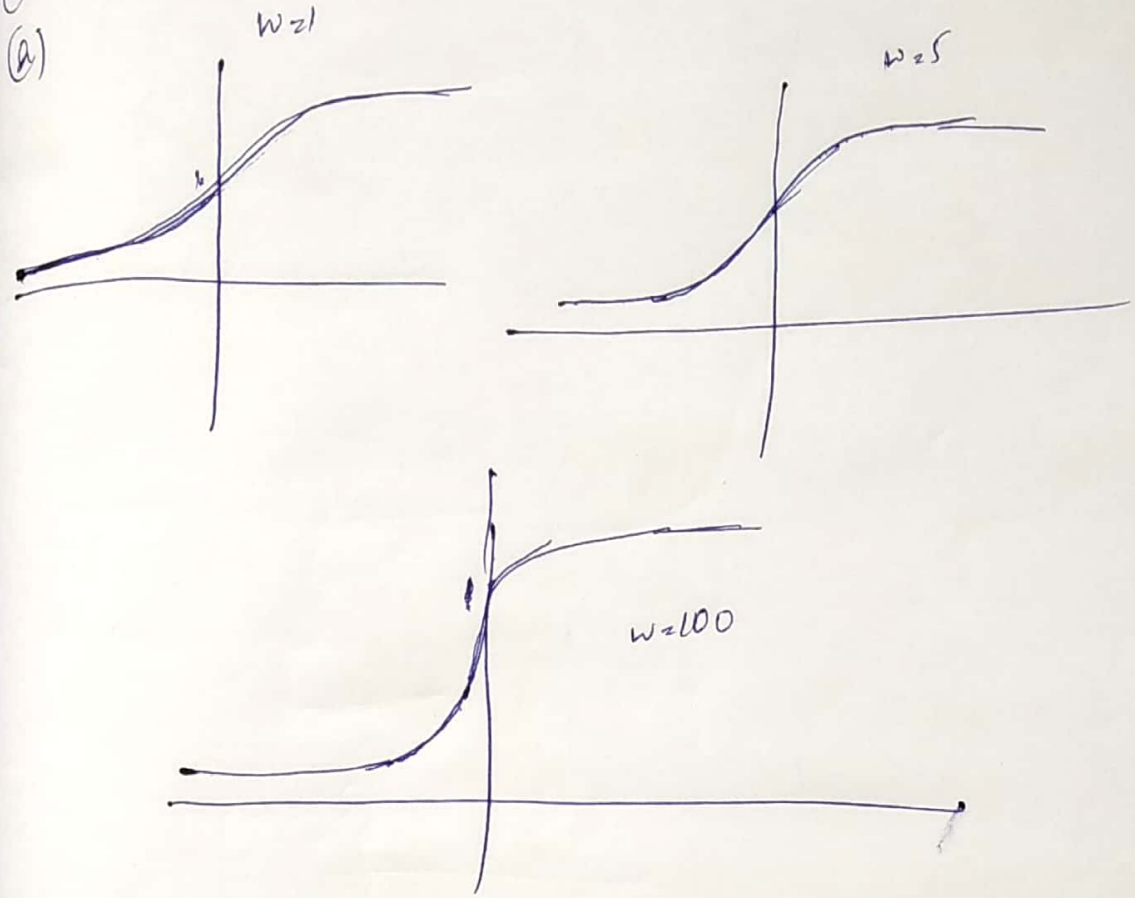
Homework - 3

G. Deepika
CS16BTECH11015

(1)

(a)



The curve is getting steeper as w increases. As steeper curve means that the model is nearly completely sure of the class, with larger weights even small changes in input can lead to a large change in probability of class thus leading to easy flipping of the predicted output.

(This is the reason we claim that it overfits)

(b) Let $w = [w_0, w_1, \ldots w_d]$

$$L(w) = \log\left[p(w) \prod_{j=1}^{n} p(y^j | x^j, w)\right]$$

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

$\therefore$ MAP estimate: $w^* = \arg\max_w L(w) = \arg\max_w \left[\sum_{j=1}^{n} \log\left(p(y^j | x^j, w)\right] - \sum_i \frac{w_i^2}{2}\right)$

Gradient ascent update rule: $w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \left.\frac{\partial L(w)}{\partial w_i}\right|_t$

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

$$= -w_i$$

⇒ Final update rule is:

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

(c) Since all probabilities must sum to 1,

$$P(Y=y_K / X) = 1 - \sum_{k=1}^{K-1} P(Y=y_k / X)$$

We can define,

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

and for $k = 1, \ldots, K-1$

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

[redacted]

Classification rule simply picks the label with highest probability:

$$y = y_{k^*} \text{ where } k^* = \arg\max_{k \in \{1, \ldots K\}} P(Y=y_k / X)$$

(d) The decision boundary between each pair of classes is linear and hence the overall decision boundary is piece-wise linear. Equivalently, since $\arg\max_i \exp(a_i) = \arg\max_i a_i$, and max of linear functions is piece-wise linear, the overall decision boundary is piece-wise linear.

(2)

(a) $k(x_i, x) = \exp \frac{\|x_i - x\|_2^2}{\sigma^2}$ (input $\bar{x}$)

for linear smoother,

$\hat{y} = \ell(x)^T \cdot y$

$$= \frac{\sum_{i=1}^{n} k(x_i, x) \cdot y_i}{\sum_{i=1}^{n} k(x_i, x)}$$

So, kernel regression is linear smoother

(b) Proof by counter example:

Lets say input : $x_i = k$ for different values of y

$(i \in \{1,2,3 \cdots y$

For optimal w makes the same number of +ve & -ve errors.

So that implies, w is the median of all y

and w is not linear in any of y's and the median changes as the scale of y changes.

∴ As w is not linear it doesnot for a linear smoother

(c) $\hat{y} = \frac{1}{|B_k|} \sum_{i: x_i \in B_k} y_i$ for $x \in B_k$ where $|B_k|$ = no. of points in B_k

* Regressogram is a linear smoother &

$$l(x) = \frac{I(x_j \in B(x))}{\sum_{i=1}^{n} I(x_i \in B(x))} \quad \text{where } |B_k| = \sum_{i=1}^{n} I(x_i \in B_k)$$

$$\boxed{l_i(x) = \frac{I(x_i \in B_k)}{|B_k|}}$$