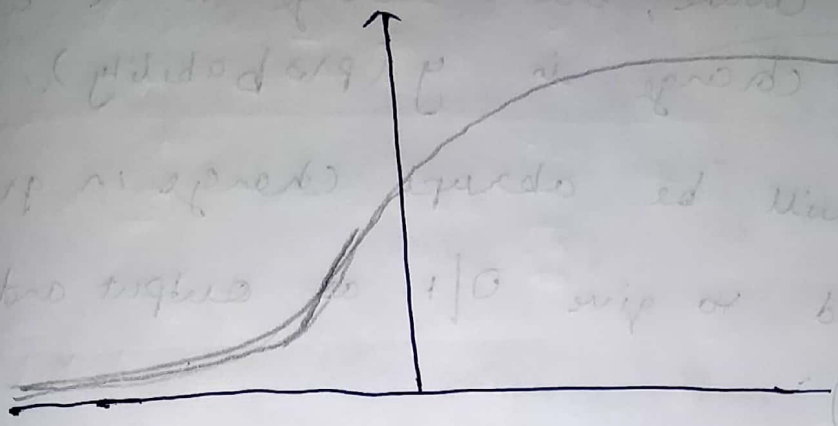
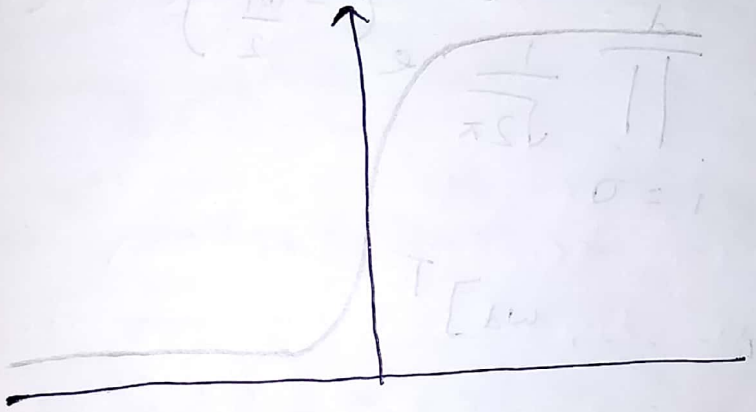
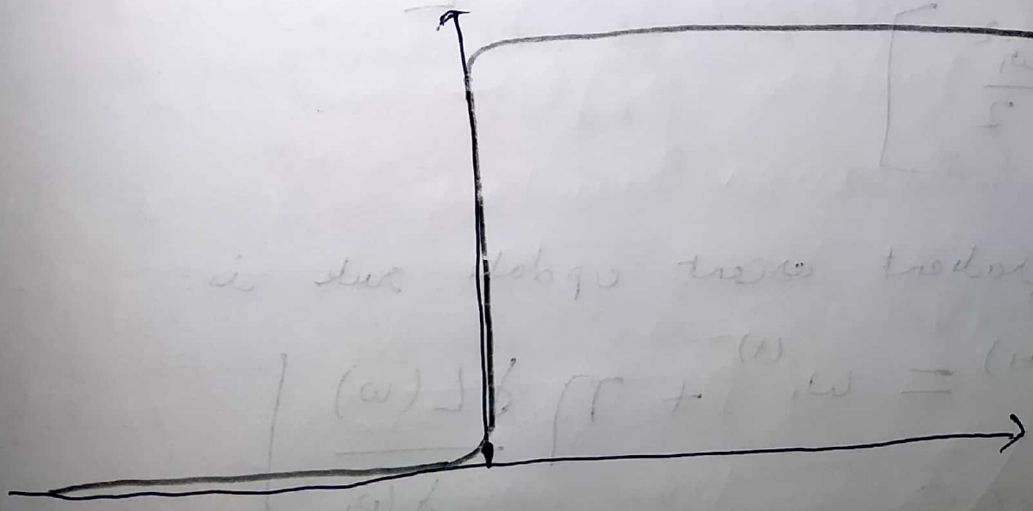
1(a) $\omega = 1$



(b) $\omega = 5$



(b) $\omega = 100$



As w increases, the curve gets steeper.

With steeper curve, same change in x will lead to very high change in y (probability).

Also there will be abrupt change in gradients

This will tend to give 0/1 as output and will overfit easily.

(b) $$L(w) = \log\left(p(w) \prod_{j=1}^{n} P(y^j / x^j, w)\right)$$

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

$$w = [w_0, \ldots, w_d]^T$$

MAP estimate is

$$w^* = \arg\max_w L(w) = \arg\max_w \left[\sum_{j=1}^{n} \log\left(P(y^j / x^j, w)\right) - \sum_i \frac{w_i^2}{2}\right]$$

The gradient ascent update rule is:

$$w_i^{(t+1)} = w_i^{(t)} + \eta \left.\frac{\partial L(w)}{\partial w_i}\right|_t$$

The gradient of the log conditional posterior is:

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

2nd term in above eqn is:

For unregularized

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

Thus finally update rule

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

(c) $P(Y = y_k | X) \propto \exp(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)$ for

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

$P(Y = y_k | X)$ is for same model

Since all the probabilities sum up to 1

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

Adding another set of weights is redundant

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

and for $k = 1, 2, \ldots k-1$

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

The classifier will simply pick the label with highest probability

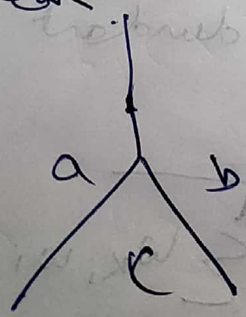
$$y = y_k^* \quad \text{where} \quad k^* = \operatorname{argmax}\left(P(Y = y_k / X)\right)$$

$$k \in \{1, \ldots k\}$$

(d) The decision boundary between each pair of classes is linear and hence overall decision boundary is piece wise linear

Equivalently since

$\operatorname{argmax} \exp(a_i) = \operatorname{argmax}(a_i)$, and max of linear function is piece wise linear the overall decision boundary is piece wise linear



(Decision boundary resulting from multiclass logistic regression)

2 (a) The given kernel function is a semi gaussian kernel function

Kernel regression

$$\hat{y} = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} \quad \text{Avg weight}$$

Now, for a point x_i, the w_i will be replaced by an kernel

$$w_i = \exp\left(\frac{||x_i - x||^2}{\sigma^2}\right)$$

$$l_i(x) = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

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

The given kernel Regression is linear smoother

(b) We are fitting a linear regression model sum of $||Hw - y||^2$, we know absolute value of residuals

Proof of not a linear smoother.

There is no closed form solution for w that minimizes the sum of absolute value of error

Yet, solution can be seen to be similar to median. An optimal w make the same

number of positive & negative error

Counter example

Consider a constant input where each training point has $x_i = 1$ for different y value so it is the median of all y.

w is not linear in any of y's since the median changes as the rank of y changes

(c) Dividing a, b in m equally spaced bins $B_1, B_2 \dots B_K$

$$\hat{y} = \frac{1}{|B_K|} \sum_{i=x_i=B_K} y_i \text{ for } x \in B_K$$

$|B_K|$ = no. of points in B_K

Definition of linear smoother $\left[\hat{y} = l(X^T | y)\right]$

we have $\hat{y} = \frac{1}{|B_K|} \sum_{i=x_i \in B_K} y_i$

By analysing func. we have $lg(x) = \frac{I(x_i \in B_i)}{|B_i|}$

Then linear smoother condition is satisfied.

Hence our regressogram is linear smoother

$$lg(x) = \frac{I}{|B_K|} \; ; x \in B_K$$