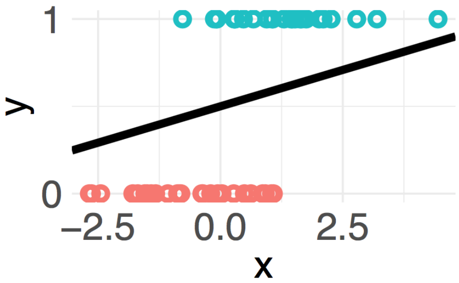
DS 100/200: Principles and Techniques of Data Science **Date: November 20, 2019**

Discussion #13

Name:

Logistic Regression

1. (a) 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. Note: this question refers to a binary classification problem with a single feature x.



(b) Suppose you use gradient descent to train a logistic regression model on two design matrices $\mathbb{X}_a$ and $\mathbb{X}_b$ and use some arbitrary threshold T. After training, you find that the training accuracy for $\mathbb{X}_a$ is 100% and the training accuracy for $\mathbb{X}_b$ is 98%. What can you say about whether the data is linearly separable for the two design matrices?

2. Suppose we are given the following dataset, with two features ($\mathbb{X}_{:,1}$ and $\mathbb{X}_{:,2}$) and one response variable (y). *(Note, this is the same dataset from Discussion 11, just formatted slightly differently.)*

$\mathbb{X}_{:,1}$	$\mathbb{X}_{:,2}$	y
1	1	0
1	-1	1

Here, **x** corresponds to a single row of our data matrix, not including the y column. For instance, $\mathbf{x}_1 = [1 \quad 1]^T$

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

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

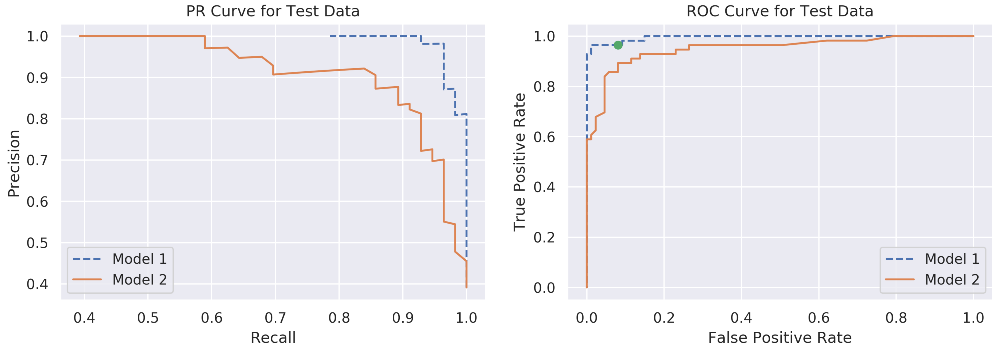
Your algorithm returns $\hat{\beta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$

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

(b) Recall from Discussion 11 that the empirical risk for $\hat{\beta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$ and the two observations above is $\frac{1}{2}\log(2 + 2e^{-1})$. Does this fitted model minimize cross-entropy loss?

Evaluating Classifiers

3. Consider the PRC curves (left) and ROC curves (right) for two models.

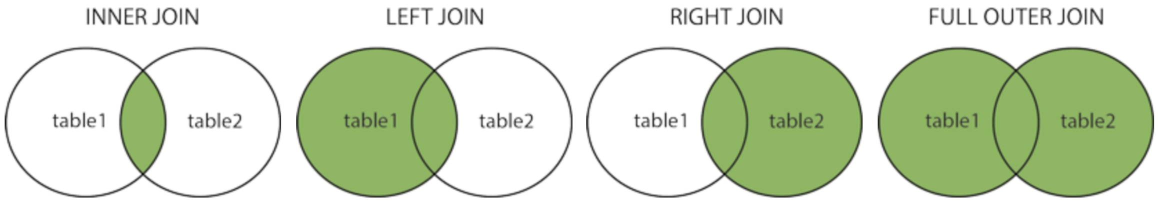


(a) Which model is better according to the ROC curves? Why?

(b) Which model is better according to the PRC curves? Why?

(c) For the ROC curve, suppose that the green dot represents the threshold T = 0.2. Should we use a higher or lower threshold than 0.2?

SQL



Note: You do not always have to use the JOIN keyword to join sql tables. The following are equivalent:

```
SELECT column1, column2
FROM table1, table2
WHERE table1.id = table2.id;

SELECT column1, column2
FROM table1 JOIN table2
ON table1.id = table2.id;
```

4. Describe which records are returned from each type of join in the figure above. How does a cross join relate to these types of joins?

5. Consider the following real estate schema:

```
Homes(home_id int, city text, bedrooms int, bathrooms int,
area int)
Transactions(home_id int, buyer_id int, seller_id int,
transaction_date date, sale_price int)
```

```
Buyers(buyer_id int, name text)
Sellers(seller_id int, name text)
```

Fill in the blanks in the SQL query to find the id and selling price for each home in Berkeley. If the home has not been sold yet, **the price should be NULL**.

```
SELECT ____________________________
FROM _________________________________
__________________ JOIN __________________
ON ___________________________
WHERE ___________________________;
```

6. Examine this schema for these two tables:

```
CREATE TABLE owners (            CREATE TABLE cats (
    ownerid integer,                 catid integer,
    name text,                       owner integer,
    age integer,                     name text,
    PRIMARY KEY (userid)             breed text,
);                                   age integer,
                                     PRIMARY KEY (catid),
                                     FOREIGN KEY (owner) REFERENCES owners
                                 );
```

(a) Write a SQL query to get a random sample of 5 random Maine Coons (a cat breed) with a name that starts with the letter A.

(b) Write a SQL query to create an almost identical table as cats, except with an additional column ’Nickname’ that has the value ’Kitten’ for cats less than or equal to the age of 1, ’Catto’ for cats between 1 and 15, and ’Wise One’ for cats older than or equal to 15.

(c) Write a SQL query to select all rows from the `cats` table that have cats of the top 5 most popular cat breeds.