CSCD01 Assignment 2

Group cswasamistake

2021-03-06

**Bug #1:** “Regression in GP standard deviation where y\_train.std() == 0”

(Source: <https://github.com/scikit-learn/scikit-learn/issues/18318>)

Overview:

The first bug we have selected to fix is an issue regarding an unexpected error that occurs in the Gaussian Process module, when creating a GaussianProcessRegressor object with the parameter normalize\_y set to True. The parameter normalize\_y indicates that the dataset provided by the user will be normalized. Inspecting the code shows that the formula used to do this is done by subtracting by the mean of the dataset, and then dividing by the standard deviation of the dataset. The problem arises when the standard deviation of the dataset is zero; there is a division by zero error. This occurs when there is only one point in the dataset, or when all points in the dataset are the same.

Solution:

The solution implemented will check for the edge cases, and then raise a clear and concise error. Note that we considered an alternate solution of adding an insignificant value to the standard deviation to avoid division by zero. However, this hides the error from the user, changing the expected behaviour of the normalization done by the module. We have also considered simply automatically setting normalize\_y to False, and continuing without normalizing the dataset. This also hides the error from the user, and so we decided to raise an appropriate error to form the user instead.

Changes:

In the file gaussian\_process/\_gpr.py, approximately line 203; a case is added to check for when standard deviation is equal to zero.

Testing:

The test cases are implemented within the existing test suite of the library, under the gaussian\_process module.

In the file gaussian\_process/test\_gpr.py, approximately line 276; two test cases were added under the function named test\_no\_standard\_deviation. The first test case constructs a NumPy array with only one data point. The second test case constructs a NumPy array with three identical data points. The two test cases together exhaust the possibilities of when there is only one datapoint, and when there are multiple datapoints (and all of them are identical).

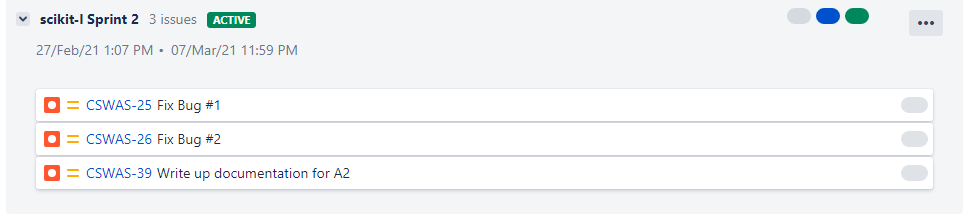
**Use of tools:**

Github:

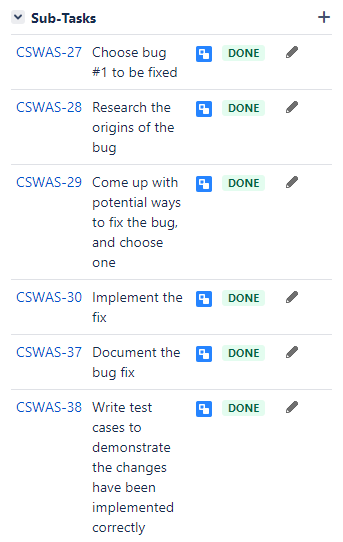
Each assignment is under the correct folder, and one copy of the scikit-learn library is cloned. The changes are all made in the cloned library, and each bug fix is done on a separate branch before being merged with master.

Jira:

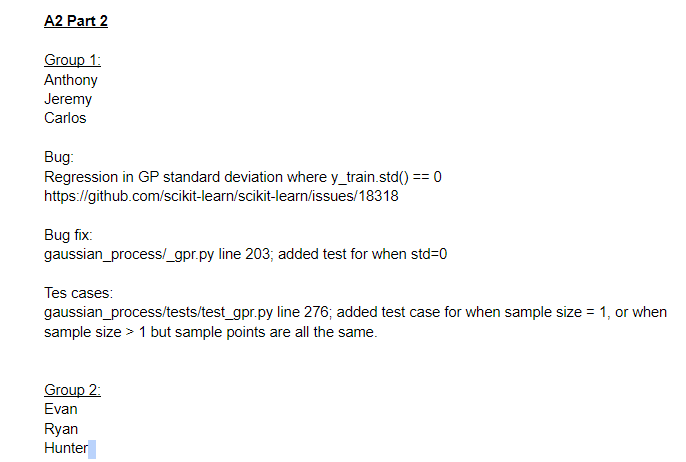
We have planned our workflow of assignment 2 as the second sprint of our project. Each bug fix is shown as an issue:



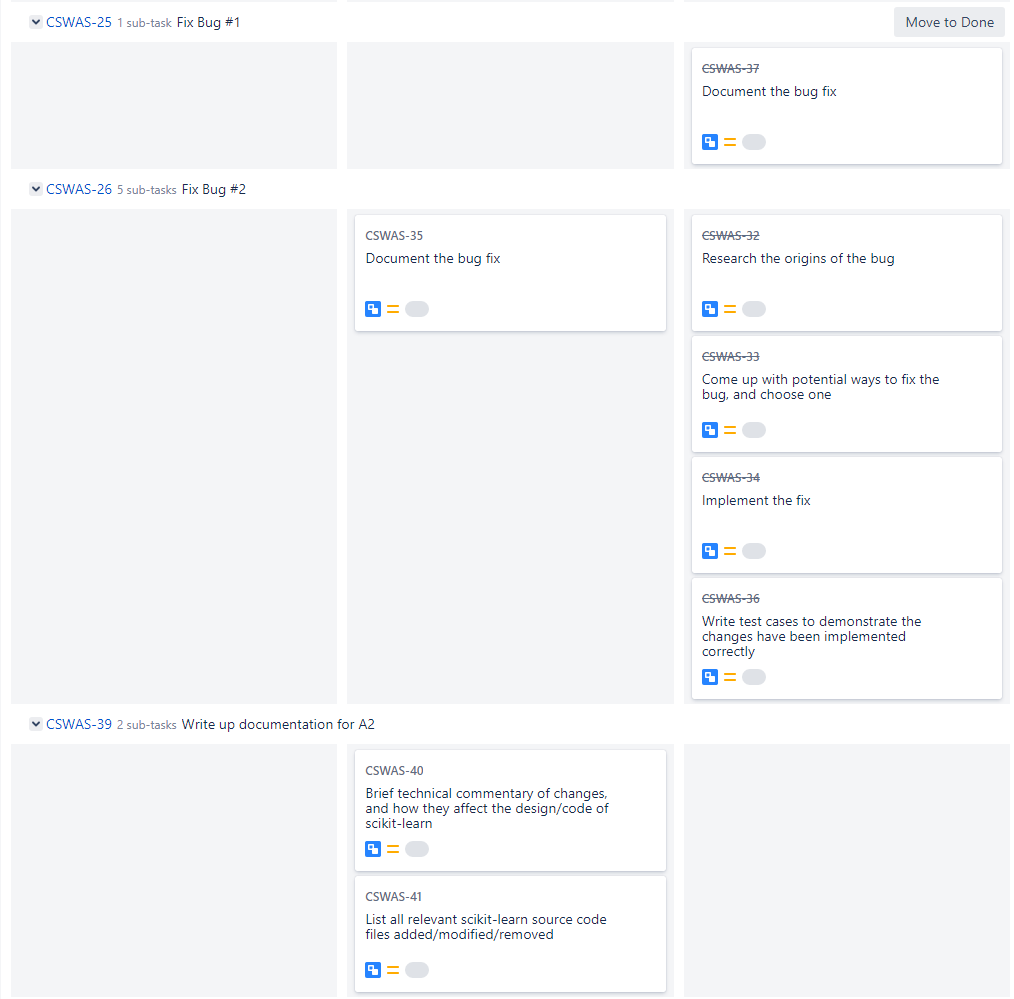
Also, each issue (bug fix) has six sub-tasks assigned to it:



The bug fixes do not have assignees on Jira. This is because Jira does not allow multiple people to work on the same issue. As such, we have assigned groups of three to work on each bug through Discord and a shared Google docs file:

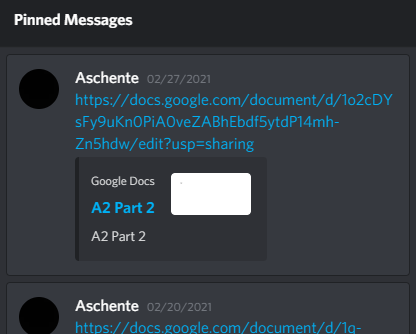


As of the creation of this document, our workflow progress is as follows:



Discord:

We use pins to link to shared Google docs files for referral:



Our Discord server is the main means of communication, and previous chat logs show that we meet up every 2-3 days. Below are snippets of our chat logs.

