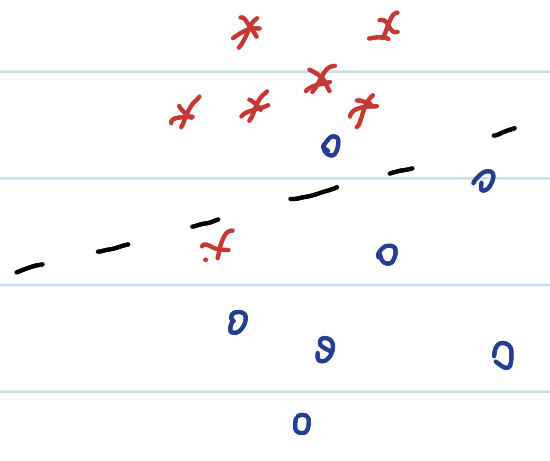
2. Overlapping Class Distributions

In part 1, we assume that training data points are linear separable. But, this assumption is strong and usually inpractical. However, we still want to make use of Maximum Margin Classfier's good properties.

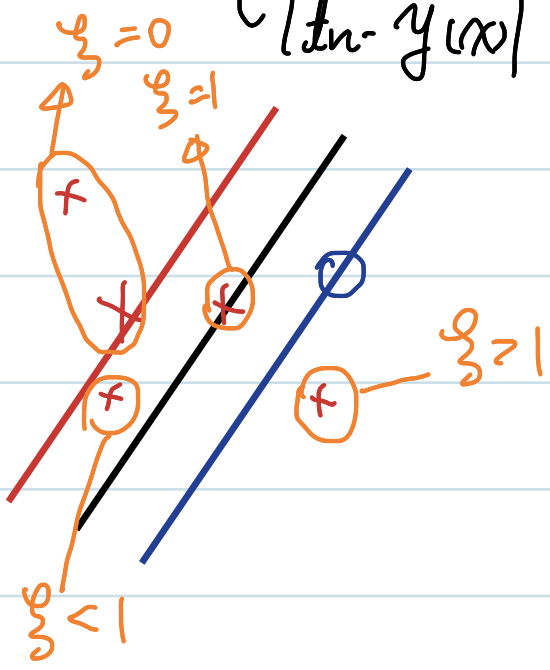
Our target is to enable some mis-classified points. In case below, the dashed line can separate most points properly.



but leave some miss-classified

To deal with that, we introduce slack variables $\xi_n \geq 0$

$\xi_n = \begin{cases} 0 & \text{if point } n \text{ is in the correct margin} \\ |t_n - y(x)| & \text{miss classified.} \end{cases}$ So, we have



$$\begin{cases} \xi = 0 & \text{in the correct margin} \\ \xi \in (0,1) & \text{in the correct decision region} \\ \xi = 1 & \text{on the decision plane} \\ \xi > 1 & \text{miss-classified.} \end{cases}$$

Remember! Even through we will show that slack variable can deal with miss-classified point, but this also make model Sensitive to outliers

Now, we will minimize.

$$\underset{w,b,\xi}{\operatorname{argmin}} \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$

C: a constant control the sensitive to miss-classified points.

$$\text{S.t.} \quad t_n y_n(x) \geq 1-\xi_n \quad \forall n$$
$$\xi_n \geq 0 \quad \forall n$$

If $C \to \infty$, we will turn to separable data's case.

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

Also use the Lagrangian dual method.

$$\mathcal{L}(w,b,\xi,\lambda,\mu) = \frac{1}{2}\|w\|_2^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\lambda_n\{t_n y(x_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

minimize it over w, b, ξ.

$$\frac{\partial L}{\partial w} = w - \sum_{n=1}^{N}\lambda_n t_n \phi(x_n) = 0 \Rightarrow w^* = \sum_{n=1}^{N}\lambda_n t_n \phi(x_n)$$

$$\frac{\partial L}{\partial b} = \sum_{n=1}^{N}\lambda_n t_n = 0$$

$$\frac{\partial L}{\partial \xi_n} = C - \lambda_n - \mu_n = 0 \Rightarrow \quad \mu_n = C - \lambda_n$$

Now, we have

$$g(\lambda) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n \lambda_m t_n t_m \phi(x_n)^T \phi(x_m)$$

$$- \sum_{n=1}^{N} \lambda_n \left[t_n \left(\sum_{m=1}^{N} \lambda_m t_m \phi(x_m) \right)^T \phi(x_n) - 1 \right]$$

$$= \sum_{n=1}^{N} \lambda_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n \lambda_m t_n t_m k(x_n, x_m)$$

The lagrangian turns to be the same as separable case, but we have different constraints.

$\because \lambda_n \geq 0$, $\mu_n \geq 0$ and $\lambda_n = C - \mu_n$

we have $0 \leq \lambda_n \leq C$

for optimal b, we have $\sum_{n=1}^{N} \lambda_n t_n = 0$

Fortunately, solving the dual problem is also not difficult.