DS 100: Principles and Techniques of Data Science **Date: October 24, 2018**

Discussion #8

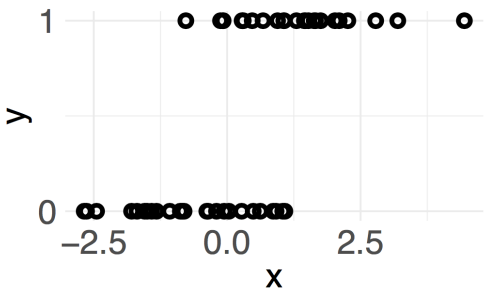
Name:

Logistic Regression

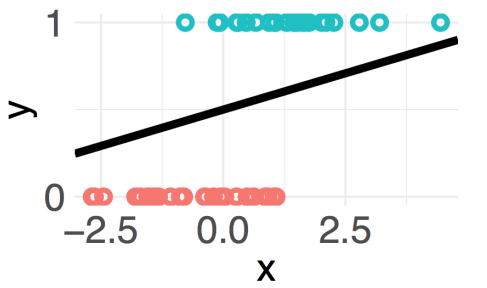
1. State whether the following claims are true or false. If false, provide a reason or correction.
   (a) A binary or multi-class classification technique should be used whenever there are categorical features.
   (b) A classifier that always predicts 0 has test accuracy of 50% on all binary prediction tasks.
   (c) In logistic regression, predictor variables are continuous with values from 0 to 1.
   (d) In a setting with extreme class imbalance in which 95% of the training data have the same label it is always possible to get at least 95% testing accuracy.

The next two questions refer to a binary classification problem with a single feature x.

2. Based on the scatter plot of the data below, draw a reasonable approximation of the logistic regression probability estimates for $\mathbb{P}(Y = 1 \mid x)$



3. Your friend argues that the data are linearly separable by drawing the line on the following plot of the data. Argue whether or not your friend is correct.



4. You have a classification data set:

x	y
1	0
-1	1

You run an algorithm to fit a model for the probability of $Y = 1$ given x:

$$\mathbb{P}(Y = 1 \mid x) = \sigma(\phi^T(x)\theta)$$

where $\phi(x) = \begin{bmatrix}1 & x\end{bmatrix}^T$. Your algorithm returns $\hat{\theta} = \begin{bmatrix}-\frac{1}{2} & -\frac{1}{2}\end{bmatrix}^T$

(a) Calculate $\hat{\mathbb{P}}(Y = 1 \mid x = 0)$

(b) Recall that the average cross-entropy loss is given by

$$\begin{aligned} L(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} -\mathbb{P}(y_i = k \mid x_i)\log\hat{\mathbb{P}}(y_i = k \mid x_i) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i\phi_i^T\theta + \log(\sigma(-\phi_i^T\theta))\right] \end{aligned}$$

where $\phi_i = \phi(x_i)$. Let $\theta = \begin{bmatrix}\theta_0 & \theta_1\end{bmatrix}$. Explicitly write out the (empirical) loss for this data set in terms of θ_0 and θ_1.

(c) Calculate the loss of your fitted model $L(\hat{\theta})$.

(d) Are the data linearly separable? If so, write the equation of a hyperplane that separates the two classes.

(e) Does your fitted model minimize cross-entropy loss?