linear regression: outputs were in $\mathbb{R}$

logistic regression: CLASSIFICATION : predicting 1 or 0

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

Discussion #8

Name: surajrampure.com

Logistic Regression

prob $\in [0, 1]$

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. False: can use one-hot encoding

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

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

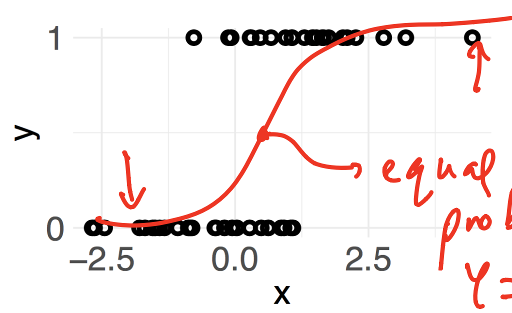
   (d) In a setting with extreme class imbalance in which 95% of the training data have the same label it is always possible to get at least 95% testing accuracy. False

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 $\mathbb{P}(Y = 1 \mid x)$

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

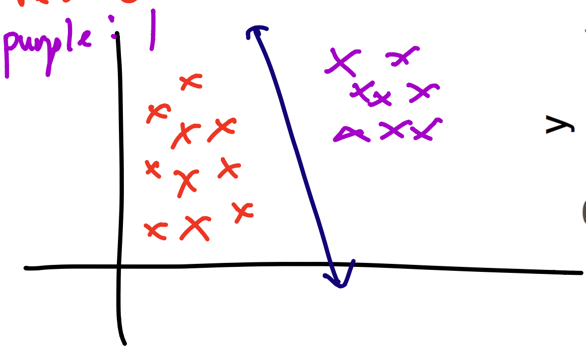
$$f_\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}$$

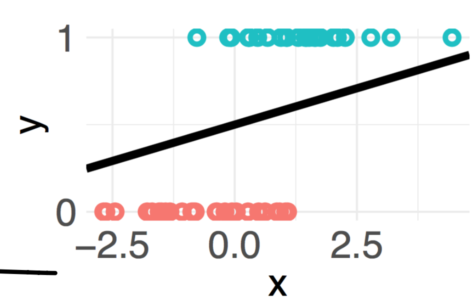


equal prob that $Y=1$ or $y=0$

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.

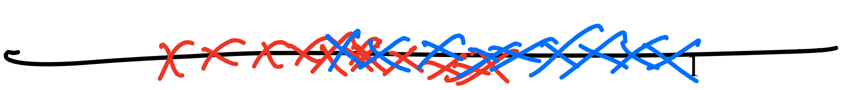
red: 0
purple: 1





not lin. separable

each pt:
value
class



$(1, 0)$

$(-1, 1)$

4. You have a classification data set:

$P(y=0 \mid x) = 1 - \sigma(\phi^T(x)\theta)$

x	y
1	0
-1	1

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

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

$\phi^T(x)\hat{\theta}$

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

$= \underbrace{-\frac{1}{2} - \frac{1}{2}x}$

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

$= \sigma(\phi^T(0)\theta) = \sigma(-\frac{1}{2}) = \boxed{\frac{1}{1+e^{1/2}}}$

$P(y=1)$

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

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

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.

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

$\phi(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}$

(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?

not necessarily, doesn't matter

$(1, 0)$

$y_i\phi_i^T\theta = 0$

$-\phi_i^T\theta = -[1 \quad 1]\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$

$= -(\theta_0 + \theta_1)$

$(-1, 1)$

$y_i\phi_i^T\theta$

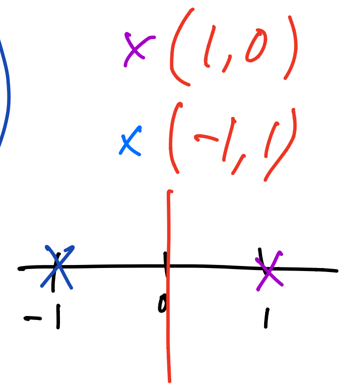
$= 1\,[1 \quad -1]\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$

$= \theta_0 - \theta_1$

$-\phi_i^T\theta = \theta_1 - \theta_0$

$$l(\theta) = -\frac{1}{2}\Big(0 + \log\big(\sigma(-\theta_0-\theta_1)\big) + \theta_0 - \theta_1 + \log\big(\sigma(\theta_1-\theta_0)\big)\Big)$$

$$= -\frac{1}{2}\left(\log\left(\frac{1}{1+e^{\theta_0+\theta_1}}\right) + \theta_0 - \theta_1 + \log\left(\frac{1}{1+e^{\theta_0-\theta_1}}\right)\right)$$



c) substitute $\theta_0 = \theta_1 = -\frac{1}{2}$