DS 100/200: Principles and Techniques of Data Science **Date: April 5, 2019**

Discussion #9

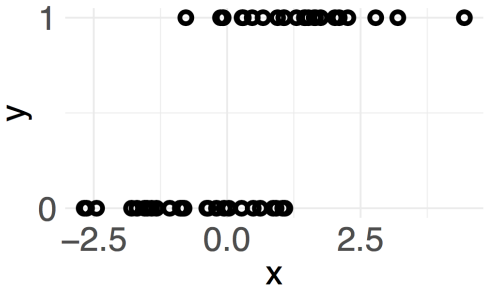
Name:

Logistic Regression

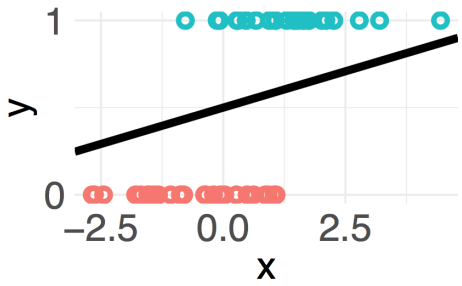
1. State whether the following claims are true or false. If false, provide a reason or correction.
   (a) A binary or multi-class classification technique should be used whenever there are categorical features.
   (b) A classifier that always predicts 0 has a test accuracy of 50% on all binary prediction tasks.
   (c) For a logistic regression model, all features 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 consisting of two (x, y) pairs $(1, 0)$ and $(-1, 1)$. The covariate vector $\mathbf{x}$ for each pair is a two-element column vector $\begin{bmatrix}1 & x\end{bmatrix}^T$. You run an algorithm to fit a model for the probability of $Y = 1$ given $\mathbf{X}$:

$$\mathbb{P}(Y = 1 \mid \mathbf{X}) = \sigma(\mathbf{X}^T\beta)$$

where

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Your algorithm returns $\hat{\beta} = \begin{bmatrix}-\frac{1}{2} & -\frac{1}{2}\end{bmatrix}^T$

(a) Calculate $\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{X} = \begin{bmatrix}1 & 0\end{bmatrix}^T\right)$

(b) The empirical risk using log loss (a.k.a., cross-entropy loss) is given by:

$$\begin{aligned} R(\beta) &= \frac{1}{n}\sum_{i=1}^{n} -\log \hat{\mathbb{P}}(Y = y_i \mid \mathbf{x_i}) \\ &= -\frac{1}{n}\sum_{i=1}^{n} y_i \log \hat{\mathbb{P}}(Y = 1 \mid \mathbf{x_i}) + (1 - y_i)\log \hat{\mathbb{P}}(Y = 0 \mid \mathbf{x_i}) \end{aligned}$$

And $\hat{\mathbb{P}}(Y = 1 \mid \mathbf{x_i}) = \frac{\exp(\mathbf{x_i}^T\beta)}{1+\exp(\mathbf{x_i}^T\beta)}$ while $\hat{\mathbb{P}}(Y = 0 \mid \mathbf{x_i}) = \frac{1}{1+\exp(\mathbf{x_i}^T\beta)}$. Therefore,

$$\begin{aligned} R(\beta) &= -\frac{1}{n}\sum_{i=1}^{n} y_i \log \frac{\exp(\mathbf{x_i}^T\beta)}{1 + \exp(\mathbf{x_i}^T\beta)} + (1 - y_i)\log \frac{1}{1 + \exp(\mathbf{x_i}^T\beta)} \\ &= -\frac{1}{n}\sum_{i=1}^{n} y_i \mathbf{x}_i^T\beta + \log(\sigma(-\mathbf{x}_i^T\beta)) \end{aligned}$$

Let $\beta = [\beta_0 \quad \beta_1]$. Explicitly write out the empirical risk for the data set $(1, 0)$ and $(-1, 1)$ as a function of β_0 and β_1.

(c) Calculate the empirical risk for $\hat{\beta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$ and the two observations $(1, 0)$ and $(-1, 1)$.

(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?