SUPPORT VECTOR MACHINE

→ The objective of the support vector machine is to find a hyper plane in an N dimentional space (N - number of features) that distinctly classifies the data points.

→ Our objective is to find a plane that has the maximum margin

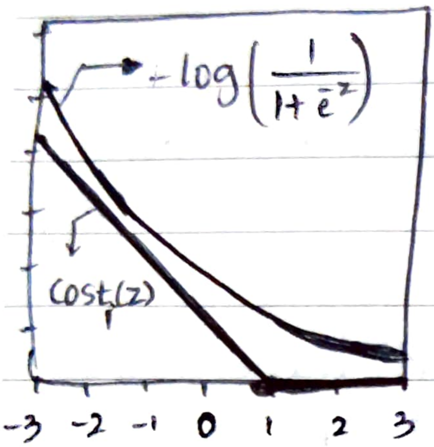
→ If the number of features is 2, then hyperplane is just a line. If the number of features is 3, then the hyperplane is just a 2-D plane.

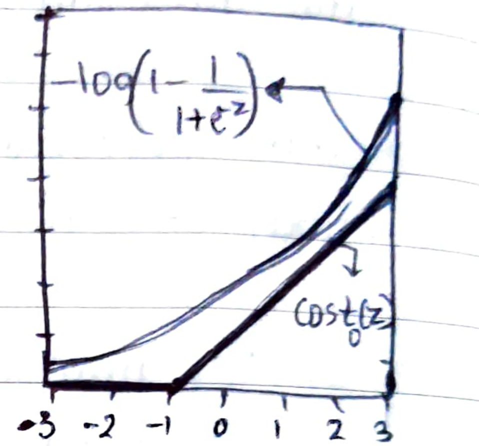
→ Support vectors are datapoints which influence the position and orientation of hyperplane i.e which are closer to hyperplane.

→ Cost function, which has to be minimized in

LOGISTIC:

$$\min_{\theta} \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\left(-\log h_\theta(x^{(i)})\right) + (1-y^{(i)})\left(\left(-\log\left(1-h_\theta(x^{(i)})\right)\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$





SUPPORT VECTOR:

$$\min_{\theta} \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\,\text{cost}_1(\theta^T(x^{(i)})) + (1-y^{(i)})\,\text{cost}_0(\theta^T(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

$$\Leftrightarrow \boxed{\min_{\theta} C\sum_{i=1}^{m}\left[y^{(i)}\text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)})\,\text{cost}_0(\theta^T x^{(i)})\right] + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2}$$

→ The hypothesis of SVM is straight forward,

if $\theta^T x \geq 0$; predict 1

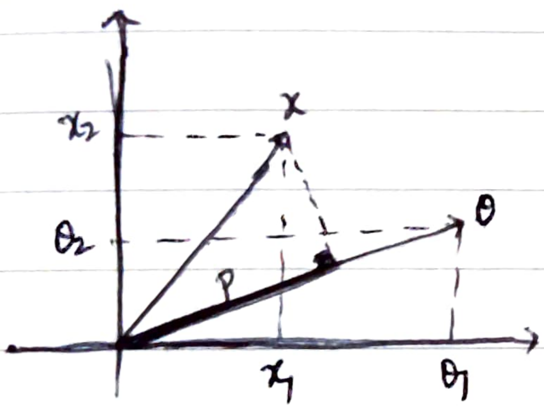
if $\theta^T x < 0$; predict 0

MORE INTO COST FUNCTION:

Cost function can be transformed as; for very large C as:

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 \quad \text{where;} \quad \begin{aligned} &\theta^T x^{(i)} \geq 1 \text{ if } y^{(i)} = 1 \\ &\theta^T x^{(i)} \leq -1 \text{ if } y^{(i)} = 0. \end{aligned}$$

UNDERSTANDING $\theta^T x^{(i)}$



$$\theta^T x^{(i)} = \underbrace{p^{(i)}}_{\text{Projected length}} \underbrace{\|\theta\|}_{\text{Length of vector } \theta} = \theta \left(x^{(i)}\right)^T$$

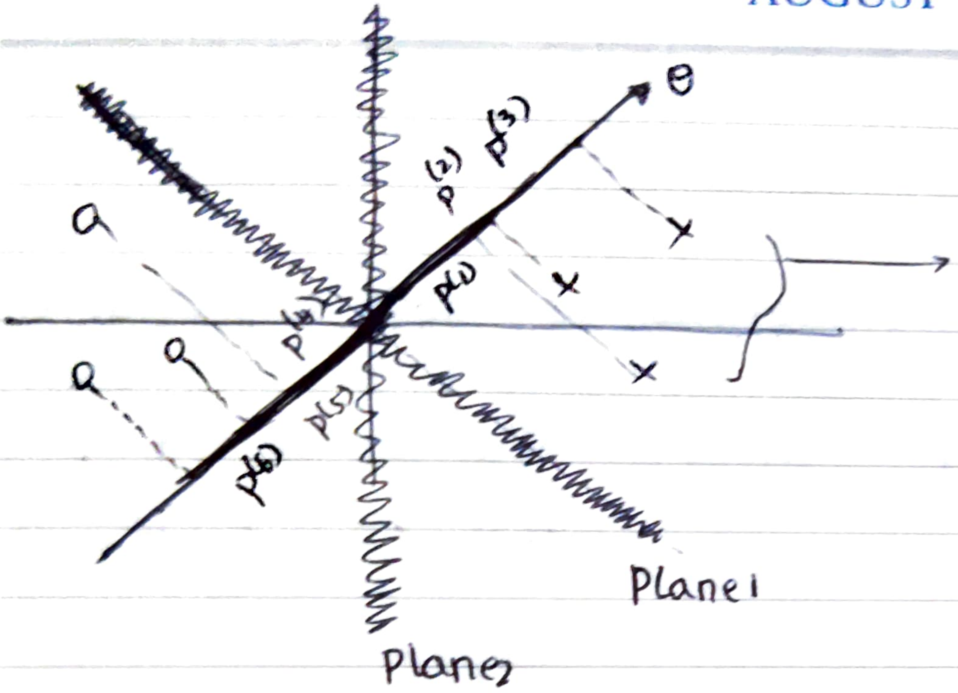
So;

For very large C.

$$\left\{ \min_{\theta} \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 = \frac{1}{2}\|\theta\|^2 \right. \quad \text{where;} \quad \begin{aligned} &p^{(i)} \cdot \|\theta\| \geq 1 \text{ if } y^{(i)} = 1 \\ &p^{(i)} \cdot \|\theta\| \leq -1 \text{ if } y^{(i)} = 0 \end{aligned}$$

UNDERSTANDING LARGE-MARGIN CLASSIFIER.

For simplicity let us take $\theta_0 = 0$.



$p^{(i)} ||\theta|| \geq 1$

But we will minimize $||\theta||$ so, $p^{(i)}$ should be greater so:

plane 2 will be found using SVM ie with more margin

NAIVE BAYES

- Naive Bayes classifiers are a collection of classifiers based on Bayes's Theorem.

- The fundamental assumption is that each feature makes an:
  - Independent
  - Equal

  contribution to outcome

- Bayes Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- So, wrt the dataset, we can apply Baye's Theorem as:

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$$

where, y is class variable & x is dependent feature vector (of size n) where:

$$x = (x_1, x_2, x_3 \dots, x_n)$$

$\rightarrow$ $$P(y|x_1, x_2, \ldots, x_n) = \frac{P(x_1|y) \cdot P(x_2|y) \ldots P(x_n|y) \cdot P(y)}{P(x_1) \cdot P(x_2) \ldots P(x_n)}$$

$$= \frac{P(y) \prod_{i=1}^{n} P(x_i|y)}{P(x_1) \cdot P(x_2) \ldots P(x_n)}$$

$$\propto P(y) \prod_{i=1}^{n} P(x_i|y).$$

$\rightarrow$ In gaussian Naive Bayes, each feature is assumed to be distributed according to gaussian.

i.e, $$\boxed{P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\left(\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)}}$$

$\rightarrow$ Other Naive Bayes Classifiers are:

1) Multinomial Naive Bayes
2) Bernoulli Naive Bayes