Question 1. Logistic Regression - Theory.

a) Due to binarity of y we can rewrite the definition as

$$P(y = y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}) = y^{(i)}\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}) + (1 - y^{(i)})(1 - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))$$

One of the summands will always cancel, reducing the expression to the given definition of the distribution.

b) Let's first tackle the derivative $\frac{\partial \log(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))}{\partial\theta_j}$.

$$\frac{\partial \log(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})))}{\partial\theta_j} = \frac{\partial \log(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))}{\partial(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))}\frac{\partial(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))}{\partial(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})}\frac{\partial(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})}{\partial(\theta_j)}$$

Using the derivative of the sigmoid identity (page 31, lecture 3) we proceed

$$= \frac{1}{\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})}\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})(1 - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))x_j^{(i)} = (1 - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))x_j^{(i)}$$

Similarly we have

$$\begin{aligned}\frac{\partial \log(1 - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})))}{\partial\theta_j} &= \frac{1}{1 - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})}\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}) - 1)x_j^{(i)} \\ &= -\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})x_j^{(i)}\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial(\log P(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}))}{\partial\theta_j} &= \sum_i y^{(i)}\frac{\partial \log(\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))}{\partial\theta_j} + (1 - y^{(i)})\frac{\partial \log(1 - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))}{\partial\theta_j} \\ &= \sum_i y^{(i)}(1 - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))x_j^{(i)} - (1 - y^{(i)})\sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)})x_j^{(i)} \\ &= \sum_i x_j^{(i)}(y^{(i)} - \sigma(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}))\end{aligned}$$

c) For total degree of i there exists $i+1$ combinations of x_1 and x_2 powers (via elementary combinatorics). Therefore the total size of the new $\hat{\boldsymbol{\theta}}$ is going to be

$$\sum_{i=0}^{d} i + 1 = 1 + \frac{d(1+d)}{2}$$

We also expand $\hat{\boldsymbol{x}}$ to accommodate the combinations of the higher powers of features and it is going to have the same size as $\hat{\boldsymbol{\theta}}$. To order the elements of the two vectors we can map $x_1^i x_2^j$ to $\hat{x}_{i\cdot d+j}$ with the corresponding weight $\hat{\boldsymbol{\theta}}_{i\cdot d+j}$.

Question 2. Logistic Regression Implementation.

a) See Figure 1.

For b), c), d) See the attached files

e) See Figure 2.

f) See Figure. Although asked to plot a decision boundary for either degree 3 or 5 polynomials, the accuracy obtained with a degree 2 polynomial was 1.

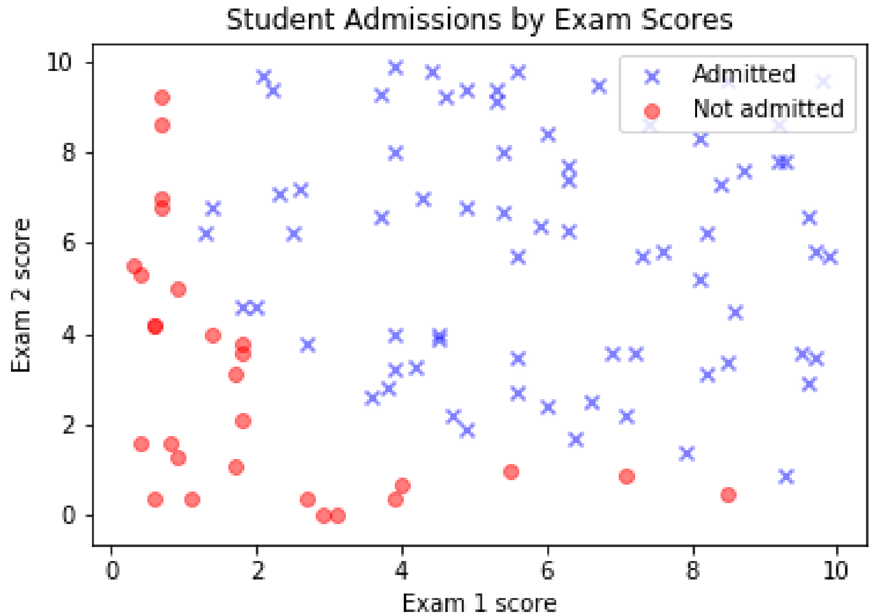

Figure 1: Scores on exam 2 plotted against scores on exam 1. Markers indicate admission outcome. Data from the training dataset.

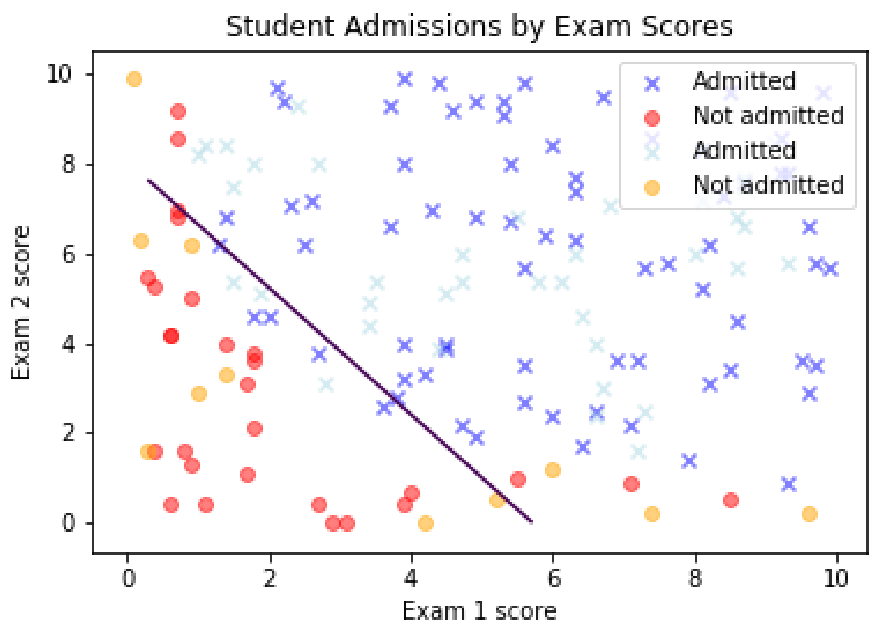

Figure 2: Linear decision boundary. Data from both training and test datasets. Accuracy obtained was 0.86.

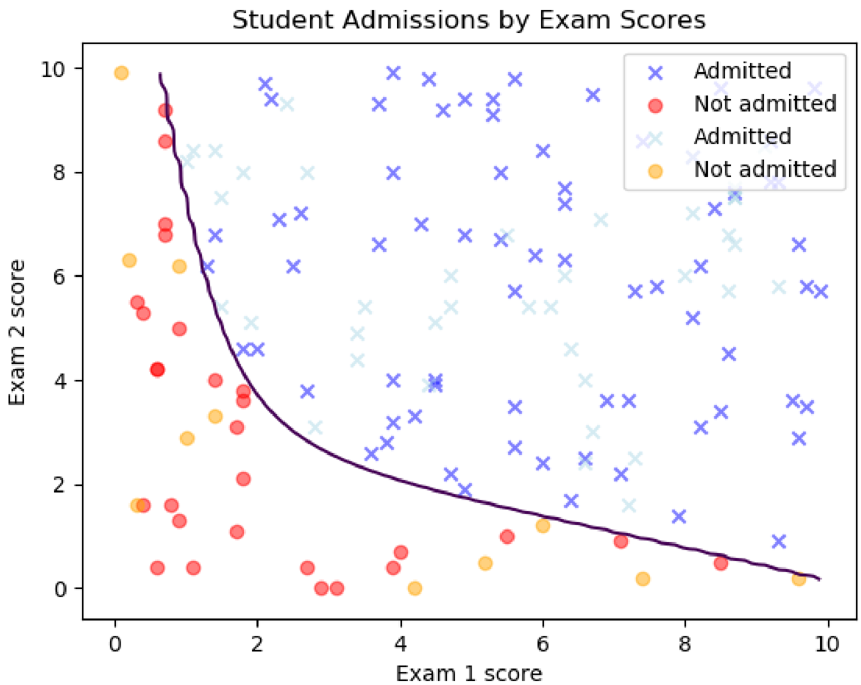

Figure 3: Decision boundary for a degree 3 polynomial. Accuracy obtained was 1.