Classification, Logistic Regression, Probability

Data 100 Review Discussion

Slides by Suraj Rampure

Fall 2018

Regression vs. Classification

Regression is the problem of creating a model that takes in a point and outputs a real number. We've seen regression in the form of Ordinary Least Squares, Ridge Regression, and LASSO Regression.

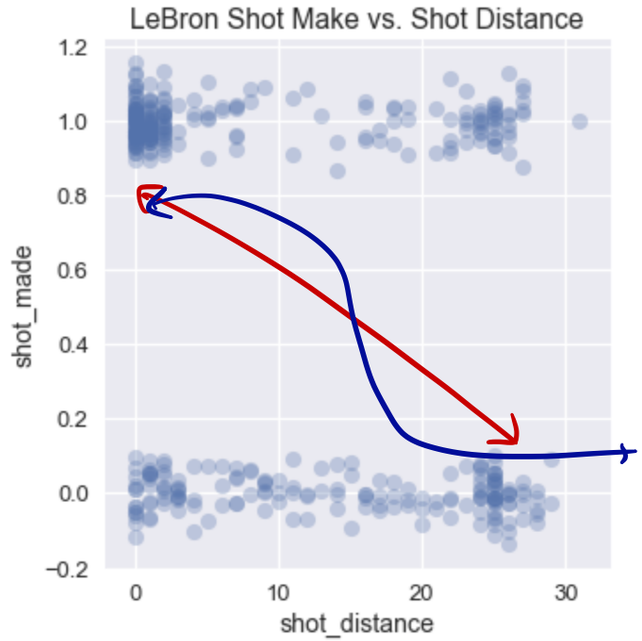
On the other hand, **classification** is the problem of creating a model that takes in a point and outputs a discrete **label**. In this course, we explored Logistic Regression (which, despite the name, is a classification technique), and previous courses, you may have seen k-Nearest Neighbors and k-Means Clustering, all of which are classification techniques.

A very basic example:

- Regression would allow us to predict a student's final exam grade, given their grades on the midterm and homeworks.
- Classification would allow us to predict whether or not that student will pass the exam.

Example (from textbook): Suppose we have LeBron's shot data for a particular season. Specifically, we have the distance from which the shot was taken, and whether or not it went in.

We want to build a model that will allow us to predict whether or not a new shot will go in.



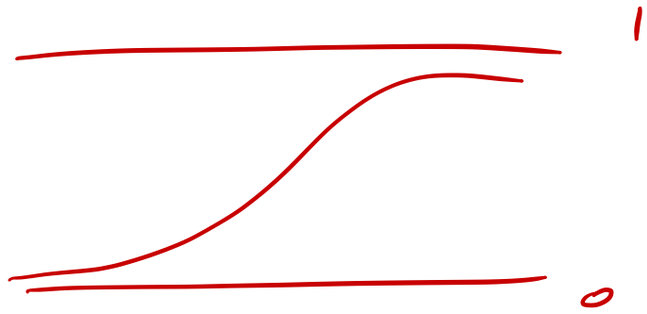
Sure, we *can* fit a standard linear regression model to this, and interpret the output as the *probability that the shot will go in*. For example, we can say if the predicted value is over 0.5, he will make the shot. **However, values aren't restricted to** $[0, 1]$**!**

Probability Recap, Sigmoid Function

A probability density function $f(x)$ is valid iff it satisfies the following conditions:

- $0 \leq f(x) \leq 1$, for all $x \in \mathbb{X}$
- $\sum_{x \in \mathbb{X}} f(x) = 1$

We need some function that maps $\mathbb{R} \rightarrow [0, 1]$. Our choice is $\sigma(x)$:



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

This is known as the **sigmoid** function, and it satisfies the following property:

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

With standard linear regression, we find the θ that minimizes

$$\min_{\theta} \frac{1}{n} ||y - \phi\theta||_2^2$$

obs predicted

Then, to make a new prediction, we evaluate $\phi(x)^T\theta$, where $\phi(x)$ is the feature vector representing our test point x.

However, with **logistic regression**, we instead model using the following (assuming our classification is binary):

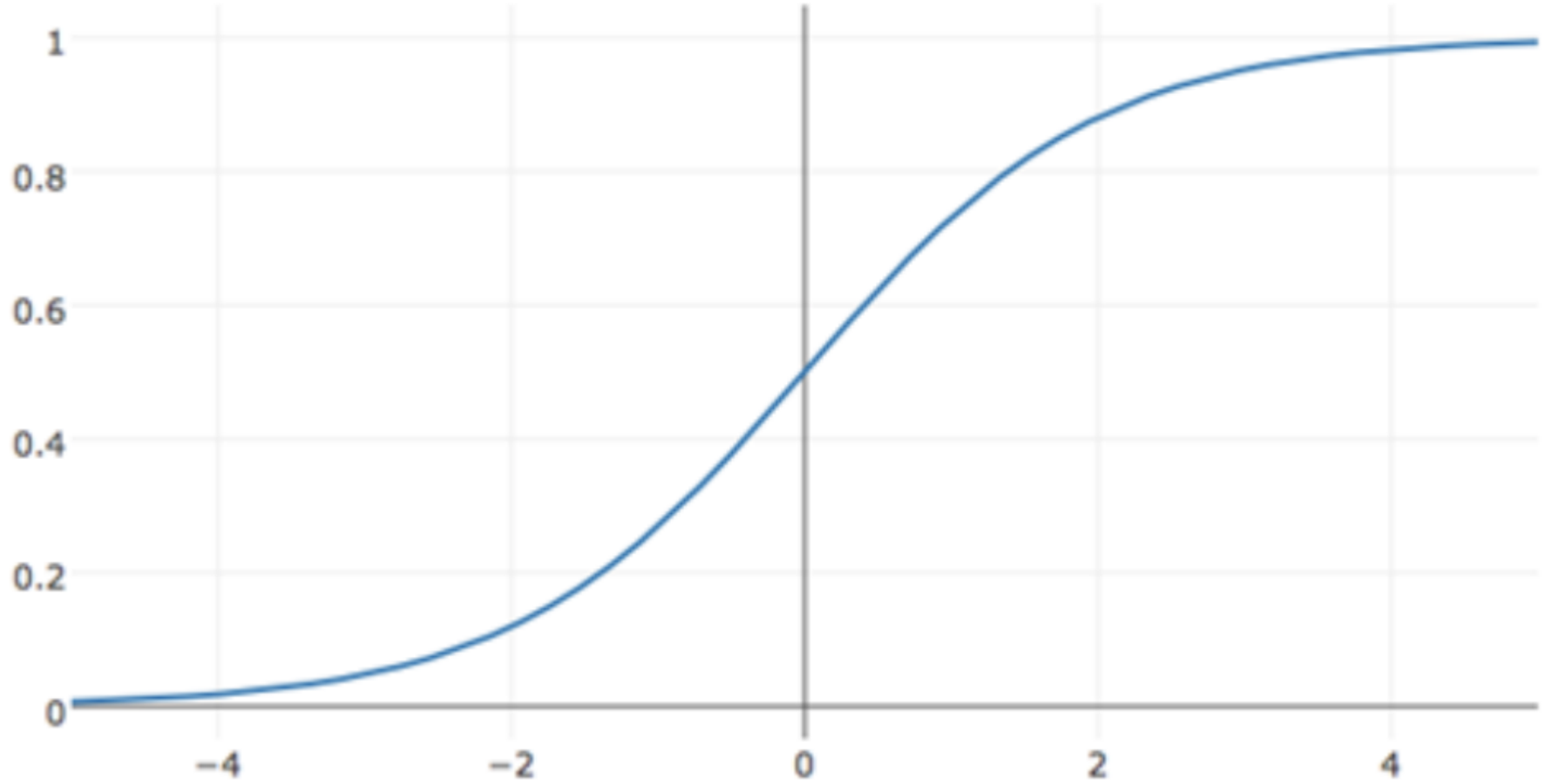
$$\mathbb{P}_\theta(Y = 1|x) = \sigma(\phi(x)^T\theta) = \frac{1}{1 + \exp(-\phi(x)^T\theta)}$$

learning

feature vec

This is the probability that our test point x belongs to class 1 (as opposed to class 0). We decide the *cutoff*, or decision boundary.

Output of $\sigma(x)$



Loss Function for Logistic Regression

The loss function we use for logistic regression is what is known as **cross entropy loss**. This is inspired by **KL Divergence**, which is defined between two probability distributions as follows:

$$D(p||q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

observed

predicted

Average cross entropy loss is of the form

$$L(\theta) = -\frac{1}{n} \sum_{i=1}^{n} (y_i \phi(x_i)^T \theta + \log(\sigma(-\phi(x_i)^T \theta)))$$

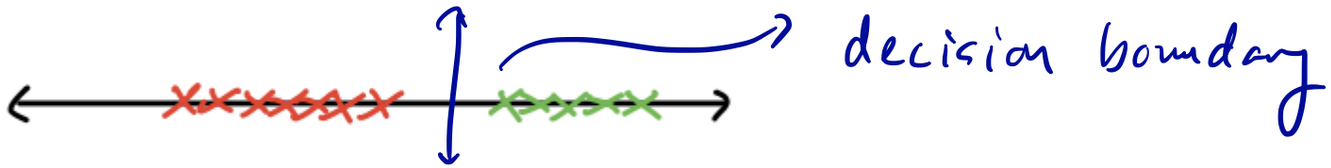
This cannot be determined analytically, unlike the solution to OLS. We must use a numerical method, such as gradient descent, to determine θ^*.

See lecture 17 slides for the derivation.

Linear Separability

The goal of linear regression is to model the probability of a point belonging to a class. We only do this when there is some level of uncertainty, i.e. overlap, in our training set.

In some cases, we are able to draw a *linear decision boundary* to separate our data.

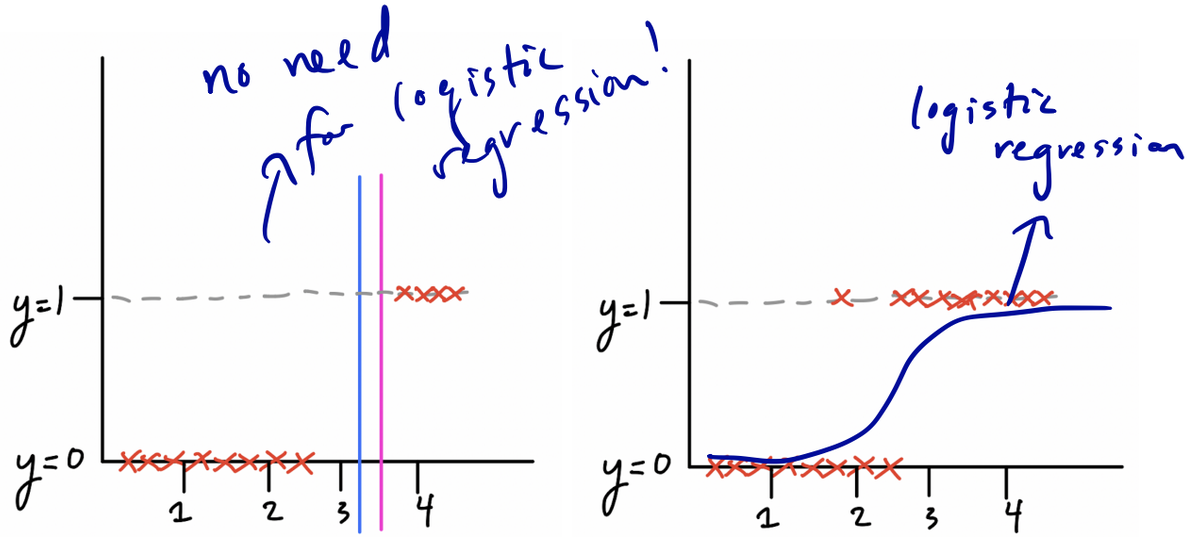


Above, we see that our data is **linearly separable**. We can draw a line (infinitely many lines, in fact) between the clusters of red dots and green dots.

This is not the case in the second example.

here, our data is still one dimensional!

Again, let's look at data in 1D, but plotted in 2D (one dimension is the value of our variable, the second dimension is the label, 0 or 1 --- this is equivalent to the drawings on the previous slide).



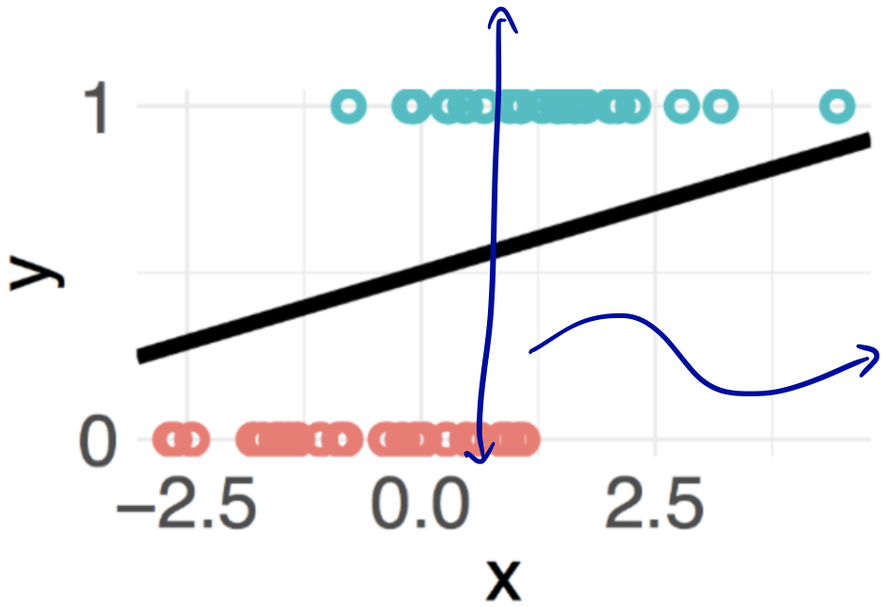
On the left, we have an example of linearly separable data. In the blue and purple are two possible hyperplanes that can separate our data. There would be no use in using logistic regression here.

However, on the right, we have non-linearly separable data. In this case we would use a tool like logistic regression to model probabilities.

IMPORTANT: Linear Separability Depends on the Dimension of the Data!

A set of d-dimensional points is linearly separable iff we can draw a degree $d - 1$ hyperplane to separate the points.

class label is represented by color and y-axis



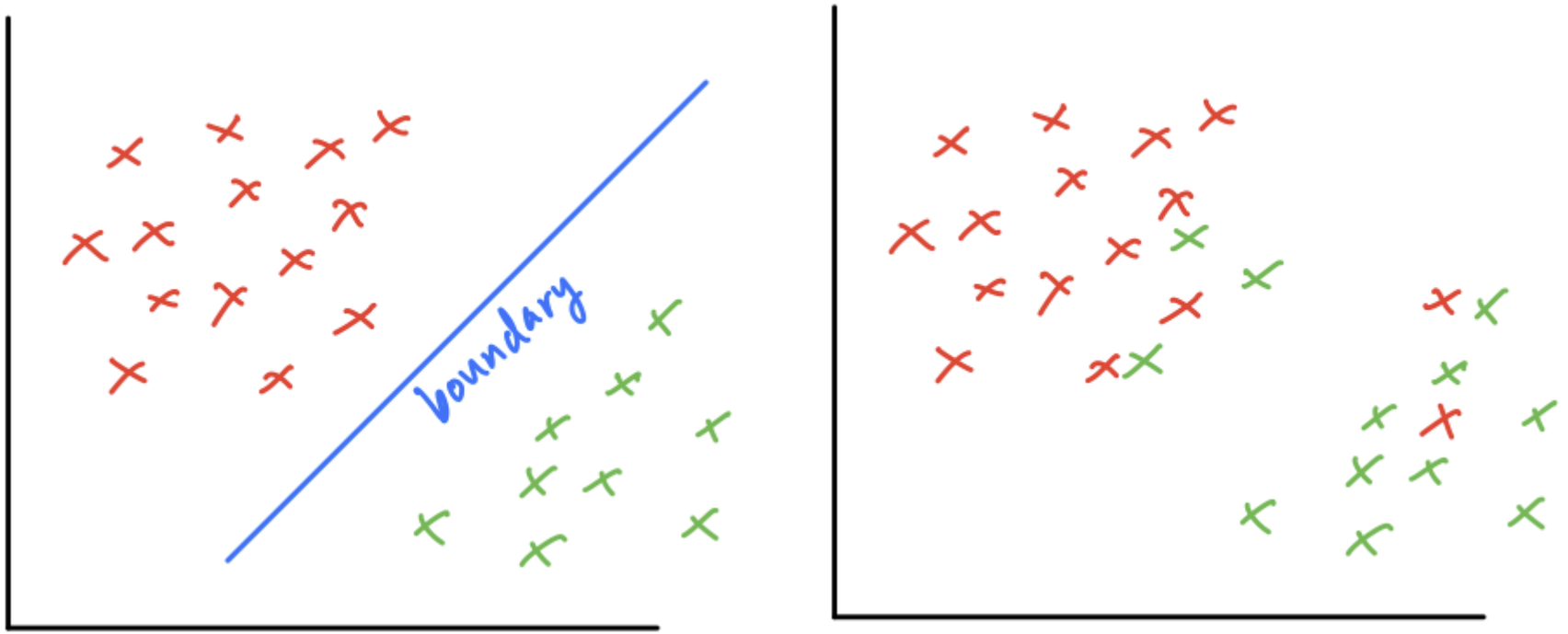
$x = a$

$y = ax + b$

no such decision boundary

This data **is not** linearly separable. This problem (from Disc 8) is intentionally misleading; the points are in 1D, however the class labels are represented in two ways (y axis and color). We cannot draw a degree 0 line (i.e. something of the form $x = a$) to separate this data.

Linear Separability in Two Dimensions



Here, we have examples of both linearly separable and non-linearly separable data in two dimensions. Here, our data is truly two dimensional, as our feature space has two components an x and a y. The class is represented by the color.

Evaluating the Effectiveness of a Classifier

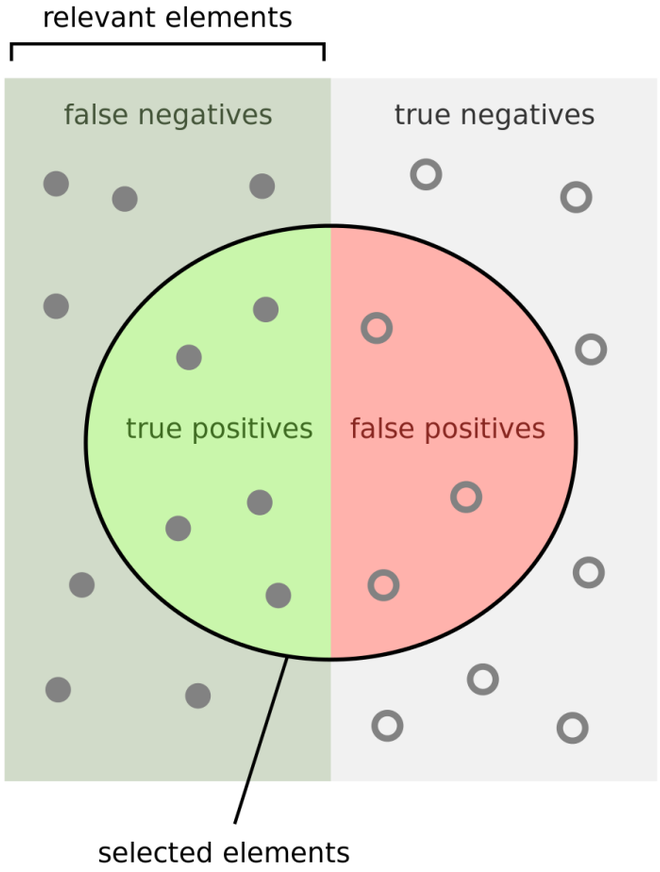
Suppose we train a binary classifer, and suppose `y` represents actual values, and `y_pred` represents predicted values.

Recall (no pun intended) the following definitions:

- True Positives: `TP = np.count_nonzero((y == y_pred) & (y_pred == 1))`
- True Negatives: `TN = np.count_nonzero((y == y_pred) & (y_pred == 0))`
- False Positives: `FP = np.count_nonzero((y != y_pred) & (y_pred == 1))`
- False Negatives: `FN = np.count_nonzero((y != y_pred) & (y_pred == 0))`

Then, we have the following definitions:

- Precision = $\frac{TP}{TP+FP}$ → of all predictions that were positive, what proportion were actually positive
- Recall = $\frac{TP}{TP+FN}$ ↪ proportion of positives that were successfully identified



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Example

Suppose you create a classifier to predict whether or not an image contains a picture of a goat. You test it on 23 images.

- There were 12 true images of goats. Your classifier predicted 9 of them to be goats, and 3 to not be a goat.
- There were 11 images that did not contain goats. Your classifier predicted 3 of them to be goats, and the remaining 8 to not be goats.

Determine the precision and recall of your goat classifier.

More Practice Problems

The following are from past exam problems.

Practice: Classification (T/F)

True/False: In logistic regression, predictor variables (x) are continuous, with values in the range $[0, 1]$.

no constraint on x

True/False: In two-class ~~logistic regression~~ binary classification, the response variable (y) is continuous, with values in the range $[0, 1]$.

either 0, or 1 (not continuous)

True/False: In logistic regression, we calculate the weights θ^* as $\theta^* = (\phi^T \phi)^{-1} \phi^T y$ and then fit responses as $y_i = \sigma(\phi_i(x)^T \theta)$.

have to optimize cross-entropy loss

Practice: Logistic Regression (T/F)

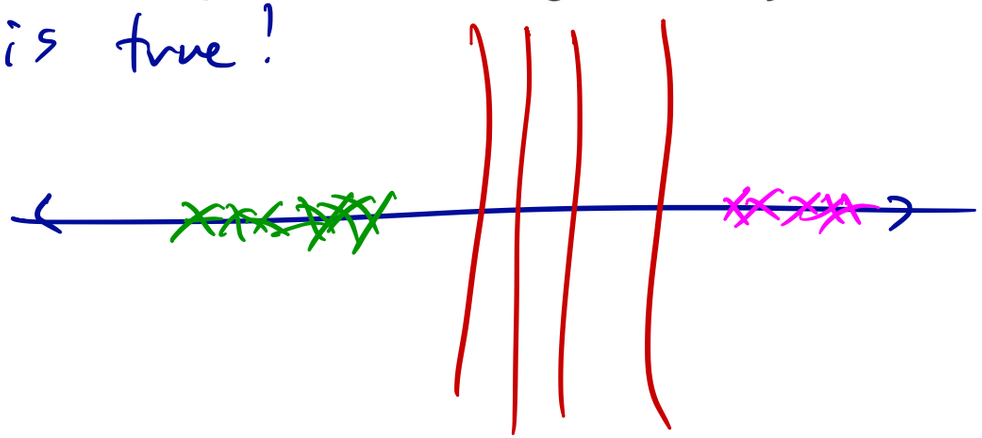
True/False: If no regularization is used and the training data is linearly separable, the parameters will tend towards positive or negative infinity.

infinitely many θs

True/False: L_1 reguarlization can help us select a subset of the features that are important.

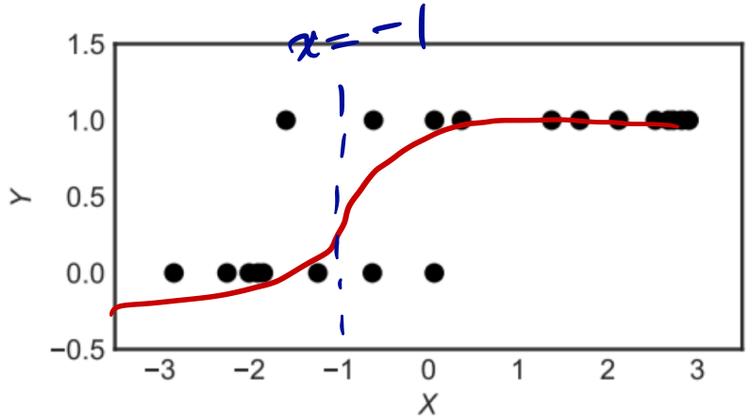
True/False: After regularization, we expect the training accuracy to increase and the test accuracy to decrease.

opposite is true!



Practice: Interpreting Logistic Regression

[2 Pts] 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 $y_i \in \{0, 1\}$.

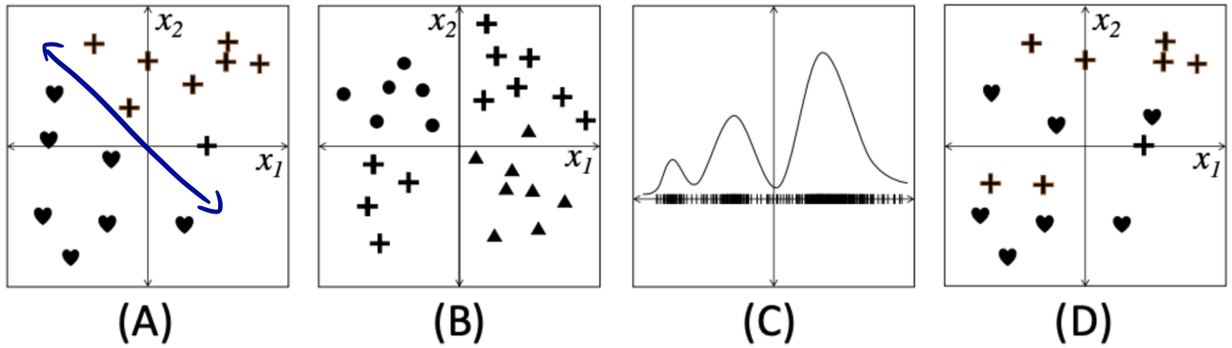


Given this data, the value $\mathbf{P}(Y = 1 \mid x = -1)$ is likely closest to:

○ 0.95 ○ 0.50 ○ 0.05 ○ ~~-0.95~~

prob. can't be negative

Practice: Separability



(1) Which of the above plots represents a **linearly separable binary classification** task?

● (A) ○ (B) ○ (C) ○ (D)

(2) Which of the above plots represents a **binary classification** task that is **not linearly separable**?

○ (A) ○ (B) ○ (C) ● (D)

(3) Which of the above plots represents a **multi-class classification task**?

○ (A) ● (B) ○ (C) ○ (D)

Practice: Probability

There are 32 participants in a randomizied clinical trial: 8 are male and 24 are female. 16 are assigned to treatment and the others are put into the control group. What is the probability that none of the men are in the treatment group if

a) The treatment was assigned using stratified random sampling, grouping by gender?
b) The treatment was assigned using simple random sampling?
c) The treatment was assigned using cluster random sampling of 2 groups of 8 using clusters as defined below?

Cluster	Male	Female
A	0	8
B	3	5
C	5	3
D	0	8

a) 0

b) $\frac{\binom{24}{16}}{\binom{32}{16}} = \frac{24!}{16!8!} \cdot \frac{16!16!}{32!}$

c) $\binom{4}{2}$ ways to select clusters

-> $\frac{1}{6}$

Sources

- Discussion 8
- Spring 2018 Final
- Fall 2017 Practice Final
- Prof. Hug's slides (slide 30)
- Wikipedia article on precision and recall
- http://textbook.ds100.org