DS 100: Principles and Techniques of Data Science **Date: April 6, 2018**

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.

   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 test accuracy of 50% on all binary prediction tasks.

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

   (c) In logistic regression, predictor variables are continuous with values from 0 to 1.

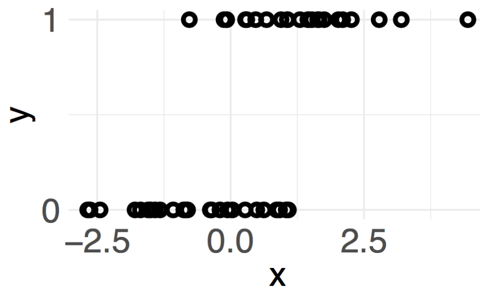
   Solution: False. There is no such constraint on the values that predictor variables might take.

   (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.

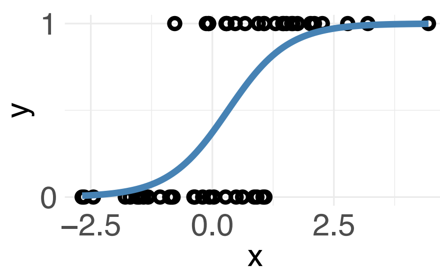
   Solution: False. The test accuracy could be much lower depending on the class imbalance in the test data.

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

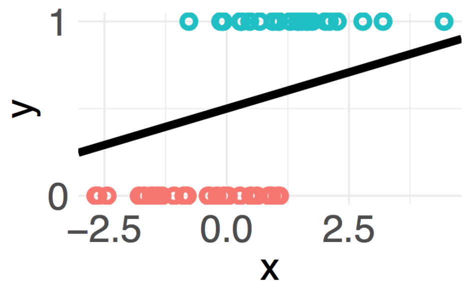
2. Based on the scatter plot of the data below, draw a reasonable approximation of the logistic regression probability estimates for $\mathbb{P}(Y = 1 \mid x)$



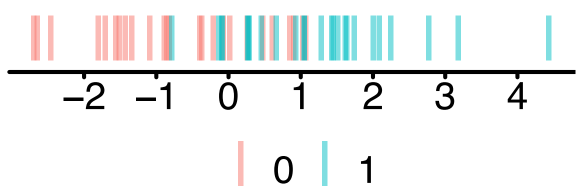
Solution:



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:

x	y
1	0
-1	1

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

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

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

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

Solution:

$$\begin{aligned}
\hat{\mathbb{P}}\left(Y = 1 \mid x = 0\right) &= \sigma(\phi^T(0)\hat{\theta}) \\
&= \sigma\left(\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}\right) \\
&= \sigma\left(1 \times -\frac{1}{2} + 0 \times -\frac{1}{2}\right) \\
&= \sigma\left(-\frac{1}{2}\right) \\
&= \frac{1}{1 + \exp(\frac{1}{2})}
\end{aligned}$$

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

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

where $\phi_i = \phi(x_i)$. Let $\theta = [\theta_0 \quad \theta_1]$. Explicitly write out the (empirical) loss for this data set in terms of θ_0 and θ_1.

Solution:

$$\phi_i^T \theta = \phi^T(x_i)\theta = \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \theta_0 + \theta_1 x_i$$

For the data point (1, 0):

$$y_i \phi_i^T \theta = 0 \times (\theta_0 + \theta_1 \times 1) = 0$$

$$-\phi_i^T \theta = -(\theta_0 + \theta_1 \times 1) = -\theta_0 - \theta_1$$

For the data point (-1, 1):

$$y_i \phi_i^T \theta = 1 \times (\theta_0 + \theta_1 \times -1) = \theta_0 - \theta_1$$

$$-\phi_i^T \theta = -(\theta_0 + \theta_1 \times -1) = -\theta_0 + \theta_1$$

We can then write the loss as:

$$L(\theta) = -\frac{1}{2}\left[(0 + \log \sigma(-\theta_0 - \theta_1)) + (\theta_0 - \theta_1 + \log \sigma(-\theta_0 + \theta_1))\right]$$
$$= -\frac{1}{2}\left[\theta_0 - \theta_1 + \log \sigma(-\theta_0 - \theta_1) + \log \sigma(-\theta_0 + \theta_1)\right]$$
$$= -\frac{1}{2}\left[\theta_0 - \theta_1 + \log\left(\frac{1}{1 + \exp(\theta_0 + \theta_1)}\right) + \log\left(\frac{1}{1 + \exp(\theta_0 - \theta_1)}\right)\right]$$

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

Solution:

$$\begin{aligned}
L(\hat{\theta}) &= -\frac{1}{2}\left[\theta_0 - \theta_1 + \log\left(\frac{1}{1+\exp(\theta_0+\theta_1)}\right) + \log\left(\frac{1}{1+\exp(\theta_0-\theta_1)}\right)\right] \\
&= -\frac{1}{2}\left[-\frac{1}{2} - \left(-\frac{1}{2}\right) + \log\left(\frac{1}{1+\exp((-\frac{1}{2})+(-\frac{1}{2}))}\right) + \log\left(\frac{1}{1+\exp((-\frac{1}{2})-(-\frac{1}{2}))}\right)\right] \\
&= -\frac{1}{2}\left[0 + \log\left(\frac{1}{1+\exp(-1)}\right) + \log\left(\frac{1}{1+\exp(0)}\right)\right] \\
&= \frac{1}{2}\log(2+2e^{-1})
\end{aligned}$$

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

Solution: Yes, the line $\phi_2 = 0$ 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 θ so that cross-entropy is arbitrarily close to 0.

5. (a) Show that $\sigma(-x) = 1 - \sigma(x)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$.

Solution:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

$$\sigma(-x) = \frac{e^{-x}}{e^{-x}+1} = \frac{e^{-x}+1-1}{1+e^{-x}} = \frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}} = 1 - \sigma(x)$$

(b) Show that the derivative of the sigmoid function can be written as:

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

Solution:

$$\sigma(x) = \frac{1}{1+e^{-x}} \implies \sigma(x)(1+e^{-x}) = 1$$

Taking the derivative with respect to x on both sides yields:

$$(1+e^{-x})\frac{d}{dx}\sigma(x) + \sigma(x)\frac{d}{dx}(1+e^{-x}) = (1+e^{-x})\frac{d}{dx}\sigma(x) + \sigma(x)(-e^{-x}) = 0$$

Solving for $\frac{d}{dx}\sigma(x)$:

$$\begin{aligned}\frac{d}{dx}\sigma(x) &= \sigma(x)\frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x)\sigma(-x) \\ &= \sigma(x)(1-\sigma(x))\end{aligned}$$