Q4.

(a) Newton Raphson for Logistic Regression

- Logistic regression says

$$p(\bar{x}_i) = \frac{1}{1+e^{-\bar{\beta}^T \bar{x}_i}} = p_i$$

Consequently,

$$\log\left[\frac{p(\bar{x}_i)}{1-p(\bar{x}_i)}\right] = \bar{x}_i^T \bar{\beta} \rightarrow (1)$$

- Also,

$$P[y_1, y_2, y_3 \dots y_N \mid p_1, p_2, \dots p_N] = \prod_{i=1}^{N} \left[p_i^{y_i} (1-p_i)^{1-y_i}\right]$$

$\Rightarrow$ Log-Likelihood ℓ

$$= \sum y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

$$\boxed{\ell(\beta) = \sum_{i=1}^{N} \log(1-p_i) + y_i \log\left(\frac{p_i}{1-p_i}\right)} \rightarrow (2)$$

$$\therefore \ell(\beta) = \sum_{i=1}^{N} y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)$$

$$\boxed{\ell(\beta) = \sum_{i=1}^{N} y_i (\bar{x}_i^T \bar{\beta}) - \log(1+e^{\bar{x}_i^T \bar{\beta}})} \rightarrow (3)$$

From (1) & (2)

- Computing Gradient,

$$\nabla \ell(\beta) = \sum_{i=1}^{N} y_i \frac{\partial(\bar{x}_i^T \bar{\beta})}{\partial \beta} - \frac{\partial}{\partial \beta} \log(1+e^{\bar{x}_i^T \bar{\beta}})$$

$$= \sum_{i=1}^{N} y_i \bar{x}_i - \bar{x}_i \left(\frac{e^{\bar{x}_i^T \bar{\beta}}}{1+e^{\bar{x}_i^T \bar{\beta}}}\right)$$

$$\boxed{\nabla \ell(\beta) = \sum_{i=1}^{N} \bar{x}_i [y_i - p_i]} \rightarrow (4)$$

- Computing Hessian,

$$\nabla(\nabla l(\beta)) = -\sum_{i=1}^{N} \nabla\left(\bar{x}_i \left[\frac{e^{\bar{x}_i^T \bar{\beta}}}{1+e^{\bar{x}_i^T \bar{\beta}}}\right]\right)$$

$$= -\sum_{i=1}^{N} \nabla\left(\bar{x}_i \left[\frac{1}{1+e^{-\bar{x}_i^T \bar{\beta}}}\right]\right)$$

$$= -\sum_{i=1}^{N} \bar{x}_i \left(+\bar{x}_i^T \left(\frac{e^{-\bar{x}_i^T \bar{\beta}}}{(1+e^{-\bar{x}_i^T \beta})^2}\right)\right)$$

$$= -\sum_{i=1}^{N} \bar{x}_i \bar{x}_i^T \left(\frac{e^{-\bar{x}_i^T \bar{\beta}}}{1+e^{-\bar{x}_i^T \bar{\beta}}}\right)\left(\frac{1}{1+e^{-\bar{x}_i^T \bar{\beta}}}\right)$$

$$\boxed{\nabla^2 l(\beta) = -\sum_{i=1}^{N} p_i(1-p_i)\,\bar{x}_i \bar{x}_i^T}$$

In matrix form,

$$\nabla l(\beta) = X^T(\bar{y} - \bar{p})$$

$$\nabla^2 l(\beta) = -X^T \underbrace{\begin{bmatrix} p_1(1-p_1) & & & \\ & p_2(1-p_2) & & \\ & & \ddots & \\ & & & p_N(1-p_N) \end{bmatrix}}_{p^*(\beta)} X = -X^T p^*(\beta) X$$

<u>Methodology</u>:

The update equation is

$$\bar{\beta}_{i+1} = \bar{\beta}_i - [\nabla^2 l(\beta)]^{-1}(\nabla l(\beta))$$

$$\boxed{\bar{\beta}_{i+1} = \bar{\beta}_i + (X^T \bar{p}^*(\beta) X)^{-1} X^T(\bar{y} - \bar{p})}$$

→ We update $\bar{\beta}$ till $\nabla l(\beta) < \varepsilon$, where ε is some low error ≈ 0 that we decide.

(b) In the Newton Raphson update scheme, each parameter's "importance" is calculated via the gradient. That is, the more "important" a given parameter value is to the output, the higher the partial derivative of the output wrt the parameter would be.

- Since this is the case, the "more important" parameter is varied more [courtesy of the Newton-Raphson update equation] than those that are "less important".
- This "importance" is nothing but the weight associated with a parameter, as explained in Q 3.(c).
- At every iteration, a weighted error is calculated and the parameter values are varied accordingly.
- Hence, the Newton-Raphson method, ~~iterat~~ calculating weighted errors at each iteration, with the weights changing at each iteration (re-weighted), is known as ITERATIVE RE-WEIGHTED LEAST SQUARES METHOD.

~~(c) The error fu~~

(c) To show that the error function is ~~concave~~ convex [and hence has a unique minimum], we need to show that the Hessian matrix (derived in 4(a)) is <u>negative semi-definite</u>.

Error is a function of $-H(\beta)\{-\nabla^2 \ell(\beta)\}$. Hence, $-H(\beta)$ MUST be POSITIVE SEMI-DEFINITE. i.e. $H(\beta)$ must be NEGATIVE SEMI-DEFINITE

▲ For error to be CONVEX

$$H(\beta) = -X^T \bar{P}^x(\beta) X.$$

$$= -\sum P_i(1-P_i)\,\bar{x}_i\,\bar{x}^T$$

→ The condition for this is that $\forall\ \bar{V} \in (N \times 1)$,

$\bar{V}^T H(\beta) \bar{V}$ is always ≤ 0.

$$\bar{V}^T H(\beta)\bar{V} = -\sum \underbrace{\overbrace{\bar{V}^T \bar{x}_i}^{A_1}\,\overbrace{\bar{x}^T \bar{V}}^{A_2}}_{A}\ \underbrace{P_i(1-P_i)}_{b}$$

~~→ A has entirety~~

→ • $A_1 = A_2$, since $\bar{V}^T \bar{x}_i = \bar{x}_i^T \bar{V}$ - dot product of $\bar{V}$ & $\bar{x}_i$ ↳ (1)

• ~~as~~ $0 \leq P_i \leq 1$, $\therefore$ $0 \leq b \leq 1$ → (2)

→ From (1) & (2), we have

$$0 \leq \bar{V}^T \bar{x}_i\,\bar{x}_i^T \bar{V}\ P_i(1-P_i) \leq 1$$

$$\Downarrow$$

$$-1 \leq -(\bar{V}^T \bar{x}_i\,\bar{x}_i^T \bar{V}\ P_i(1-P_i)) \leq 0$$

$$\Downarrow$$

$$-N \leq \boxed{-\sum_{i=1}^{N} (\bar{V}^T \bar{x}_i\,\bar{x}_i^T \bar{V}\ P_i(1-P_i)) \leq 0}$$

$$\boxed{\therefore\ \bar{V}^T H(\beta)\bar{V} \leq 0\ \forall\ \bar{V} \in (N \times 1)}$$

→ $H(\beta)$ is negative semi-definite ⇒ The error function has a unique minimum