Lab -- Separators

Instructions

This lab is focused on getting you started in 6.036 and introducing you to some of the kinds of problems we will be thinking about in this class.

For this first week only, the lab includes introductions by an instructor at the start of the lab section.

In 6.036 a large part of the learning happens by discussing lab questions with partners. You will be assigned to a breakout room after introductions. Once you are in your breakout rooms, please complete this group self-partnering question with other people in your breakout room.

Group information

Create Your Group

Once you and your partner(s) are in a breakout room, you need to create a group.

Instructions:

(1) **One** of you should click here to create a group.

(2) Everyone else should then enter the given new group name (including trailing numbers) into the box below (and press enter when done).
Name of group to join:

Reload this page anytime to refresh your group status (and to get a delete button, if you want to remove yourself from a group you've joined).

Groups are limited to 3 persons max each, so that checkoff discussions can be inclusive.

You are not currently in any group

This 'Group Information' box will be similar to what you will see at the top of future labs as well.

Note: Write all your answers down somewhere you can share them with a staff member. Be prepared to discuss your answers!

Checkoff 1:
Check-in with a staff member, to confirm your attendance and that the setup is working for you.

Ask for Help | Ask for Checkoff

Now let's get started!

1) Finding Good Models

Machine Learning is about using data that represent examples of a phenomenon in the world (such as how individual humans respond to a medical treatment), in order to make predictions about new examples (how different humans will respond to the same treatment).

In order to make predictions, however we must make the assumption that there is an underlying structure in the data and that this structure extends to new, unseen data as well. But how can we know whether a prediction rule we have identified that agrees with previous data, will be accurate on future data?

In this lab, we will practice making predictions from data and explore conceptual and more technical strategies for evaluating the effectiveness of prediction rules.

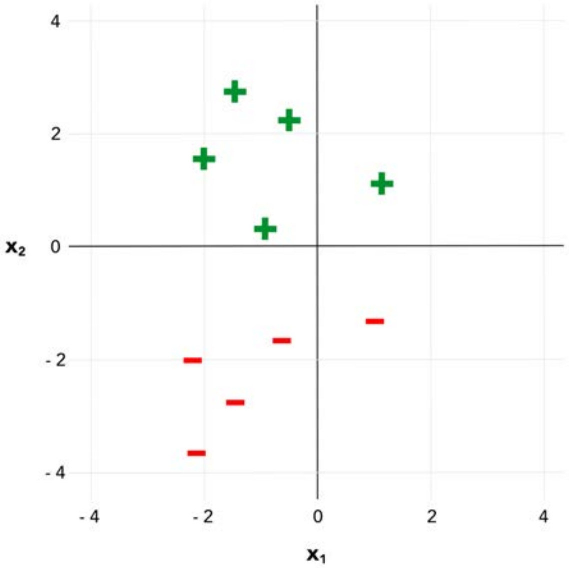
1.1) Power in Numbers

Acme Apparel is a company that has data from previous catalogs indicating which past customers ordered something from the catalog and they are trying to decide which customers to mail catalogs to (yes, that’s still a thing!) this holiday season.

Each customer is represented in terms of two features (they could represent anything, such as age, address, previous buying history, etc.), so we can plot a point in 2D coordinates representing each person and label it according to whether they bought something last time: positive (shown as a green plus sign) and negative (shown as a red minus sign). We want to predict whether a brand new customer, characterized in terms of these same two features, will make a purchase.

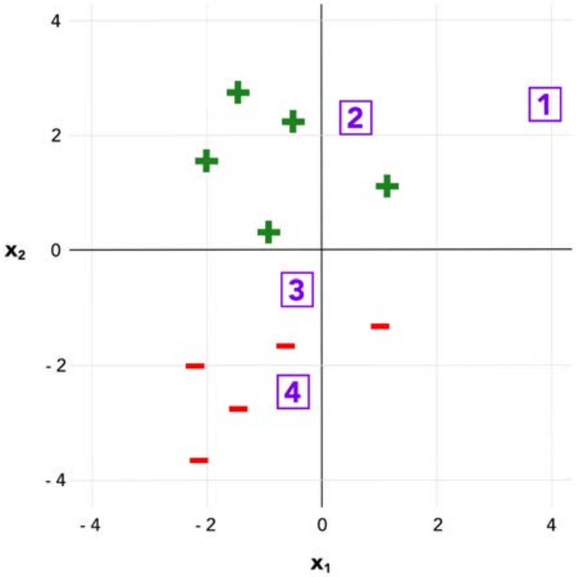
Our historical data looks like this:

Note: Here we use x_1 and x_2 as axes instead of the typical x, y axis because as you'll see later, we use 'y' to denote something else!



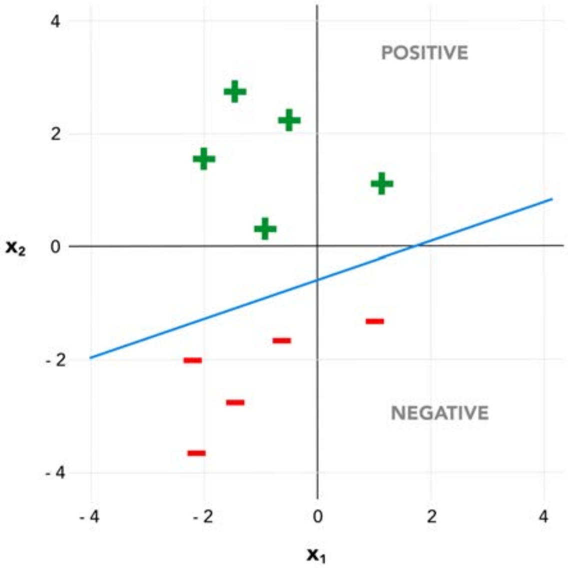
1.1.1)

Suppose there are 4 people who have not recieved a catalog. We collect information from them and represent them in the image below using the four points marked as purple numbers. Which people do you predict will buy something? Why or why not?



1.1.2)

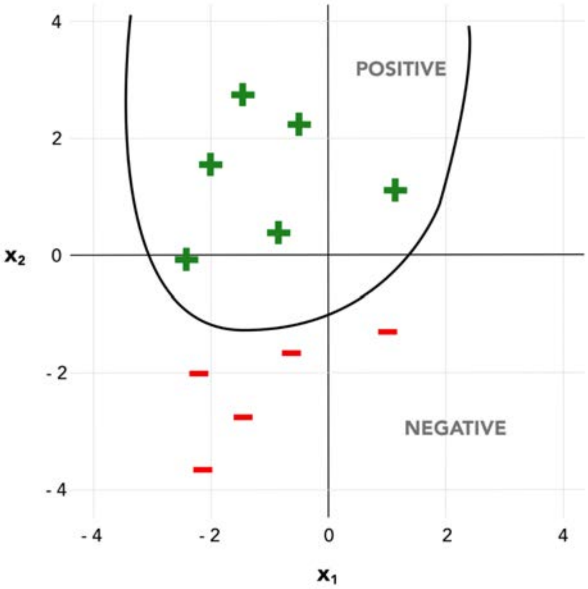
Acme's marketing department studied the data and came up with a hypothesis about which potential customers will make a purchase, which is illustrated by the blue line below: points on one side are positive whereas points on the other side are negative.



How do you think this hypothesis will perform on future customers?

1.1.3)

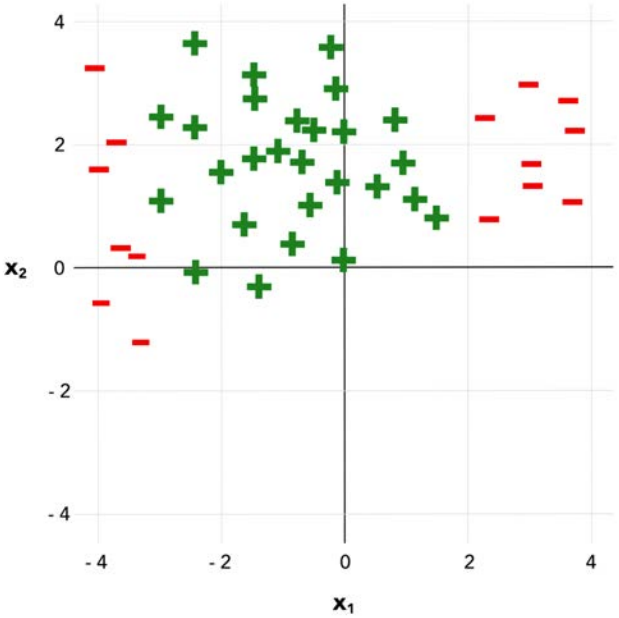
The Acme sales department comes up with a different hypothesis, shown in black below.



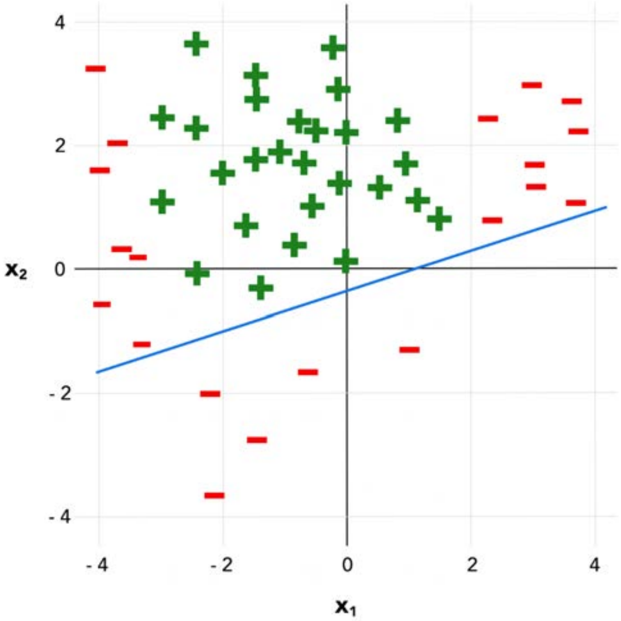
How do you think the black hypothesis will perform? If you had to choose between the blue and the black hypotheses, which would you choose to use for future predictions? Why?

1.1.4)

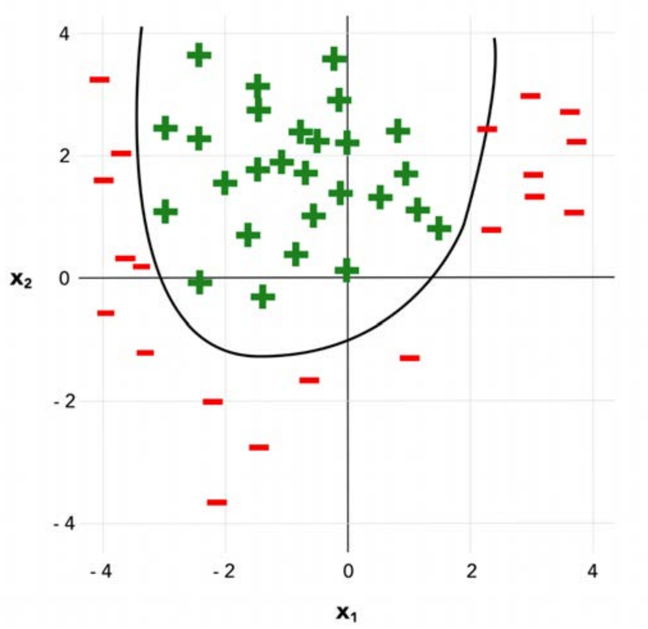
Meanwhile, at Acme Apparel, the marketing department won the argument and so they mail out catalogs based on the first, simple (blue) linear hypotheses. After the holiday rush, they make a plot of their new data and see this:



Marketing department (linear) hypothesis with both old and new data



Sales department (non-linear) hypothesis with both old and new data



Did using the simple,linear hypothesis to mail out catalogs result in a good outcome? Why or why not?

1.1.5)

Which hypothesis should Acme use next holiday season? What might they have done differently to avoid mailing so many catalogs to uninterested customers?

1.1.6)

The engineering department decides they want to try their hand at developing a new hypothesis. The department collects both new and old data from sales and marketing and divides it at random into two separate data sets, called A and B. They use data set A to develop a new hypothesis *and don't use data set B to develop the hypothesis*.

Data set A (training) has 10,000 data points. Data set B (test) has 5,000 data points.

The engineers' hypothesis, when applied to data set A (training), makes wrong predictions on 5% of the data. When that same hypothesis is applied to data set B (test set), it makes wrong predictions 10% of the time. What would you estimate the error rate of applying this hypothesis to brand new data (neither A nor B) to be?

1.1.7)

For the same setup as described above, how well do you think your hypothesis will perform in the real world if your data set B had 5 data points, versus if it had 5,000 data points?

1.1.8)

For the same setup as described above, how well do you think your hypothesis will perform in the real world if your data set A had 10 points, versus if it had 10,000 points? Why? Note that the same setup as described above applies here.

1.2) Comparing Apples to Oranges

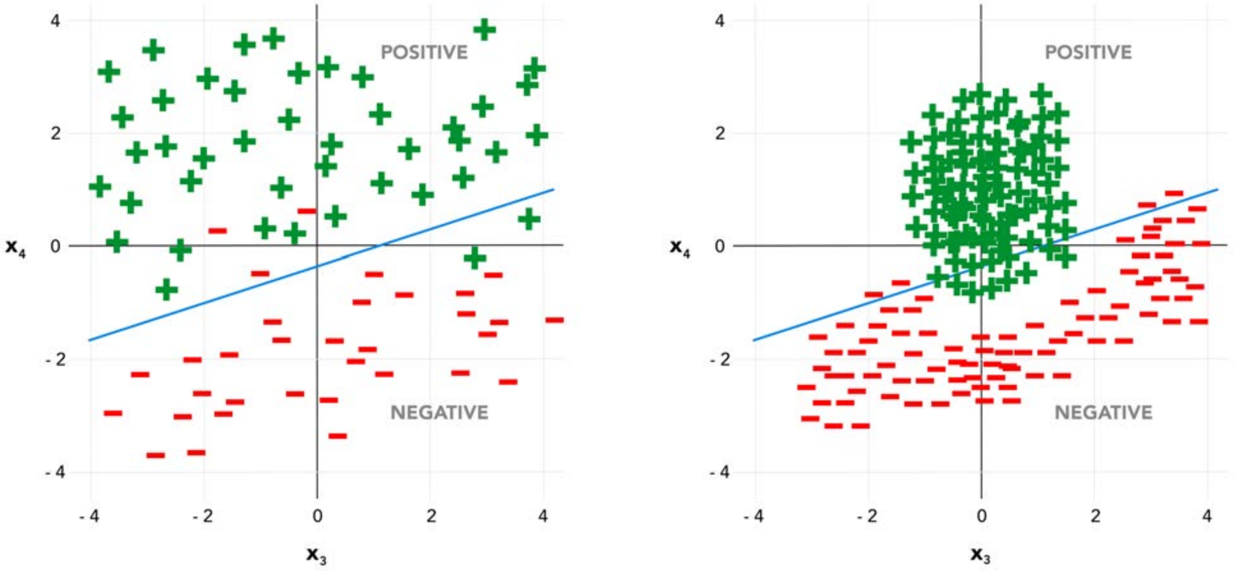
Acme Apparel (AA) is looking to expand to a new demographic (engineering college students), so they decide to use one of the hypotheses their departments have developed (using data from existing AA college student customers) to determine which college students to send their catalog to. AA also hires some consultants to gather some college student purchasing patterns data from a rival company specializing in engineering-themed apparel, Zenith Zapatos (ZZ).

1.2.1)

They use a hypothesis developed by the engineering department which has a 5% error rate on the data set A and had 10% error rate on data set B of existing Acme college student customers. They apply it to the college student customer data from ZZ and find it has a 35% error rate. What could account for this problem?

1.2.2)

Here is some actual data from AA current college student customers (left) and ZZ college-student customers (right). The engineering department finds another hypothesis, shown as the blue line below, that has the same error rate on both the AA data and the ZZ data. What do you think about this hypothesis? Should we use it for both populations? If not, what hypothesis or hypotheses would be better and why?



Checkoff 2:
Have a check-off conversation with a staff member, to discuss your answers until this point! While you're waiting, please continue working on the lab.

Ask for Help Ask for Checkoff

2) A More Rigorous Evaluation

AA and ZZ have merged to form A2Z and hired you as their chief data scientist! Luckily they have found a source of a huge amount of data. Now, your job is to come up with some procedures for predicting how well classifiers you come up with will perform.

Note the following notation and definitions, used throughout this problem:

- A generator $\mathcal{G}$ is a function that takes as input n, the number of samples desired, and returns a $(\mathcal{X}, y)$ pair where $\mathcal{X}$ is a d by n array of randomly sampled data points and y is a 1 by n array of their corresponding labels $\{+1, -1\}$.
- A training dataset $\mathcal{D}_{\text{train}}$ is a set of labeled samples $\mathcal{X}, y$ generated by a generator $\mathcal{G}$, where x^i represents the features of an object to be classified (vector of real and/or discrete values), and y^i represents the label of x^i. (You can think of i as the index for point x^i.)

2.1) Evaluating a classifier

Imagine that you have a generator $\mathcal{G}$ that pulls from a finite dataset of millions of points.

Let's assume that $\mathcal{D}_{\text{train}}$ is one such output of the generator $\mathcal{G}$.

Consider the situation in which you have run a machine learning algorithm on some training dataset $\mathcal{D}_{\text{train}}$, and it has returned to you a specific classifier h. Your job is to design (but not implement yet!) a procedure for evaluating h's effectiveness as a classifier. (Want more on classifiers? Check the the notes on Linear Classifiers.)

Assume we have a `eval_classifier` function that takes a classifier h, dataset D - a tuple of data and labels: (X, y) - and returns the percentage of correctly classified data points as a decimal between 0 and 1. We'll package it as follows:

```
def eval_classifier(h, D):
    test_X, test_y = D
    return score(h, test_X, test_y)
```

2.1.1)

Percy Eptron suggests reusing the training data to assess h:

```
eval_classifier(h, D_train)
```

Explain why Percy's strategy might not be so good.

2.1.2)

Now write down a better approach for evaluating h (which may use $h, \mathcal{G}$) and create a score function for h. The syntax is not important, but do write something down. *Hint: How can you incorporate $\mathcal{G}$?*

2.1.3)

Explain why your method might be more desirable than Percy's. What problem does it fix?

2.1.4)

How would your method from **2.1.2** score the classifier h on a different dataset $\mathcal{D}'_{\text{test}}$ that was generated from a different underlying distribution than $\mathcal{D}_{\text{train}}$ (i.e generated from a different generator)? Would the score be better, worse, or about the same? For example, if we're trying to predict whether or not a certain medication works, how well would a classifier trained on an adult population score on a population of children?

Check and submit this box once you've finished this section:

□ I've finished this section.

Save | Submit | Clear Answer | Ask for Help

As staff, you are always allowed to submit. If you were a student, you would see the following:
You have infinitely many submissions remaining.

Solution:

✓ I've finished this section.

Explanation:

Good job! Keep going!

3) Good Hypotheses: Beyond Accuracies

So far in this lab we have discussed the difficulties of evaluating a model and defined a slightly more rigorous process for evaluation. In all of the above cases, however, we made a few critical assumptions.

We assumed that we would always know with complete certainty whether a model made a correct or incorrect prediction on unobserved data. We also used accuracy as the ultimate quantitative metric of “goodness” of a model (albeit, accuracies on different datasets). In the real world, however, there are other metrics we should care about beyond just accuracy, such as fairness and whether the model's decisions have any ethical implications that need to be considered while evaluating the model's performance. Thus, we must be careful in how we define the "goodness" of a particular model. To illustrate this, let's look at a fictional case study.

To Loan or Not to Loan, that is the question

Feirna Sinemel was recently hired to be director at a mortgage lender company, Loan Investing Team (LIT) Corporation. LIT wants to be competitive with other industry leaders in the mortgage lending space, but has had difficulty scaling due to the limited speed at which humans can manually review loan applications.

The company is therefore interested in developing a machine learning model that can be used to determine whether or not to give someone a mortgage. They hope to use the model as an initial screening for the applications and reserve manual reviewing only for promising applicants.

Feirna knows you are learning about machine learning and reaches out to seek your advice on how to start designing a model. She suggests, however, that you approach this role with caution and reminds you that the mortgage business has a lengthy history of systemic racism and sexism , including a form of discrimination known as "redlining", which continues to this day.

The loan application assessment process is about assessing the likelihood that a person will be able to repay the loan in the given period of time.

3.1)

One of the first things to choose when developing a machine learning model is what inputs will be provided to the model and what the implications may be of including a particular input. For example, you may wish to include a person's current salary. Presumably, the company is more willing to give loans to those who have well-paying jobs. An implication of including a feature like "salary", however, is that historically certain demographic groups (ex. women, minorities, immigrants) have lower salaries, which may skew results of the model.

Other example inputs may include work history and loan repayment/approval history. What are the implications for including work history and loan history as inputs to the model? What are other inputs you may want to use for the model and what are the implications of using those inputs?

3.2)

In developing your model, would you opt to include sensitive inputs, like race or gender identity? What are the reasons you might do so, and what are the reasons you might not?

3.3)

What are the differences (if any) between using a computation model and having a human make these decisions? What are the benefits and drawbacks to each approach?

For next week's lab, we will delve deeper into this scenario and think about how to handle fairness in such situations.

Checkoff 3:
Have a check-off conversation with a staff member, to discuss your answers, and make sure you're setup for the rest of the week.

Ask for Help Ask for Checkoff