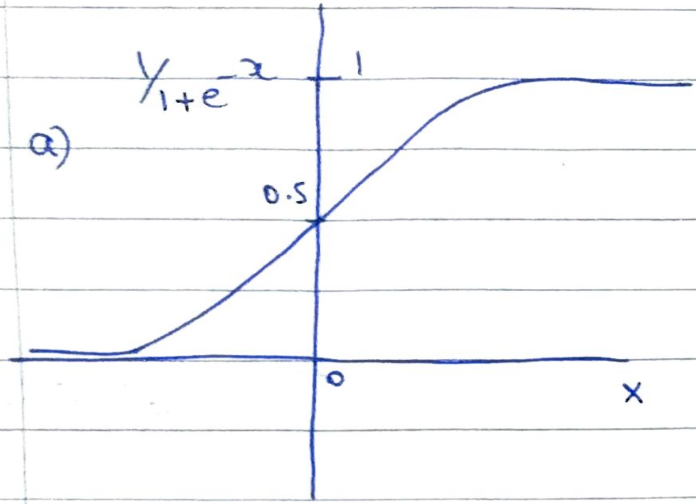
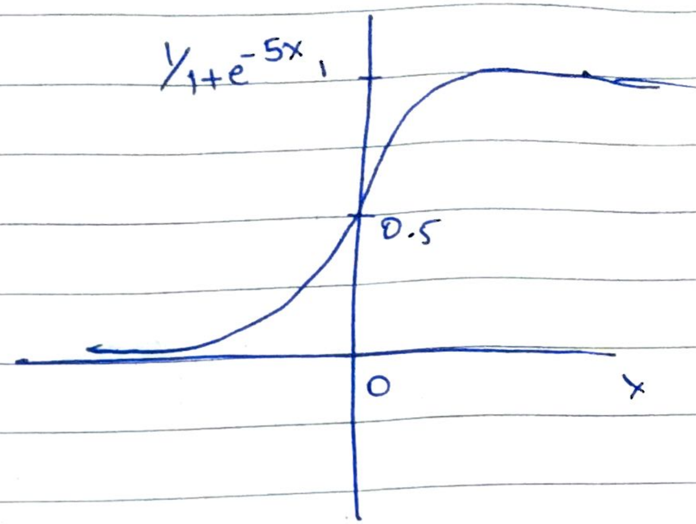
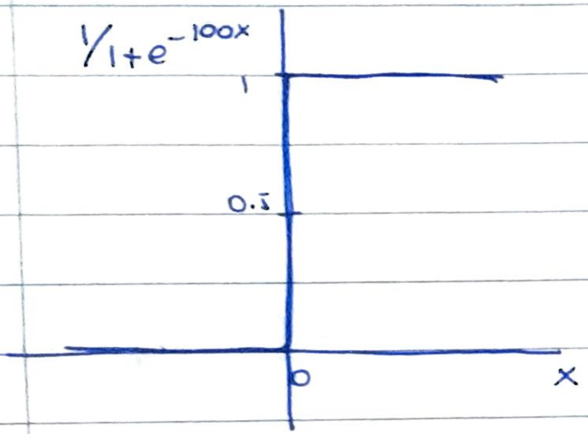
1) a)







As w increases, the curve becomes more ||el to y-axis and begins to look like step function.
The reason why it overfits as w increases is because with large value of w, small change in input leads to change in output label. ~~The point of probability never tends to affect the class~~

DATE

1) (b) MLE $\Rightarrow$ $w = \max_{w_0,..w_d} \prod_{i=1}^{n} P(Y_i / X_i, w_0, \dots w_d)$

MAP $\Rightarrow$ $w = \max_{w_0,\dots w_d} \prod_{i=1}^{n} P(Y_i / X_i, w_0, \dots w_d) \, P(w_0, \dots w_d)$.

Assume Standard Gaussian prior $N(0, I)$ Gradient Ascent?

Let $w = [w_0, \dots \dots w_d]$

For Gaussian distribution wkt

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 \, (x - \mu/\sigma)^2}$$

$$P(w) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 \, (w - \mu/\sigma)^2} \quad \text{For } N(0, I).$$

$$= \frac{1}{\sqrt{2\pi}} e^{-1/2 \, w^2}$$

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

From MLE

$$L(w) = \sum_i \left[y^i \ln P(y^i = 1 / x_i, w) + (1 - y^i) \ln P(y^i = 0 / x^i, w) \right]$$

$$+ \ln P(w)$$

$$\frac{\partial}{\partial w} L(w) = \sum_i \left[\frac{y^i}{\sigma(w^T x_i)} \sigma(w^T x_i)(1 - \sigma(w^T x_i)) \cdot x_i - \left(\frac{1 - y_i}{1 - \sigma(w^T x_i)} (1 - \sigma(w^T x_i)) \cdot \sigma(w^T x_i) - x_i \right] \right.$$

$$- w$$

$$\frac{\partial}{\partial w} L(w) = \sum_i x_i \left[y_i - P(Y = 1 / X, w) \right] - w$$

Gradient Ascent formula: $w_{t+1} \leftarrow w_t + \eta \left[\frac{\partial}{\partial w} L(w) \right]$

$$w_{t+1} \leftarrow w_t + \eta \left[-w_t + \sum_i x_i \left[y_i - P(Y = 1 / X, w) \right] \right]$$

DATE

1) (c) $\Rightarrow P(Y = y^k / X) \propto \exp\left(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i \right)$

From odd's ratio wrt $\ln\left(\frac{Pr(Y_i = 1/X_i)}{Pr(Y_i = K/X_i)} \right) = w_1 x_i$ where K is set of classes excluding 1.

$\ln\left(\frac{Pr(Y_i = 2/X_i)}{Pr(Y_i = K/X_i)} \right) = w_2 x_i$

$\vdots$

In General $\ln\left(\frac{Pr(Y_i = k/X_i)}{Pr(Y_i = K/X_i)} \right) = w_k x_i$

$\Rightarrow Pr(Y_i = k/X_i) = Pr(Y_i = K/X_i) \cdot e^{w_k x_i}$

$$Pr(Y_i = K/X_i) = 1 - \sum_{k=1}^{K-1} Pr(Y_i = k/X_i)$$

$$= 1 - \sum_{k=1}^{K-1} Pr(Y_i = K/X_i) \cdot e^{w_k x_i}$$

$$= 1 - \sum_{k=1}^{K-1} \frac{e^{w_k x_i}}{1 + \sum_{k=1}^{K-1} e^{w_k x_i}}$$ [Expand $Pr(Y_i = K/X_i)$]

$$= 1 - \frac{\left[e^{w_1 x_i} + e^{w_2 x_i} + \dots + e^{w_{K-1} x_i} \right]}{1 + \left[e^{w_1 x_i} + e^{w_2 x_i} + \dots + e^{w_{K-1} x_i} \right]}$$

$$Pr(Y_i = K/X_i) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{w_k x_i}}$$

$$Pr(Y_i = k/X_i) = \frac{e^{w_k x_i}}{1 + \sum_{k=1}^{K-1} e^{w_k x_i}}$$ [combined $w_0 + \sum_{i=1} w_{ki} X_i$ to $w_k X_i$]

The classification rule implies that we have the label with the highest probability.

$$y = y_{k^*} \text{ where } k^* = \underset{k \in \{1, \dots K\}}{\operatorname{argmax}} P(Y = y_k / X)$$

DATE

1) (d)