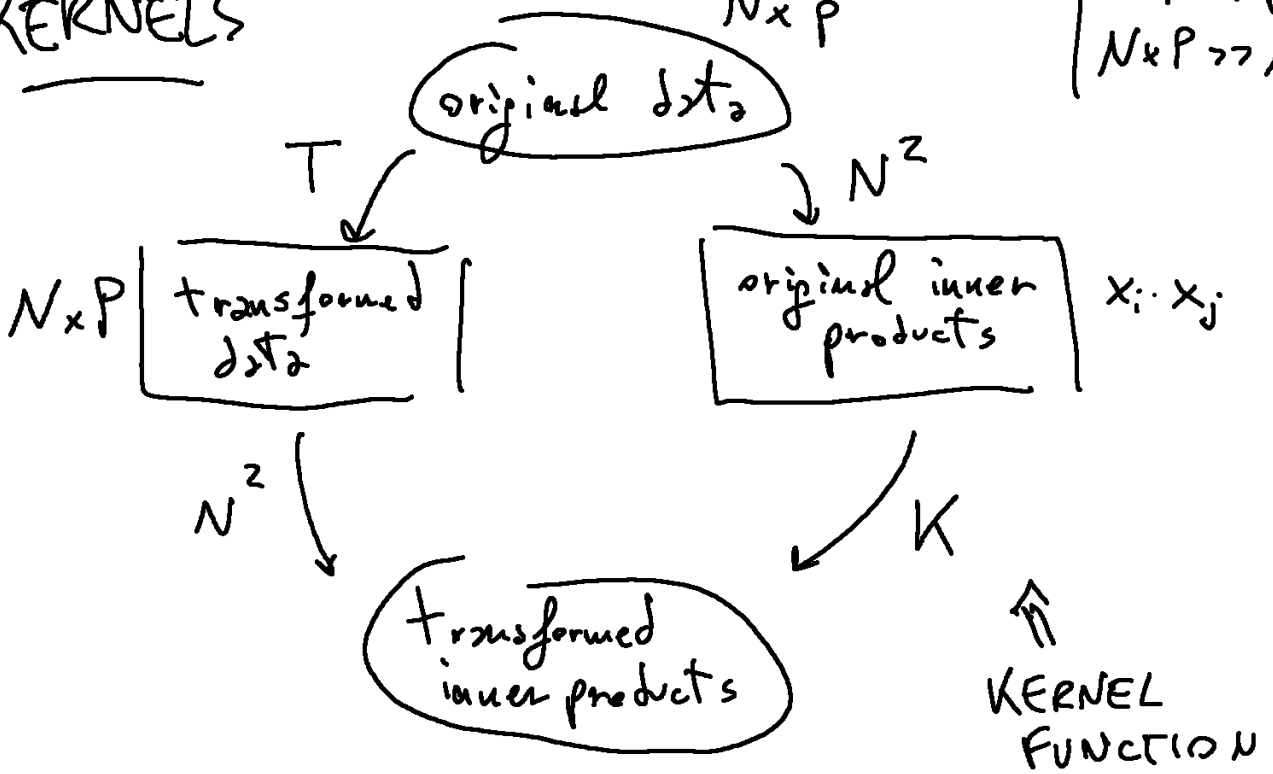
1. useful for KERNELS (see lesson)

2. with high-dimensional data,
   when $p >> N$ (attributes >> record)
   for example with images,
   in the primal we have to use the whole dataset
   so $N \cdot p$ operations, while in the dual
   we need to compute only $\langle x_i^T \cdot x \rangle$ which is
   only N^2.
   Then if $p >> N$ and the dual is better
   $Np >> N^2$

(?) I didn't get why we need Np in the primal

KERNELS

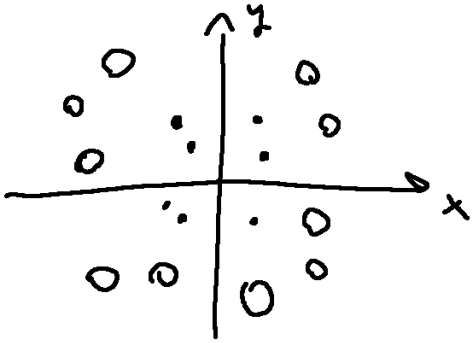
$P \gg p$
$N \times P \gg N \times p$



- What is the benefit of Kernel functions?
  We never have to send data to higher dimensional space.
- In the context of SVM, a Kernel function is any function that only uses the inner products of the original data and it's able to use those inner products and send them into transformed inner products without visiting the higher dimensional space.

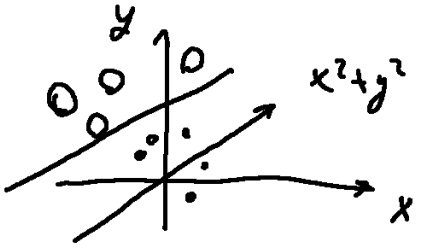
- Why do we need to transform data into higher dimensional data?

↳ Because if we had points like this,



we cannot use an hyperplane to separate the dataset and therefore we need to engineer another dimension.

In this case, it will be x^2+y^2, (which in two-dimension is a circle).



But this means transforming

$$x_i \longmapsto T(x_i)$$

- Why do we only need inner products?

↳ Because we are solving the Dual problem of SVM

PRIMAL $\longleftrightarrow$ DUAL

$$\begin{cases} \min\limits_{w,b} \ \frac{1}{2}\|w\|^2 \\ y_i(w^T x_i - b) \geq 1 \end{cases}$$

$$\begin{cases} \min\limits_{\alpha_i \geq 0} \ \sum_i \sum_j \alpha_i \alpha_j y_i y_j \boxed{x_i^T x_j} - \sum \alpha_i \\ \alpha^T y = 0 \leftarrow ? \end{cases}$$

$$x_i \cdot x_j \longmapsto T(x_i \cdot x_j)$$

or with Kernel Functions

$$\longmapsto K(x_i \cdot x_j)$$