Question 1: the posterior probability expression is as follows:

$p(y=1 \mid x, w) = \hat{y}$
$p(y=0 \mid x, w) = 1-\hat{y}$

Question 2: the decision boundary expression is given by the sigmoid function and is thus:

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

the value should intercept 0 at 0.5 given that $p(y=1) = p(y=0)$.

Question 3:

$$L_{CE} = -[y \cdot \log \hat{y} + (1-y) \log(1-\hat{y})]$$

$$\hat{y} = \sigma(a) = \frac{1}{1+e^{-a}}$$

$$a = w^T x$$

Question 4:

$$\frac{dL(w)}{dw} = \frac{dL(w)}{d\hat{y}} \frac{d\hat{y}}{da} \frac{da}{dw} \quad \leftarrow \text{chain rule}$$

$L(w) = L(w(y(a)))$

$$\frac{dL(w)}{d\hat{y}} = -\left(y \cdot \frac{1}{\hat{y}} - 1-y \cdot \frac{1}{1-\hat{y}}\right)$$

$$= -\left(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right)$$

$$= -\left(\frac{y(1-\hat{y})}{\hat{y}(1-\hat{y})} - \frac{1-y(\hat{y})}{1-\hat{y}(\hat{y})}\right)$$

$$= -\left(\frac{y - \hat{y}y - (\hat{y} - y\hat{y})}{\hat{y}(1-\hat{y})}\right)$$

$$= \frac{\hat{y}-y}{\hat{y}(1-\hat{y})}$$

$$\frac{d\hat{y}}{da} = \frac{d}{da}[1+e^{-a}]^{-1} \quad \leftarrow \text{chain rule}$$

$$= [1+e^{-a}]^{-2} \cdot -e^{-a}$$

$$= \frac{e^{-a}}{[1+e^{-a}]^2} = \frac{e^{-a}}{1+e^{-a}} \cdot \frac{1}{1+e^{-a}}$$

$$\Rightarrow \left(1 - \frac{1}{1+e^{-a}}\right) \frac{1}{1+e^{-a}} \quad \boxed{= (1-\hat{y})\hat{y}}$$

$$\Rightarrow \underbrace{\left(1 - \frac{1}{1+e^{-a}}\right)}_{\left(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right) = \frac{e^{a}}{1+e^{-a}}} \frac{1}{1+e^{-a}} = (1-\hat{y})\hat{y}$$

$$\frac{da}{w} = x$$

$$\text{thus} \quad \frac{dL(w)}{dw} = \frac{\hat{y} - y}{\hat{y}(1-y)} \cdot \hat{y}(1-\hat{y}) \cdot x = \boxed{x(\hat{y} - y)}$$

Question 5: the acurracy will be affected by an imbalanced data set. Due to the high volumn of $y=0$ the training model may be more biased towards classifying that class thus producing more FN and less TP, Lowering TPR and Precision.

there are two approaches to combat Data Imbalance that I can recall.

1. Limit the amount of $y=0$ samples
   - By limiting the abundant $y=0$ samples we can reduce the bias. But there will be less samples to train the model
2. Generate more $y=1$ samples.
   - By generating more $y=1$ samples we reduce the imbalance However depending on the method choosen to produce more $y=1$ samples we could throw off the optimization of parameters, thus further reducing our TPR/Precision.