Balancing and Feature Scaling

CS522 Assignment #1 Help

But first, some FAQs…

FAQ 1: Do I have to use Python3.8?

No! You can use any STABLE version of Python 3.8 or later.

FAQ 2: Why do I keep getting different results?

We did not seed the random number generator that controls the train_test_split method (from scikit).

In order to fix this you can add the parameter: random_state=#, # is whatever seed you would like to use. For example:

```
xtrain, xtest, ytrain, ytest = train_test_split(data, labels, test_size= 0.30, random_state=5555)
```

Some other assignment notes...

1. We are going to lower the accuracy requirement from 95% to 90%. Just note that rebalancing can allow you to have a higher accuracy.
2. If you choose to rebalance, depending on which method you may want to write the code in task 4 or task 6.
    a. Option 1: Just leave a comment in task three telling us where you inserted the code
    b. Option 2: Write a method in task 3 (which does the rebalancing) and gets called somewhere else in your code
3. You can include the package “imblearn” in addition to the ones mentioned in Environment.pdf

FAQ 3: How long do question responses need to be?

Long enough to explain your thoughts (concisely) and support your answer.

In general, 3-5 sentences is a good rule of thumb but it could be more or less depending on the question.

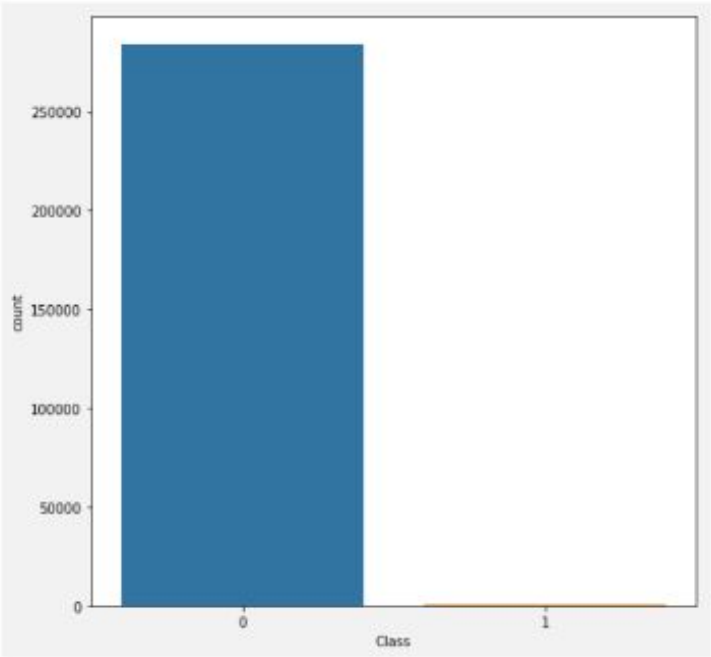
Now back to the lecture…

Balancing

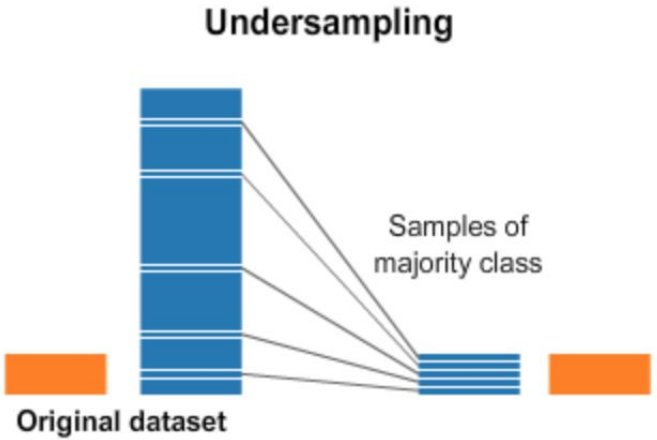
What is data imbalance?

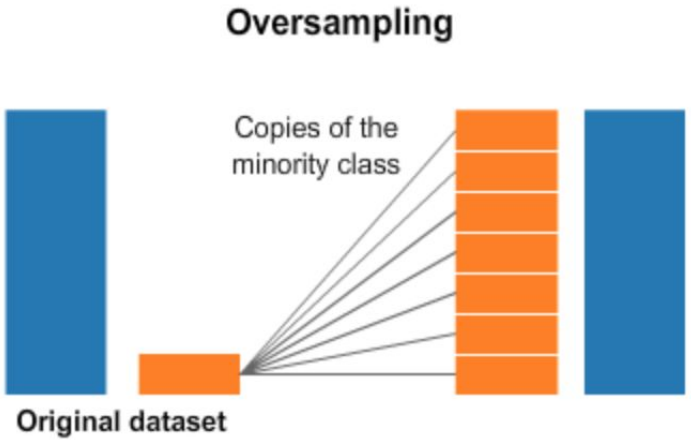
- One of the common issues found in datasets that are used for classification is imbalanced classes issue.
- Data imbalance usually reflects an unequal distribution of classes within a dataset.
    - For example, in a credit card fraud detection dataset, most of the credit card transactions are not fraud and a very few classes are fraud transactions.
    - This leaves us with something like 50:1 ratio between the fraud and non-fraud classes.

Class imbalance example



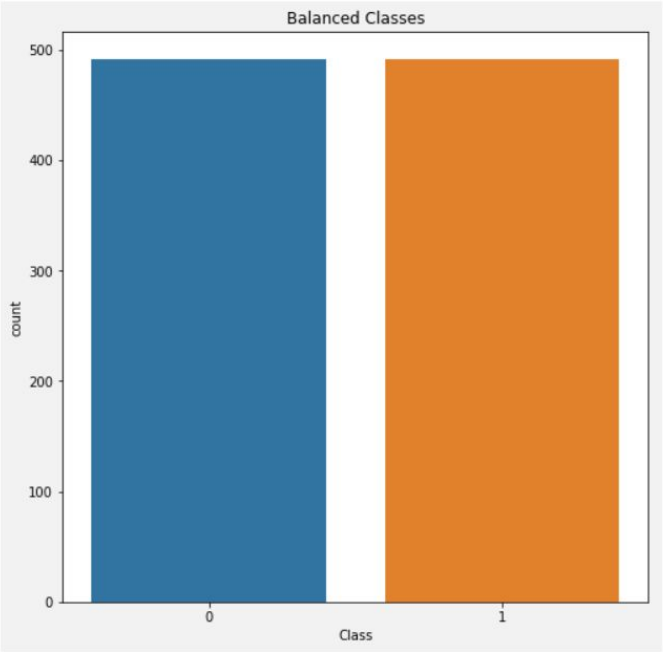
Resampling


Undersampling
Samples of
majority class
Original dataset


Oversampling
Copies of the
minority class
Original dataset

Undersampling

It is the process where you randomly delete some of the observations from the majority class in order to match the numbers with the minority class.

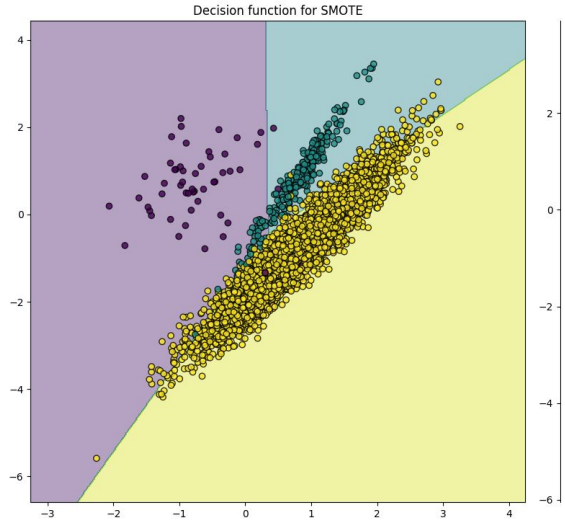


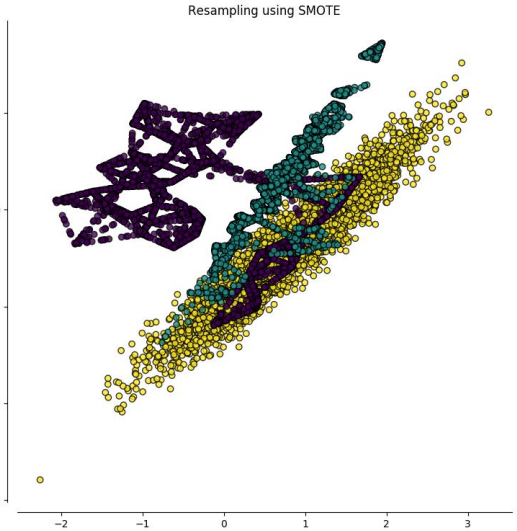
Balanced Dataset (Undersampling)

Oversampling

This process is a little more complicated than undersampling. It is the process of generating synthetic data that tries to randomly generate a sample of the attributes from observations in the minority class.

Example: SMOTE (Synthetic Minority Over-sampling Technique). It works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.





Scaling:
(I) Standardization / (II) Normalization

What is scaling?

Feature Scaling is where we force the values from different features to exist on the same scale, in order to enhance the learning capabilities of the model.

The two most common feature scaling techniques are **Standardization** and **Normalization**.

Standardization

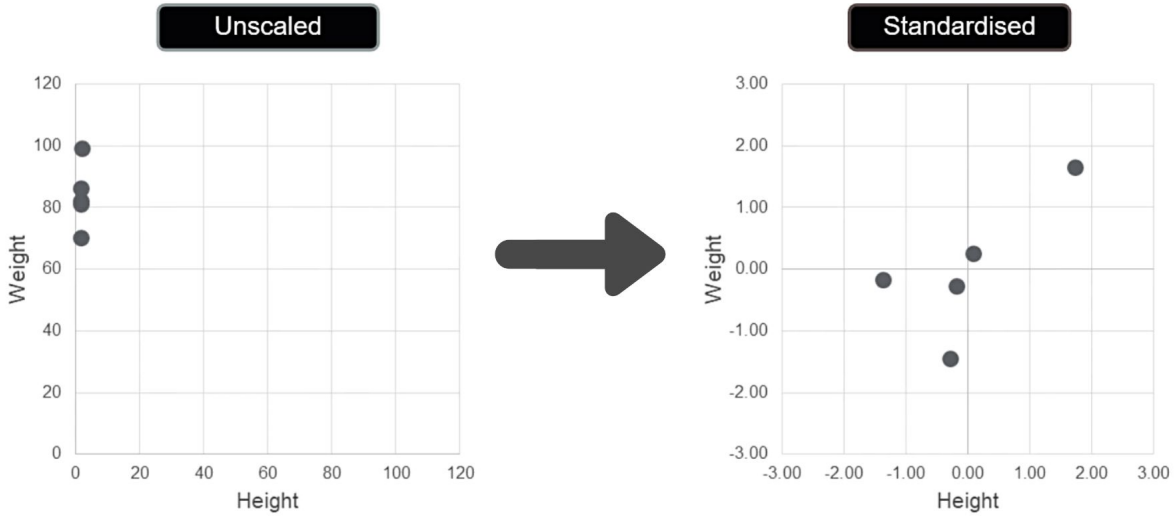
It rescales the data to have a mean of 0 and a standard deviation of 1. It assumes that your data has a Gaussian (bell curve) distribution.

Height (meters)	Weight (Kilograms)
1.98	99
1.77	81
1.76	70
1.80	86
1.64	82

Standardization

Mean	1.79	83.6
Std. Dev.	0.10954	9.35094

$$x_{standardised} = \frac{(x - mean(x))}{std.\ deviation(x)}$$



Normalization

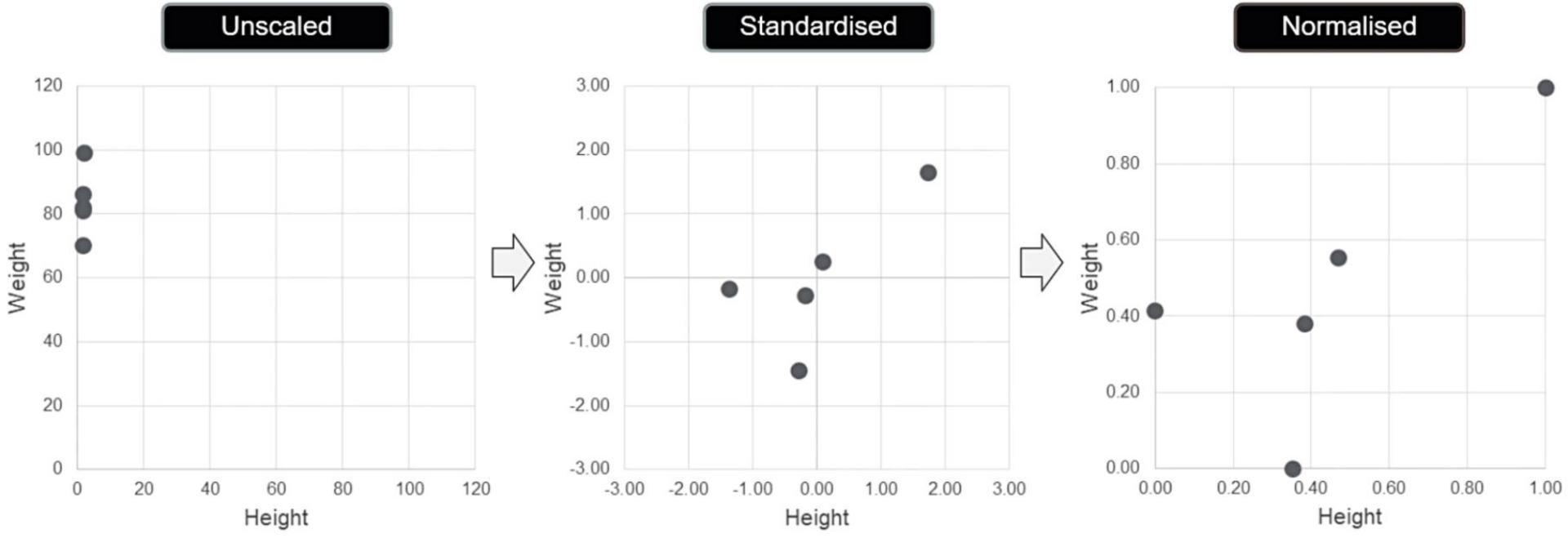
Normalization rescales data so that it exists in a range between 0 and 1. It is is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (bell curve).

Min	1.64	70
Max	1.98	99

$$x_{normalised} = \frac{(\ x - \text{minimum}(x)\ )}{(\ \text{maximum}(x) - \text{minimum}(x)\ )}$$

Which of the two techniques to use?

If you need your values to be positive, then go for normalization. But in general, it is common to go for standardization.



Final thoughts

- If you scale your values it makes it harder to understand the true meanings of any coefficients in terms of their actual values as you have transformed it to a different scale so they can exist together.
- It is often worth trying to run the model with and without scaling. If it doesn't appear to make any difference to the accuracy, then you can say that you don't need your data to be scaled and you will have an easier interpretation of your model's prediction.

Sources

- https://www.d2l.ai/
- https://python.plainenglish.io/feature-scaling-when-should-you-use-standardization-and-normalization-ea2eabb4a1d5
- https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
- Nitesh Chawla, et al. (2002) “SMOTE: Synthetic Minority Over-sampling Technique.”
- https://python.plainenglish.io/feature-scaling-when-should-you-use-standardization-and-normalization-ea2eabb4a1d5

Any other (general) questions for assignment 1?

Looking ahead to the next assignments…

Teams are made!

- I made the teams using the survey results
- I did my best to match people based on:
  1. Programming experience
  2. Machine learning experience
  3. Sensors/mobile experience
- Groups are available on Canvas (under announcements)
- You will be in the SAME group for assignment 2 and 3

Please reach out to your groups ASAP

- You will need to work together with your group to collect the data for projects 2 and 3
- We strongly suggest you get started sooner rather than later
- There are channels already created on Discord for you to use
- Please use the assigned channel (via group numbers on Canvas) in order to communicate

Group Questions?

- More details will be released about assignment 2 and 3 in the next couple weeks
- If you have any issues with your group, please reach out to your TAs ASAP