Question 4(e):

Let m denote the number of features

$$\frac{\partial J}{\partial w} * N = X^T(\mathrm{y}-\mathrm{t})$$

$$= \begin{matrix} x_1^{(1)} & \ldots & x_1^{(N)} \\ \ldots & \ldots & \ldots \\ x_m^{(1)} & \ldots & x_m^{(N)} \end{matrix} * \begin{matrix} y^{(1)} - t^{(1)} \\ \ldots \\ y^{(N)} - t^{(N)} \end{matrix}$$

$$\frac{\partial J}{\partial w} = \left(\frac{1}{N}\right) * \begin{matrix} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) x_1^{(i)} \\ \ldots . \\ \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) x_{1m}^{(i)} \end{matrix}$$

$$= \begin{matrix} \partial J / \partial w_1 \\ \ldots \\ \partial J / \partial w_m \end{matrix}$$

$$\therefore\ [\frac{X^T(y-t)}{N}]_j = (\frac{1}{N}) \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) x_1^{(i)}$$

$$\therefore\ [\frac{\partial J}{\partial w}]_j = [\frac{X^T(y-t)}{N}]_j \leftrightarrow \frac{\partial J}{\partial w_j} = (\frac{1}{N}) \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) x_1^{(i)}$$

Question 4(f):

Prove: $L_{LCE}(z,t) = L_{CE}(sigmoid(z),t)$

$$L_{CE}(y,t) = -t * log(y) - (1-t) * \log(1-y)$$

Substitute y for sigmoid(z)

$$L_{ce}(sigmoid(z),t) = -t * log\left(\frac{1}{1+e^{-z}}\right) - (1-t) * \log(1 - \frac{1}{1+e^{-z}})$$

$$= t * \log(1+e^{-z}) - (1-t) * \log(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}})$$

$$= t * \log(1+e^{-z}) - (1-t) * \log(\frac{e^{-z}}{1+e^{-z}})$$

$$= t * \log(1+e^{-z}) + (1-t) * \log(\frac{1+e^{-z}}{e^{-z}})$$

$$= t * \log(1+e^{-z}) + (1-t) * \log(1 + (e^{-z})^{-1})$$

$$= t * \log(1 + e^{-z}) + (1 - t) * \log(1 + e^{-z})$$

$$\therefore\ L_{LCE}(z, t) = L_{CE}(sigmoid(z), t)$$

Question 6(e):

I suspect that the biggest reason behind both occurrences is the fact that there are more differences between 4 and 7 then between 5 and 6. A 5 is more likely to look like a 6 and vice versa than a 4 is to a 7.

Question 6(f):

Because we only have two classes, an odd number of k prevents a tie in the decision. In each prediction, one class will hold the majority of predictions

Question 6(g):

Because numbers were designed to look different from one another and are unlikely to have many outliers in comparison to some other data sets. This lack of outliers makes it a great candidate for KNN because any given hand drawn digit is highly likely to look like another hand drawn version of the same

digit