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

Discussion #9 Solutions

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.

   > **Solution:** False. Categorical features may appear in both classification and regression settings. They are often addressed with one-hot encoding.

   (b) A classifier that always predicts 0 has a test accuracy of 50% on all binary prediction tasks.

   > **Solution:** False. Class imbalances could lead to substantially higher or lower accuracy.

   (c) For a logistic regression model, all features are continuous, with values from 0 to 1.

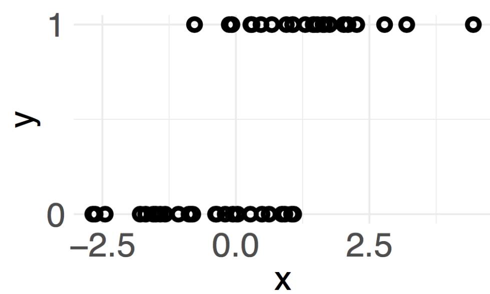
   > **Solution:** False. There is no such constraint on the features that predictor variables might take.

   (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.

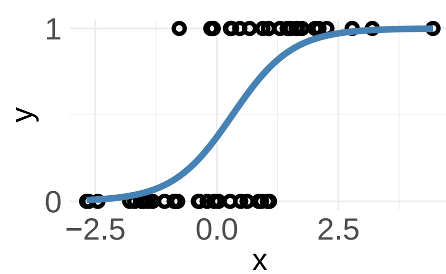
   > **Solution:** False. The test accuracy could be much lower depending on the class imbalance in the test data.

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

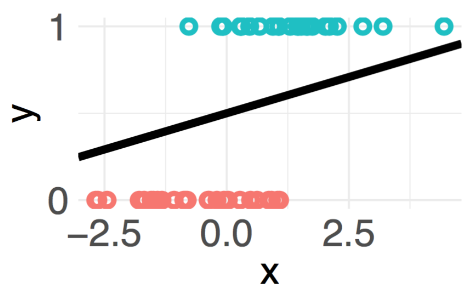
2. Based on the scatter plot of the data below, draw a reasonable approximation of the logistic regression probability estimates for $\mathbb{P}(Y = 1 \mid x)$.



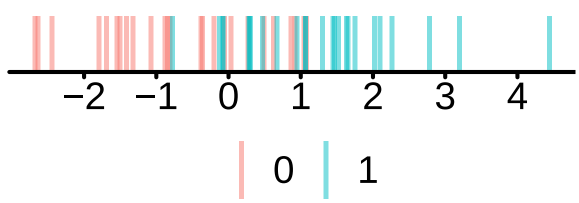
Solution:



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.



Solution: The scatter plot of x against y isn't the graph you should be looking at. The more salient plot would be the $d = 1$ representation of the features colored by class labels.



From this plot, it's clear that we can't draw a $d = 0$ plane (a point on the axis) that separates the data.

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 $[1 \quad x]^T$.

   You run an algorithm to fit a model for the probability of $Y = 1$ given $\mathbf{X}$:

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

   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 $\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{X} = [1 \quad 0]^T\right)$

Solution:

$$\begin{aligned}
\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{X} = [1 \quad 0]^T\right) &= \sigma\left(\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}-\frac{1}{2} \\ -\frac{1}{2}\end{bmatrix}\right) \\
&= \sigma\left(1 \times -\frac{1}{2} + 0 \times -\frac{1}{2}\right) \\
&= \sigma\left(-\frac{1}{2}\right) \\
&= \frac{1}{1 + \exp(\frac{1}{2})} \\
&\approx 0.38
\end{aligned}$$

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

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

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

Solution:

$$x_i^T\beta = \begin{bmatrix}1 & x_i\end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} = \beta_0 + \beta_1 x_i$$

For the data point $(1,0)$, $\mathbf{x}_i = [1 \quad 1]^T$ and $y_i = 0$, so:

$$y_i\mathbf{x}_i^T\beta = 0$$

$$-\mathbf{x}_i^T\beta = -(\beta_0 + \beta_1 \times 1) = -\beta_0 - \beta_1$$

For the data point $(-1,1)$:

$$\begin{aligned} y_i x_i^T\beta &= 1 \times (\beta_0 + \beta_1 \times -1) = \beta_0 - \beta_1 \\ -x_i^T\beta &= -(\beta_0 + \beta_1 \times -1) = -\beta_0 + \beta_1 \end{aligned}$$

We can then write the empirical risk as:

$$\begin{aligned} R(\beta) &= -\frac{1}{2}\left[(0 + \log\sigma(-\beta_0 - \beta_1)) + (\beta_0 - \beta_1 + \log\sigma(-\beta_0 + \beta_1))\right] \\ &= -\frac{1}{2}\left[\beta_0 - \beta_1 + \log\sigma(-\beta_0 - \beta_1) + \log\sigma(-\beta_0 + \beta_1)\right] \\ &= -\frac{1}{2}\left[\beta_0 - \beta_1 + \log\left(\frac{1}{1+\exp(\beta_0+\beta_1)}\right) + \log\left(\frac{1}{1+\exp(\beta_0-\beta_1)}\right)\right] \\ &= \frac{1}{2}\left[\beta_1 - \beta_0 + \log\left(1+\exp(\beta_0+\beta_1)\right) + \log\left(1+\exp(\beta_0-\beta_1)\right)\right] \end{aligned}$$

(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 $(-1, 1)$.

Solution:

$$\begin{aligned} R(\hat{\beta}) &= \frac{1}{2}\left[\beta_1 - \beta_0 + \log\left(1 + \exp(\beta_0 + \beta_1)\right) + \log\left(1 + \exp(\beta_0 - \beta_1)\right)\right] \\ &= \frac{1}{2}\left[-\frac{1}{2} - \left(-\frac{1}{2}\right) + \log\left(1 + \exp(-\frac{1}{2} + -\frac{1}{2})\right) + \log\left(1 + \exp(-\frac{1}{2} - -\frac{1}{2})\right)\right] \\ &= \frac{1}{2}\left[0 + \log\left(1 + \exp(-1)\right) + \log\left(1 + \exp(0)\right)\right] \\ &= \frac{1}{2}\log(2 + 2e^{-1}) \end{aligned}$$

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

Solution: Yes, the line $x_2 = 0$ separates the data in feature space.

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

Solution: No, since the features are linearly separable, we should be able to choose β so that cross-entropy is arbitrarily close to 0.