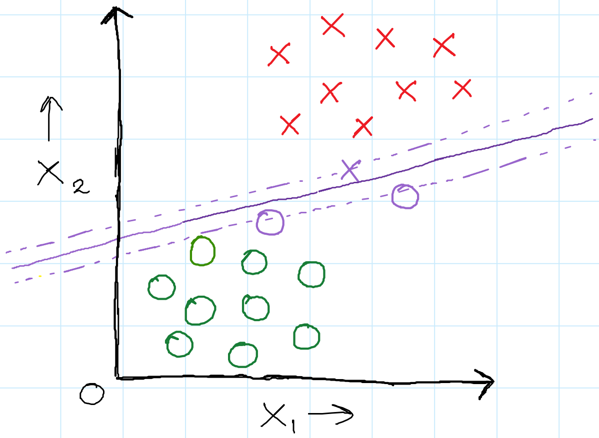
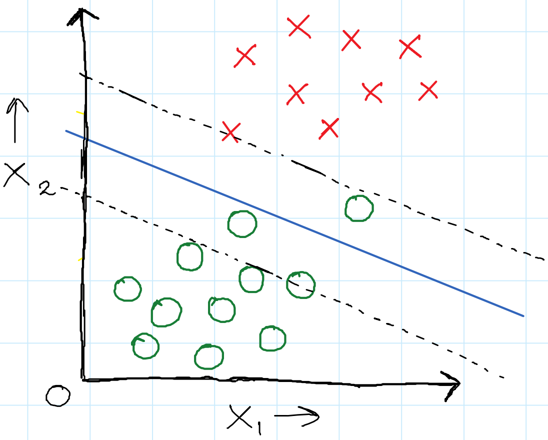
Soft Margin Classification

For soft margin classifier, we need to allow some missclassifications so that we have a tradeff between the margin width and the 'correctness' of our classifier. This indicates that we need to have some 'slack' or 'dheel' (ढील) in classifying all the data points correctly so that our margin width does not suffer because of noise in our dataset.

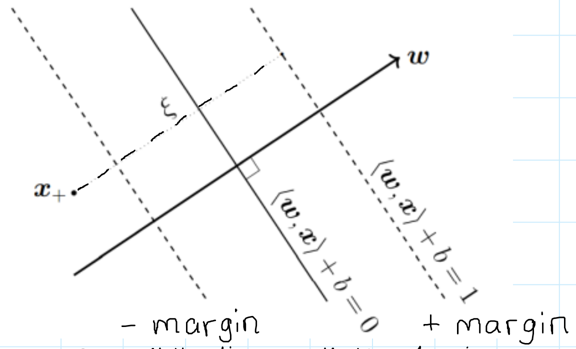
Eg: It is unlikely that any real world data will be linearly separable, even if it is, it might look somewhat like :



A soft margin SVM will/can allow some slackness in the process which will give us a wider margin at the cost of a few misclassifications.



So we aim at minimizing the total amounts of slacks while maximizing the margin width. Let's introduce a slack variable ε_i for every training example $i = 1, \ldots, m$.



Source: Mathematics for Machine Learning

Where ε quantifies the 'error' or deviation of a + sample x_+ from the + margin: $\omega \cdot x + b = +1$. For a SV, $\varepsilon_{sv} > 1$ means that it has been misclassified as it lies on the other side of the hyperplane. If $\varepsilon_{sv} < 1$, then it lies inside that margin yet correctly classified.

OPTIMIZATION:

$$\min_{\omega, b} \frac{1}{2}\|\omega\|^2 + C \sum_{i=1}^{m} \varepsilon_i$$

subject to ① $y_i(\omega \cdot x_i + b) \geq (1 - \underbrace{\varepsilon_i}_{\text{The amount of slack for the } i^{th} \text{ sample.}})$ $\quad \forall\ i = 1, 2, \ldots, m$

② $\varepsilon_i \geq 0 \quad \forall\ i = 1, 2, \ldots, m$

If C is large, then even a very little slack will increase the cost. If C is kept small, then relatively more slack can be allowed.

Thus, reducing C helps in increasing the no. of SVs inside and on the margin which results in decreasing the variance and reducing over-fitting because our model will generalize better.

New Primal Problem:

New Primal Problem :

$$\mathcal{L}(\omega, b, \varepsilon, \alpha, \mu) = \frac{1}{2}\omega.\omega + C\sum_{i=1}^{m}\varepsilon_i - \sum_{i=1}^{m}\alpha_i\left[y_i(\omega.x_i+b)-(1-\varepsilon_i)\right] - \sum_{i=1}^{m}\mu_i\varepsilon_i$$

$$\alpha_i \geq 0 \qquad \forall\ i = 1, 2, \dots, m$$
$$\mu_i \geq 0$$

$$\min_{\omega, b, \varepsilon} \ \max_{\alpha, \mu} \ \mathcal{L}(\omega, b, \varepsilon, \alpha, \mu)$$

NOTE : $y_i(\omega.x_i + b) \geq 1 - \varepsilon_i$

if $\varepsilon_i = 0$, then this i^{th} sample is correctly classified.

if $y_i(\omega.x_i + b) = 1 - 0$ then it's a correctly classified S.V.

else if $\varepsilon_i > 0$, then this sample might be misclassified.

if $y_i(\omega.x_i + b) = 1 - \varepsilon_i$ and $\varepsilon_i > 1$ then this '<u>is</u>' a misclassified data point.

else if $y_i(\omega.x_i + b) \geq 1 - \varepsilon_i$ is smaller than 1, then it's misclassified.

Partially differentiating the Primal Lagrangian wrt the Primal Variables ω, b and ε gives us :

$$\frac{\partial \mathcal{L}}{\partial \omega} = \omega - \sum_{i=1}^{m}\alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{m}\alpha_i y_i$$

$$\frac{\partial \mathcal{L}}{\partial \varepsilon_i} = C - \alpha_i - \mu_i$$

Equating them to zero gives :

$$\boxed{\omega = \sum_{i=1}^{m}\alpha_i y_i x_i} \quad - \ 1$$

$$\boxed{\omega = \sum_{i=1}^{m} \alpha_i y_i x_i} \quad - \ (1)$$

$$\boxed{\sum_{i=1}^{m} \alpha_i y_i = 0} \quad - \ (2)$$

$$\boxed{C - \alpha_i - \mu_i = 0} \quad - \ (3)$$

Substituting (1), (2) and (3) after simplifying our Primal Problem :

$$\mathcal{L}_P : \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i y_i \left(\left\{\sum_{j=1}^{m} \alpha_j y_j x_j\right\} \cdot x_i\right) - b \underbrace{\cancel{\sum_{i=1}^{m} \alpha_i y_i}}_{0}$$

$$+ \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \varepsilon_i - \sum_{i=1}^{m} \mu_i \varepsilon_i + C \sum_{i=1}^{m} \varepsilon_i$$

$$\mathcal{L}_D : -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{m} \alpha_i + \sum_{i=1}^{m} \underbrace{\cancel{(C - \alpha_i - \mu_i)}}_{0} \varepsilon_i$$

$$\boxed{\mathcal{L}_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)}$$

Clearly, $\alpha_i \geq 0$ and $\mu_i \geq 0$ $\forall\ i = 1, 2, \ldots, m$ as they are Langrangian Multipliers.

Given that $\alpha_i = C - \mu_i$ (from (3))

$$\Rightarrow \boxed{0 \leq \alpha_i \leq C}$$

All data points for which $0 < \alpha_i < C$ holds true, are our Support Vectors that lie exactly on our margin. This is explained by Karush - Kuhn Tucker conditions.

Now, solving the Dual Problem shown above, we will get our Lagrangian Multipliers back. This enables us to get back our Primal Variables ω and b which parametrize our separating hyperplane.

parametrize our separating hyperplane.

$$\boxed{w^* = \sum_{i=1}^{m} \alpha_i y_i x_i} \qquad \text{(From 1)}$$

Once we get w, we can plug in that into:

$$w^* \cdot x_{sv} + b^* = y_{sv} \qquad — \text{ 4}$$

for some SV lying exactly on our margin.

$$\Rightarrow \boxed{b^* = y_{sv} - w \cdot x_{sv}}$$

If everything was hunky-dory, we were done at this point.

BUT

It is possible that none of the SVs lie exactly on the margin i.e. all SVs violate the margin. Then we can't use 4 as our assumption fails. By some more math, we can say that in such a case we can find b^* as follows:

$$\boxed{b^* = \text{Median}\left(|y_s - w^* \cdot x_s| \right) \quad \forall \; s \in SVs.}$$