Midterm 1 Practice Problems *(find and work with a partner)*

1. *Logistic Regression.* Say we have $p = 1$ and two training examples: $(x_1, y_1) = (3, 0)$ and $(x_2, y_2) = (7, 1)$, and we would like to fit a logistic model to this dataset.

   (a) Draw these two examples on a coordinate system and sketch a logistic function that would fit them. What is the linear decision boundary that should be created?

   (b) In terms of $h_{\boldsymbol{b}}(x)$, write out the likelihood $L(\boldsymbol{b})$ of observing this data (don't need to simplify).

   (c) Say in our SGD method, we choose to analyze (x_1, y_1) first. Before starting SGD, we set $b_0 = 0$ and $b_1 = 0$. After analyzing (x_1, y_1), what are b_0 and b_1? edit: choose $\alpha = 0.1$

   (d) If instead we had analyzed (x_2, y_2) first, what would b_0 and b_1 be after this initial data point?

   (e) What family of solutions do we obtain for $\hat{b}_0$ and $\hat{b}_1$? Note: we can't really see this through paper and pencil calculations. What happens as their magnitudes increase?

2. How did we define model complexity for:

   (a) k-Nearest Neighbors

   (b) Decision Trees

   (c) Polynomial Regression

3. If we have n training examples and (correction) m test examples, and we sort all distances to find the k smallest ones for each test example, then the runtime of k-Nearest Neighbors is:

   (a) $O(nm)$
   (b) $O(mn^2 \log n)$
   (c) $O(m(n + n \log n))$

   Explain your answer.