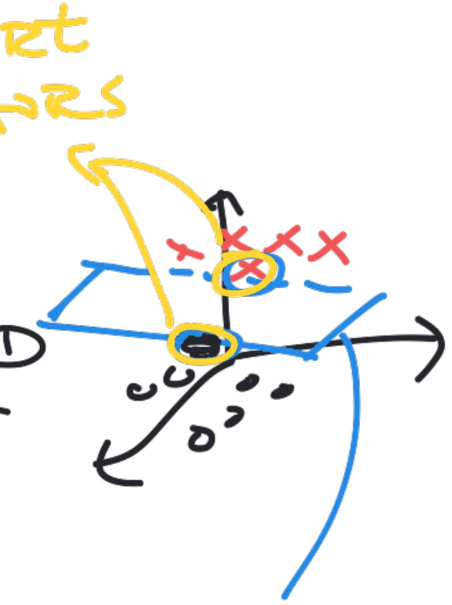
SVM

- Kernel Machine
- Maximizes Margin

$\phi : \mathbb{R}^d \to \mathbb{R}^D$

$D \gg d$



$y = w^T \phi(x) + b$

$$\arg\max_{w,b} \left\{ \frac{1}{\|w\|} \min_n \left[t_n \left(w^T \phi(x_n) + b \right) \right] \right\}$$

- Support vectors satisfy: $\boxed{t_n(w^T \phi(x_n) + b) = 1}$, then

$$\boxed{\arg\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{sub. to } \underline{t_n(w^T \phi(x_n) + b) \geq 1}}$$

- PRIMAL Lagrangian:

$$\mathcal{L}(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left[t_n (w^T \phi(x_n) + b) - 1\right]$$

$\phi(x_n)$ can be infinite dimensional!

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \end{cases} \Leftrightarrow \begin{cases} w = \sum_{n=1}^{N} a_n t_n \phi(x_n) \\ 0 = \sum_{n=1}^{N} a_n t_n \end{cases}$$

- DUAL Lagrangian: (plugging in w in $\mathcal{L}(w,b,a)$)

$$\tilde{\mathcal{L}}(a) = \sum_{n=1}^{N} a_n - \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \underbrace{k(x_n, x_m)}_{=\phi(x_n)^T\phi(x_m)}$$

GRAM MATRIX

such that $a_n \geq 0,\ n = 1, 2, \dots, N$

$$\sum_{n=1}^{N} a_n t_n = 0$$

The solution for $\tilde{\mathcal{L}}(a)$ is the same as the solution for $\mathcal{L}$.

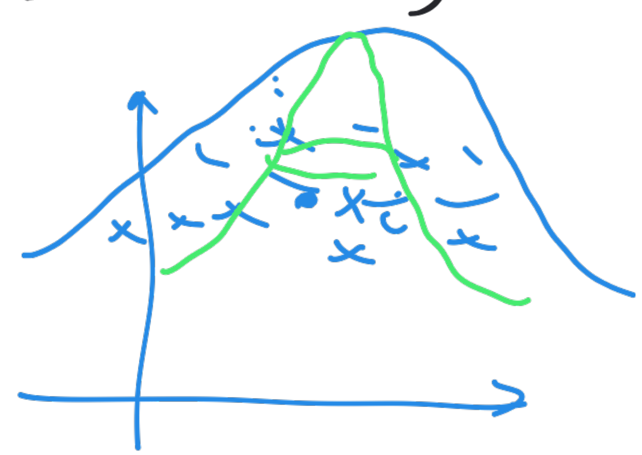
So in the dual representation, we need to compute the Gram Matrix

Gram Matrix

$$K = \begin{bmatrix} \phi^T(x_1)\phi(x_1) & \phi^T(x_1)\phi(x_2) & \cdots & \phi^T(x_1)\phi(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi^T(x_n)\phi(x_1) & \phi^T(x_n)\phi(x_2) & \cdots & \phi^T(x_n)\phi(x_n) \end{bmatrix}$$

$$\phi^T(x_i)\phi(x_j) = K(x_i, x_j) = \exp\left(-\underline{\underline{\gamma}}\, \|x_i - x_j\|^2\right)$$

e.g. we can use RBF



If γ is small $\rightarrow$ variance large

γ is large $\rightarrow$ variance small

So, γ, in the kernel function controls the neighborhood of each data sample.

→ It needs to be selected using cross-validation.

In dual rep., we use quadratic dynamic programming to solve for $\underline{\underline{a}}$.

$$y(x) = \underline{w}^T \phi(x) + b$$

using the solution for w:

$$\boxed{y(x) = \sum_{n=1}^{N} a_n t_n \boxed{K(x, x_n)} + b}$$

In training, we can define the GRAM MATRIX K for all samples

$$y(x) = \sum_{n=1}^{N} a_n t_n \underbrace{K(x, x_n)} + \underline{\underline{b}}$$

If $x \equiv x_T$, then we compute its similarity (in the kernel fct. sense) as:

$$\left[\phi^T(x_T)\phi(x_1), \phi^T(x_T)\phi(x_2), \dots, \phi^T(x_T).\phi(x_N)\right]^T$$

Linear Kernel: $K(x, y) = x^T y$

Case overlapping classes:

$$\xi_n = |t_n - y(x_n)|$$

$\xi_n = 0$: x_n is correctly classified and far away from margin

$0 < \xi_n \leq 1$: x_n lies inside the (correct side) margin

$\xi_n > 1$: x_n lies on wrong side of decision boundary

$$\boxed{\begin{aligned} &\arg\min_{w,b} \frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \xi_n \\ &\text{sub. to } \; t_n y(x_n) \geq 1 - \xi_n \\ &\qquad\quad \xi_n \geq 0, \; n = 1, \dots, N \end{aligned}}$$

$C \to \infty$: recovers the Hard margin w/out misclassification

$C \to 0$: allows for a lot misclassified samples.

(PRIMAL)

① Define Lagrangian:

$$\mathcal{L}(w,b,a) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \underset{\uparrow}{a_n}\left(t_n y(x_n) - 1 + \xi_n\right) - \sum_{n=1}^{N} \mu_n \xi_n$$

Lagrange multiplier for each sample

② Define the KKT conditions.

③ Take derivatives of $\mathcal{L}$ w.r.t. to w, b, ξ_n.

④ Construct the DUAL Lagrangian by plugging in solutions in ③.