DS 100: Principles and Techniques of Data Science **Date: July 31, 2019**

Discussion #11

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.

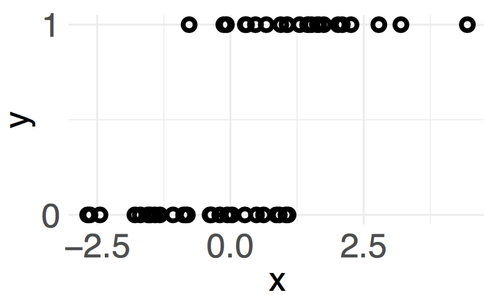
   (c) For a logistic regression model, all features are continuous, with values from 0 to 1.

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

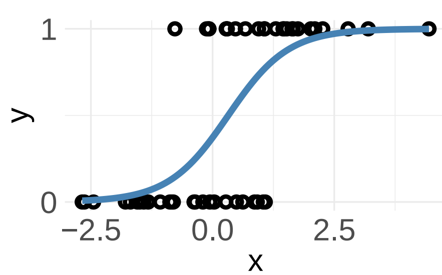
   (d) In a setting with extreme class imbalance in which 95% of the training data have the same label, it is always possible to get at least 95% testing accuracy.

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

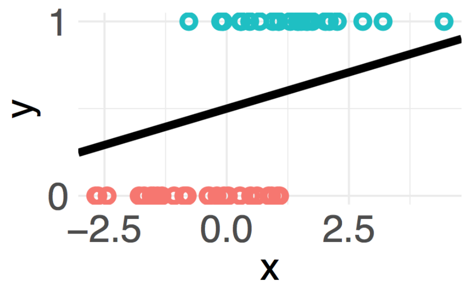
2. The next question refers to a binary classification problem with a single feature x. Based on the scatter plot of the data below, draw a reasonable approximation of the logistic regression probability estimates for $\mathbb{P}(Y = 1|x)$.



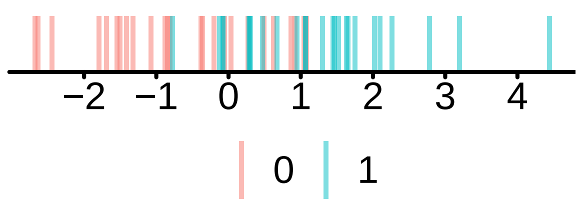
Solution:



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.

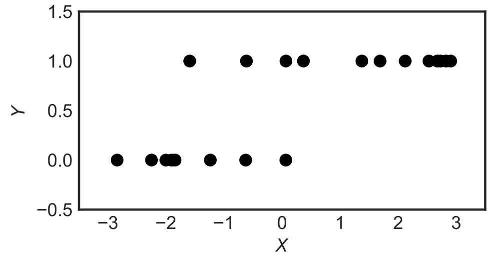


Solution: The scatter plot of x against y isn't the graph you should be looking at. The more salient plot would be the $d = 1$ representation of the features colored by class labels.



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. Suppose you are given the following dataset $\{(x_i, y_i)\}_{i=1}^n$ consisting of x and y pairs where the covariate $x_i \in \mathbb{R}$ and the response $y_i \in \{0, 1\}$.



Given this data, the value $\mathbb{P}(Y = 1|x = -1)$ is likely closest to:

☐ 0.95 √ **0.50** ☐ 0.05 ☐ -0.95

5. You have a classification data set, where x is some value and y is the label for that value:

x	y
2	1
3	0
0	1
1	0

Suppose that we're using a logistic regression model to predict the probability that $Y = 1$ given x:

$$\mathbb{P}(Y = 1|x) = \sigma(\mathbf{X}\theta)$$

(a) Suppose that $\mathbf{X} = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}^T$ and our model parameters are $\theta^* = \begin{bmatrix} 1 & 0 & -2 \end{bmatrix}^T$. For the following parts, leave your answer as an expression (do not numerically evaluate $\log$, e, π, etc).

i. Compute $\hat{\mathbb{P}}(y = 1|x = 0)$.

Solution: $\frac{1}{1+\exp(-1)}$

ii. What is the loss for this single prediction $\hat{\mathbb{P}}(y = 1|x = 0)$, assuming we are using Cross Entropy as our loss function (or equivalently that we are using the cross entropy as our loss function)?

Solution: $\log(1 + \exp(-1))$

6. Suppose we train a binary classifier on some dataset. Suppose y is the set of true labels, and $\hat{y}$ is the set of predicted labels.

y	0	0	0	0	0	1	1	1	1	1
$\hat{y}$	0	1	1	1	1	1	1	0	0	0

Determine each of the following quantities.

(a) The number of true positives

Solution: 2

(b) The number of false negatives

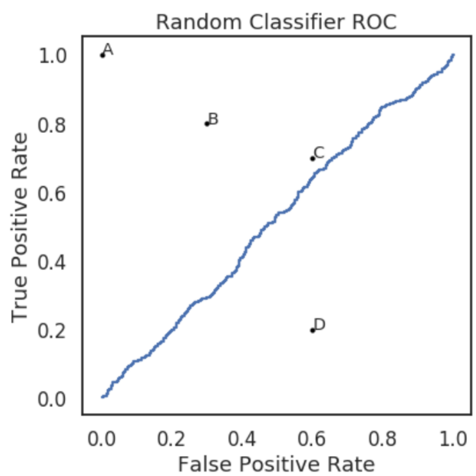
Solution: 3

(c) The precision of our classifier. Write your answer as a simplified fraction.

Solution: $\frac{2}{2+4} = \frac{1}{3}$

ROC Curves

7. State whether the following claims are true or false. If false, provide a reason or correction.



(a) Point A (0, 1) represents our ideal classifier.

Solution: True. This gets everything perfect. This classifier commits no false positive errors and gets all true positives. In general, we want points as far northwest as possible because it means that we have a higher TP rate, lower FP rate, or both.

(b) Point C performs well compared to a classifier that guesses each class randomly.

> **Solution:** False. This classifier performs just about the same as the Random Classifier. We want a classifier that can predict better than one that chooses randomly!

(c) The classifier at Point B performs better than the one at Point D.

> **Solution:** True. We never want a classifier that falls under the $y = x$ line because it performs worse than random guessing. The triangle under this line is therefore generally empty in ROC graphs.