Problem 2

Minimizing the L2 penalized logistic regression cost function

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)})) + \frac{\lambda}{2m}\sum_{j=1}^{d}\theta_j^2$$

1. (True/False) $J(\theta)$ has multiple locally optimal solutions.
False, because when minimized, $J(\theta)$ will have a unique global maximum because the Hessian is positive definite.

2. (True/False) Let $\theta^* = argmin_\theta J(\theta)$ be a global optimum. θ^* is sparse.
False, θ^* is not sparse when $\theta^* = argmin_\theta J(\theta)$ is a global optimum because the L2 penalized logistic regression just mimizes the values of theta so that they become close to 0, but don't actually become 0 unless the regularization term is infinity. This is due to the nature of the error term as a squared error (quadratic) will start taking smaller step sizes towards zero once the weights approach zero, therefore never actually reaching zero but can result in theta terms that are very small.

3. (True/False) If the training data is linearly separable, then some coefficients θ_j might become infinite if $\lambda = 0$.
True, if the training data is linearly separable, then the data should be able fully separated by an infinite number of lines (hyperplanes) graphically. The θ_j coefficients should be able to map to that line and the cost function, when minimized, should accurately point to all these hyperplanes. This means that the θ_j coefficients can map to an infinite number of hyperplanes, of which some coefficients θ_j could very well be infinite.

4. (True/False) The first term of $J(\theta^*)$ always increases as we increase λ.
False, the first term of $J(\theta^*)$ will may not always increase as we increase λ. Since $J(\theta^*)$ is the minimum cost corresponding to the optimal theta, as λ increases, the cost gets lower as we subtract the regularization term from the cost. Therefore, as λ increases, the cost function most likely actually gets lower, and once a minimum is found, it can be either greater or lower than a different value of λ, depending on how the function looks. Most likely, however, the cost will be lower as it started at a lower position.