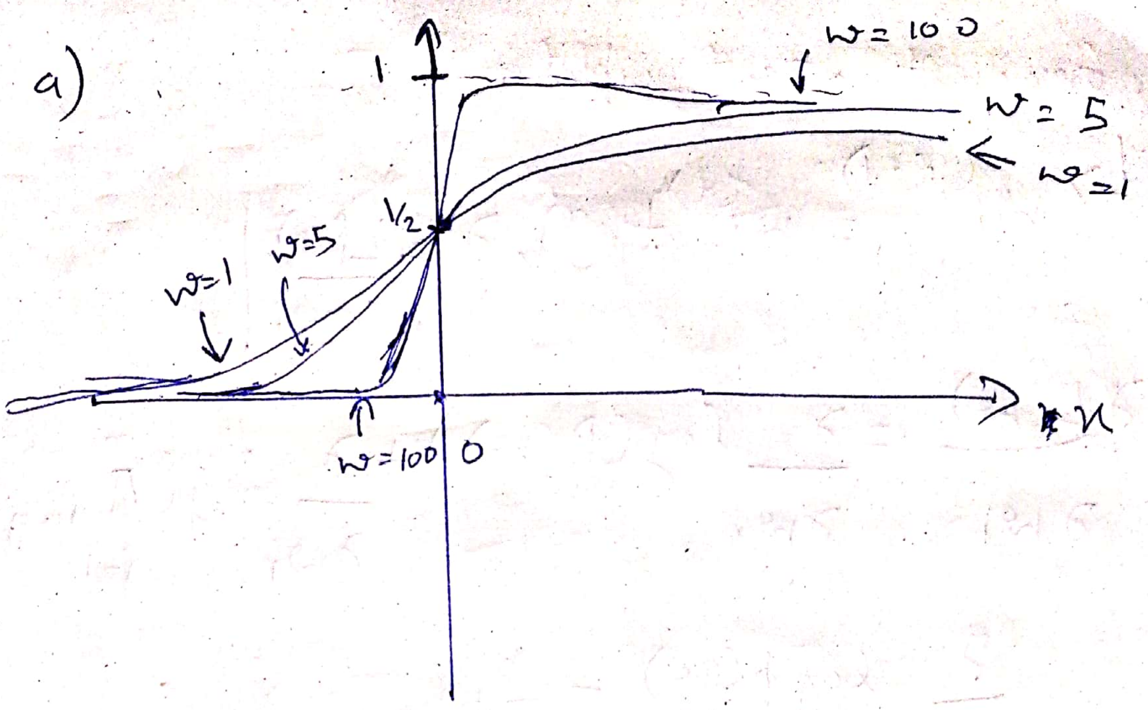
A1 a)



As w increases the curve gets steeper, so the model would be almost sure of the class (with almost 1 probability.) Even a little change in x the class ~~probabilitie~~ probabilites may change by a lot (when x near 0) when w is large. So a solution with large weights can cause overfiting.

A1 (b) $W_{MAP} = \underset{w_0,\dots,w_d}{\arg\max} \prod_{i=1}^{n} P(Y_i|X_i, w_0, \dots w_d) P(w_0, \dots, w_d)$

$$p(w) = \prod_{i=0}^{d} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-w_i^2}{2}\right)$$

$$w^* = \underset{w}{\arg\max}\ \log(W_{MAP})$$

$$= \arg\max \left[\sum_{j=1}^{n} \log(P(y^j|x^j, w)) - \sum_{i=0}^{d} \frac{w_i^2}{2} \right]$$

Gradient ascent update

$\hookrightarrow w_i^{(t+1)} = w_i^{(t)} + \eta \frac{\partial L(w)}{\partial w_i}\Big|_t$

$$\frac{\partial L(w)}{\partial w_i} = \frac{\partial}{\partial w_i} \log(P(w)) + \frac{\partial}{\partial w_i} \log\left(\prod_{j=1}^{n} P(y^j | x^j, w)\right)$$

$$\frac{\partial}{\partial w_i} \log p(w) = -w_i$$

Update rule:

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta\left(-w_i^{(t)} + \sum_j x_i^j \left(y^j - P(y=1 | x^j, w^{(t)})\right)\right)$$

A1(c) Probabilites of all the classes add up to 1

$$\Rightarrow P(y = y_k | X) = 1 - \sum_{k=1}^{K-1} P(Y = y_k | X)$$

$$P(Y = y_k | X) \propto \exp\left(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i\right), \text{ for } k = 1 \cdots K-1$$

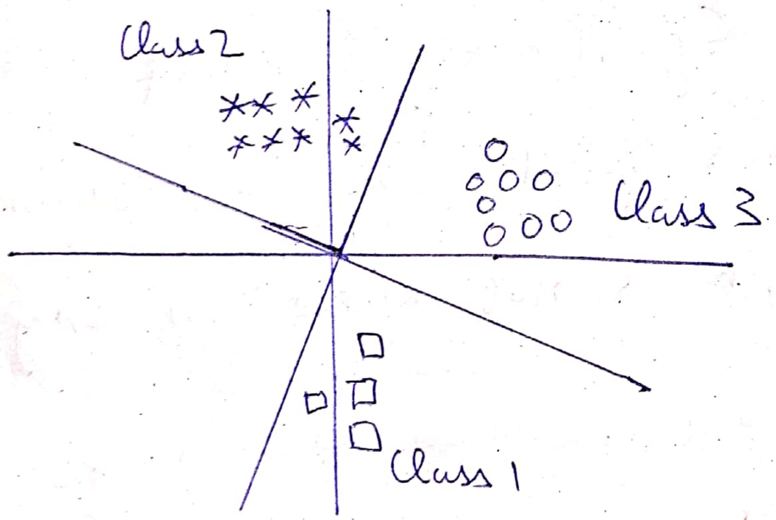
$$P(Y = y_k | X) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp\left(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i\right)}$$

for $k = 1, 2 \cdots K-1$

$$P(Y = y_k | X) = \frac{\exp(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}{1 + \sum_{k=1}^{K-1} \exp(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}$$

$y = y_{k^*}$ where $k^* = \underset{k \in \{1 \cdots K\}}{\arg\max} \; P(y = y_k | X)$.

A1d) The decision boundary between each class is linear $\Rightarrow$ overall decision ~~bounda~~ boundary is piece wise linear.



A2(a) $\hat{y} = \bar{y}(G + dI)^{-1}\bar{z}$

where $z = \langle \phi(x), \phi(\bar{x}) \rangle = K(x, \bar{x})$

z is a function of $\bar{x}$

$(G + dI)^{-1}$ is not a function of $\bar{x}$

$l(\bar{x}) = (G + dI)\bar{z}$

$\hat{y} = \bar{y}^T \cdot l(x)$ which is a dot product hence a linear smoother.

A2 b) No, it is not a linear smoother.

Counter Eg: Constant input where each point has $x_i = 1$. 'w' is median of y's.

Clearly, w is not linear in any of y's, ~~since~~ the median changes as the rank of y's does.

A2(c) Yes, it's a linear smoother.

$$l(x) : l(\bar{x})_i = \begin{cases} \frac{1}{|B_k|} & \text{if } x_i \in B_k \\ 0 & \text{else} \end{cases}$$

Range is divided into bins.

$f(x) \cdot \bar{y} \rightarrow$ Outputs the avg of value of the bin that $\bar{x}$ belongs.

Hence this a linear smoother.

)

dot

r.

each

y's.

y's,

rank