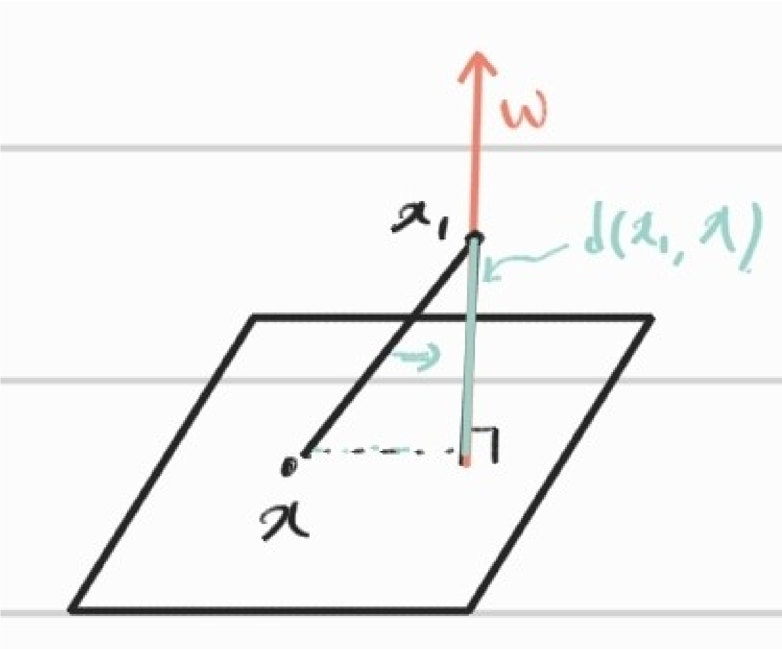
Support Vector Machine

Linear SVM for separable cases

- Normal vector $\mathbf{w}$: vector which is perpendicular to hyperplane.
- Hyperplane : Decision boundary of $y_1 \text{and } y_2$ which can be represented as

$$\begin{aligned} &\mathrm{w}'(x - x_0) = 0 \\ &\rightarrow \mathrm{w}'x + b = 0 \end{aligned}$$

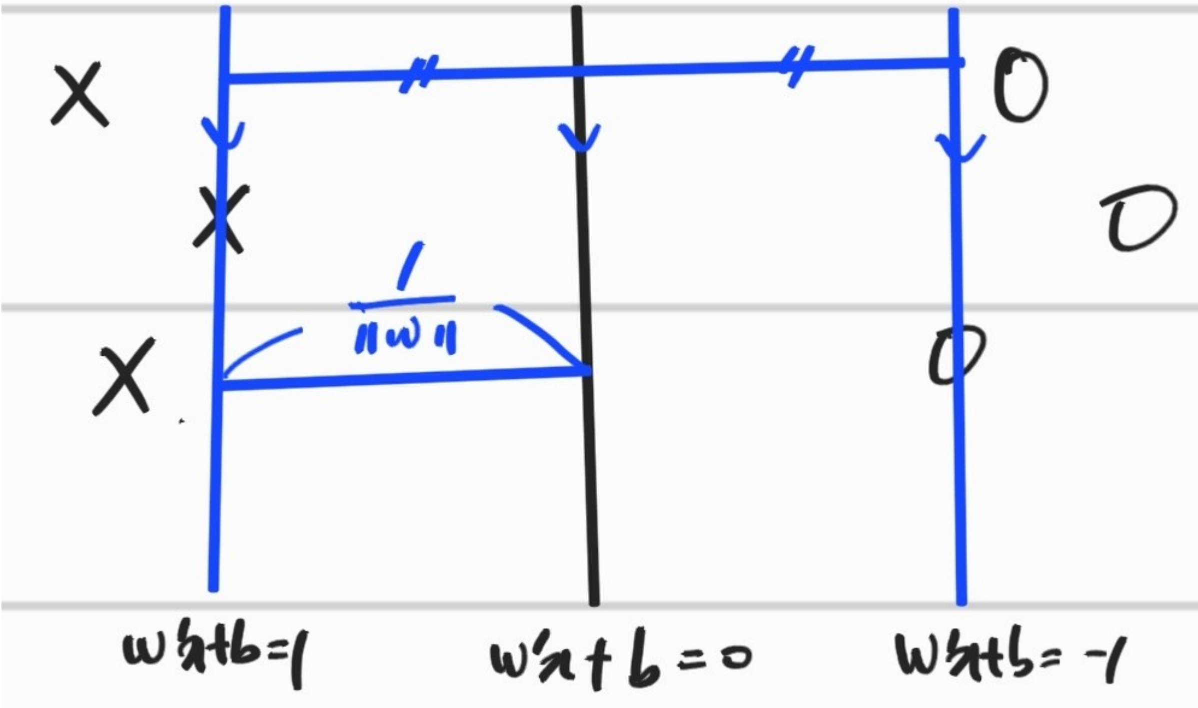
- Distance between a point and hyperplane : norm of distance vector between a pointx_1 and any point x which lies in hyperplane which is projected onto normal vector w



$$\begin{aligned} d(x, x_1) &= \|\mathrm{w}(\mathrm{w}'\mathrm{w})^{-1}\mathrm{w}'(x - x_1)\| \\ &= \frac{\|\mathrm{w}\|}{\|\mathrm{w}\|^2}|\mathrm{w}(x - x_1)| \\ &= \frac{1}{\|\mathrm{w}\|}|\mathrm{w}'x - \mathrm{w}'x_1| \\ &= \frac{|\mathrm{w}'x_1 + b|}{\|\mathrm{w}\|} \end{aligned}$$

- Assumption
  1. Binary classes of response variable can be classified perfectly by one linear decision boundary
  2. Move hyperplane in parallel until the hyperplane touches one side and make the value at that point 1 or -1. Then the real hyperplane have value of 0.

     $\rightarrow$ distance between first touched point and hyper plane becomes $\frac{1}{\|\mathrm{w}\|}$



- Margin : distance between points which is firstly touched by hyperplane moved in parallel
- Objective : maximize margin

$$\max_{\mathrm{w}} \frac{2}{\|\mathrm{w}\|} \text{ subject to } \mathrm{w}'x_i + b \geq 1, \forall i : y_i = 1, \text{ and } \mathrm{w}'x_j + b \leq -1, \forall j : y_j = -1$$

$$\iff \min_{\mathrm{w}} \frac{1}{2}\|\mathrm{w}\|^2 \text{ subject to } y_i(\mathrm{w}'\mathbf{x}_i + b) \geq 1, \forall i$$

$\rightarrow$ **We can get Optimized value by KKT condition!!**

- Dual function(Lagrangian function) of this problem:

$$h(\alpha) = L_p((\mathrm{w}, b), \alpha) = \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{i=1}^{n} \alpha_i \{y_i(\mathbf{x}_i'\mathrm{w} + b) - 1\}, \ \ \alpha_i \geq 0$$

- KKT conditions for this problem

$$\begin{aligned}
&\text{i) } y_i(\mathbf{x}_i'\mathrm{w} + b) \geq 1, \forall i(\text{feasibility}) \\
&\text{ii) } \alpha_i \geq 0, \ \forall i(\text{Lagrange Multiplier}) \\
&\text{iii) } \alpha_i(y_i(\mathbf{x}_i'\mathrm{w} + b) - 1) = 0, \ \forall i(\text{Complementary Slackness}) \\
&\text{iv) } \frac{\partial L_p}{\partial \mathrm{w}} = \mathrm{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0, \ \ \frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0(\text{first derivative have to be 0}) \\
&\iff \mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \ , \ \ \sum_{i=1}^{n} \alpha_i y_i = 0
\end{aligned}$$

- Plug in KKT iv) to dual function

$$\begin{aligned}
h(\alpha) &= -\frac{1}{2}\mathrm{w}'\mathrm{w} + \sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2}\alpha' Y X X' Y \alpha + \mathbf{1}\alpha \\
\text{where } \ Y &= diag\{y_1, \cdots, y_2\}
\end{aligned}$$

- Dual problem becomes :

$$\max_{\alpha \geq 0, y'\alpha = 0} h(\alpha)$$

- Support vectors : The solution of the problem satisfies KKT $\text{iii})$ with non-zero α_i. That is, vectors satisfies $y_i(\mathbf{x}_i'\mathrm{w}+b)=1$

  $\rightarrow$ **Points where constraints are active!!!**

- By solving dual problem, optimal point $\alpha^\star$ can be calculated.

  and let $\mathcal{S}=\{i:\alpha_i>0\}$ (= index set of Support Vector)

  $\rightarrow \mathrm{w}^\star$ can be obtained : $\mathrm{w}^\star=\sum_{i=1}^{n}\alpha_i^\star y_i\mathbf{x}_i=\sum_{i\in\mathcal{S}}\alpha_i^\star y_i\mathbf{x}_i\ (\because \alpha_i=0,\forall i\notin\mathcal{S})$

  $\rightarrow b^\star$ can be obtained : $y_i(\mathbf{x}_i'\mathrm{w}^\star+b)-1=0$

- Theoretically, We can get $\mathrm{w}^\star, b^\star$ with one Support vector, but for numerical stability, the average of all the solutions can be used.

- SVM method uses **Support vectors only. Not the vectors(points) beyond support vectors**

- Strength of SVM

  1. The Optimal Separating Hyperplane is obtained by inner product between $\mathbf{x}$ and $\mathbf{x}_i$ which means it is easy to generalize.

     $\rightarrow$ If y's can not be classified linearly, then we can take inner product in high-dimension feature space and separate y's linearly. By taking them back to original dimension, we can classify response variable non-linearly.

  2. In dual problem, the number of variables (dimension of α) is always the same as the sample size.

     $\rightarrow$ When $p>>n$ case, no matter how large the dimension of $\mathbf{x}$ is, we can obtain OSH easily.

Linear SVM for non-separable cases

- Addition to separable cases, we take care of another variables $\xi_i\geq 0$(slack variables) which means the distance between support vector and points located in opposite direction of the classification.

- Then, the constraints become relaxed.:

$$y_i(\mathrm{w}'\mathbf{x}_i+b)\geq 1-\xi_i,\ \forall i$$

- Penalty to misclassification :

$$\xi_i>1 \iff i\text{th cas is misclassified}$$
$$\rightarrow \text{total \# of misclassified cases} < \sum\xi_i$$

- Optimization problem :

$$\min_{\mathrm{w},b,\xi}\frac{1}{2}\|\mathrm{w}\|^2+C\sum_{i=1}^{n}\xi_i$$
$$\text{subject to }\begin{cases}y_i(\mathrm{w}'\mathbf{x}_i+b)\geq 1-\xi_i,\ \forall i\\ \xi_i\geq 0,\ \forall i\end{cases}$$

- Meaning of Tuning parameter C : Balances the margin and the misclassification error
  - large C : discourage any positive ξ_i, makes the margin small
  - small C : allow positive ξ_i , bigger margin
  - C must be adaptively chosen by data (e.g. Cross-Validation.)
- Dual function

$$L_p((\mathrm{w},b,\xi),(\alpha,\beta))=\min_{\mathrm{w},b,\xi}\frac{1}{2}\|\mathrm{w}\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=0}^{n}\alpha_i\{y_i(\mathrm{w}'\mathbf{x}_i+b)-1+\xi_i\}-\sum_{i=1}^{n}\beta_i\xi_i$$

- KKT conditions

$$\begin{aligned}
&\text{i) } y_i(\mathbf{x}_i'\mathrm{w}+b)-1+\xi_i \geq 0, \forall i(\text{feasibility})\\
&\text{ii) } \alpha_i \geq 0, \beta_i \geq 0,\ \forall i(\text{Lagrange Multiplier})\\
&\text{iii) } \alpha_i(y_i(\mathbf{x}_i'\mathrm{w}+b)-1+\xi_i)=0, \beta_i\xi_i=0\ \forall i(\text{Complementary Slackness})\\
&\text{iv) } \frac{\partial L_p}{\partial \mathrm{w}}=\mathrm{w}-\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i=0,\ \ \frac{\partial L_p}{\partial b}=-\sum_{i=1}^n \alpha_i y_i=0,\ \frac{\partial L_p}{\partial \xi}=C1-\alpha-\beta=0(\text{first derivative}),\\
&\quad \iff \mathrm{w}=\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i\ ,\ \ \sum_{i=1}^n \alpha_i y_i=0, C1=\alpha+\beta
\end{aligned}$$

- Dual functions become

$$h(\alpha)=-\frac{1}{2}\alpha' YXX'Y\alpha+\mathbf{1}\alpha$$
$$\text{which is same as separable cases!!}$$

- By solving Dual functions, we can get $\alpha^\star$
  $\rightarrow$ obtain $w^\star=\sum_{i\in\mathcal{S}}\alpha_i^\star y_i \mathbf{x}_i$
  $\rightarrow$ obtain $\beta_i^\star=C-\alpha_i^\star$
  $\rightarrow$ obtain support vectors where $\alpha_i^\star>0,\ \beta_i^\star>0$ by obtaining $b^\star$:

$$y_i(\mathbf{x}_i'\mathrm{w}+b)-1=0$$
$$(\because \beta_i\xi_i=0\rightarrow \xi_i=0)$$

- As in separable cases, We can get $\mathrm{w}^\star, b^\star$ with one Support vector, but for numerical stability, the average of all the solutions can be used.

SVM as a Penalization Method

- Objective function of SVM is:

$$\min_{\mathrm{w},b,\xi}\frac{1}{2}\|\mathrm{w}\|^2+C\sum_{i=1}^n \xi_i$$
$$\text{subject to } \begin{cases} y_i(\mathrm{w}'\mathbf{x}_i+b)\geq 1-\xi_i,\ \forall i\\ \xi_i\geq 0,\ \forall i\end{cases}$$

- Which is equivalent to

$$\min_{\mathrm{w},b,\xi}\frac{1}{2}\|\mathrm{w}\|^2+C\sum_{i=1}^n \xi_i \text{ subject to } \xi_i\geq\{1-y_i(\mathrm{w}'\mathbf{x}_i+b)\}_+,\ \forall i$$
$$\iff \min_{\mathrm{w},b,\xi}\frac{1}{2}\|\mathrm{w}\|^2+C\sum_{i=1}^n\{1-y_i(\mathrm{w}'\mathbf{x}_i+b)\}_+$$

- Which is same as the problem of minimizing

$$\min_{\mathrm{w},b,\xi}\frac{1}{n}\sum_{i=1}^n\{1-y_i(\mathrm{w}'\mathbf{x}_i+b)\}_+ +\lambda\|\mathrm{w}\|^2$$

  $\rightarrow$ **First term can be viewed as objective(Loss) function and Second term as ridge penalty**
- First term is called hinge loss and this has look of $L(1-yf(x))$
- Some examples of loss functions having look of $1-yf(x)$:
  1. hinge loss : $L(y,f(x))=(1-yf(x))_+$

2. squared error loss : $L(y, f(x)) = (y - f(x))^2 = (1 - yf(x))^2$, when y is coded as ± 1
3. binomial deviance : $L(y, f(x)) = \log[1 + \exp{(-yf(x))}]$,when y is coded as ± 1
4. exponential loss : $L(y, f(x)) = \exp{(-yf(x))}]$