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

Discussion #9

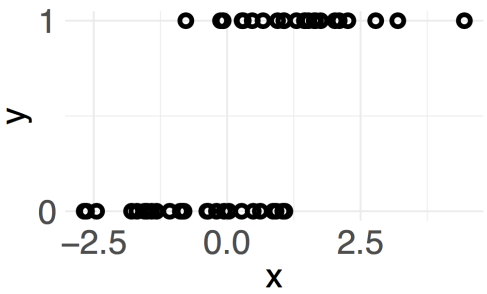
*Name:*

# Logistic Regression

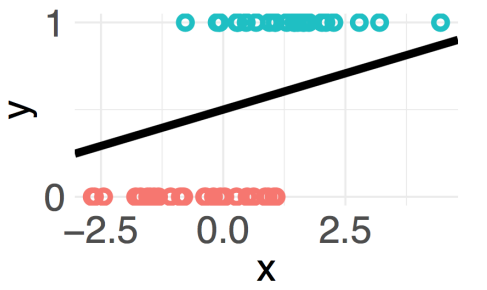
1. State whether the following claims are true or false. If false, provide a reason or correction.
   (a) A binary or multi-class classification technique should be used whenever there are categorical features.
   (b) A classifier that always predicts 0 has test accuracy of 50% on all binary prediction tasks.
   (c) In logistic regression, predictor variables 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:

| x | y |
|---|---|
| 1 | 0 |
| -1 | 1 |

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

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

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

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

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

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

where $\phi_i = \phi(x_i)$. Let $\theta = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix}$. Explicitly write out the (empirical) loss for this data set in terms of $\theta_0$ and $\theta_1$.

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

(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?

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

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

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

# Bootstrap Review

We take an i.i.d. random sample of size 9 from a population. We write all the values on pieces of paper and stick them in a box:

| 1 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|

The numbers in the box have the following summary statistics:

| Statistic | Sum | Sum of Squares | Mean | Median |
|---|---|---|---|---|
| Value | 27 | 93 | 3 | 3 |

6. For each of the following, answer the following questions: Is this value calculable from the information given? If so, either calculate it by hand or describe how you would calculate this value. If not, then suggest an estimate for the quantity.
   (a) The expected value of a single draw from the box.
   (b) The expected value of the average of nine draws with replacement from this box
   (c) The exact variance of the tickets in the box
   (d) The exact variance of a single draw from the box
   (e) The exact variance of the average of nine draws with replacement from the box
   (f) The exact variance of the average of nine draws with replacement from the population

7. Describe a bootstrap procedure to estimate the variance of the average of nine draws with replacement from the population.

8. What are the sources of error in the bootstrap procedure?

9. Which of the following could be valid bootstrap resamples? Provide reasons for the ones that are not.
   (a) 1, 2, 2, 3, 3, 4, 4, 5, 6
   (b) 1, 2, 2, 2, 3, 3, 3, 4, 4, 5
   (c) 1, 1, 1, 1, 1, 1, 1, 1, 1
   (d) 2, 2, 3, 3, 3, 4, 4, 4
   (e) 1, 2, 3, 3, 3, 4, 4, 5, 5
10. What are some assumptions we are making when performing the bootstrap?
11. You generate 10 bootstrap resamples (you would normally take many more). They are sorted and printed below:

```
[1, 2, 2, 2, 4, 4, 4, 4, 5] [1, 2, 3, 3, 3, 3, 3, 4, 4]
[1, 2, 2, 3, 3, 3, 4, 4, 5] [1, 1, 2, 3, 4, 4, 4, 5, 5]
[2, 3, 3, 3, 4, 4, 4, 5, 5] [2, 3, 3, 3, 3, 3, 3, 4, 4]
[1, 1, 1, 1, 2, 2, 3, 4, 5] [2, 2, 3, 4, 4, 4, 4, 4, 5]
[1, 2, 2, 3, 3, 3, 4, 4, 4] [1, 2, 2, 2, 3, 3, 3, 4, 5]
```

    Construct a 60% confidence interval for the population $40^{th}$ percentile of the population.

# Properties of the Bootstrap

In the bootstrap, we have a sample $\{X_1, \ldots, X_n\}$ from which we sample with replacement $n$ times to obtain $\left\{\widetilde{X}_1, \ldots, \widetilde{X}_n\right\}$. Most likely, some of the values $\{X_1, \ldots, X_n\}$ will show up more than once.

12. For a really big sample, how likely are we to observe a data point $X_1$ in a particular bootstrap sample? Write down a guess below.
13. Let's see how we would answer this question analytically. First, pick a fixed sample size $n$. What is the probability that $X_1$ appears on the second draw of a bootstrap sample?
14. What is the probability that $X_1$ appears in a particular bootstrap sample?
15. What is the limit of this probability as $n$ approaches $\infty$?

    Hint: Define $y = \mathbb{P}\left(X_1 \text{ doesn't appear in the bootstrap sample}\right)$. Then take the natural log of both sides.
16. Approximately what is the limit above equal to numerically?
17. How many times does a data point $X_1$ show up on average in the bootstrap sample?