(This file got updated)

PART III: Putting it together and writing the report

1. Code

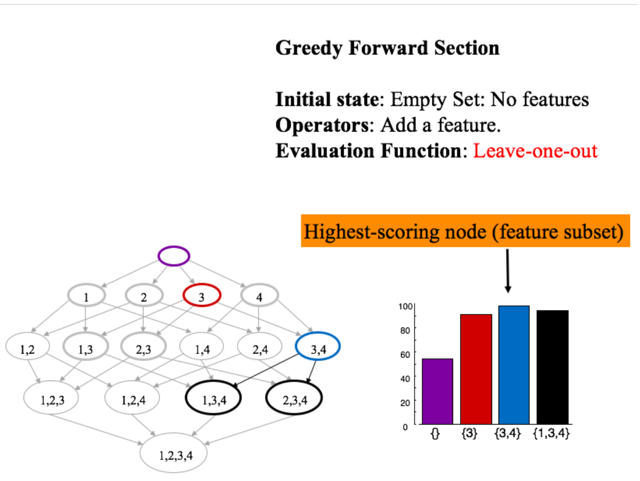
Remember that your feature search algorithm in part I didn't use real data, a real classifier, or a real evaluation function. Instead it only worked with feature numbers and assigned random accuracies to feature subsets.

Now, you do have a classifier (nearest neighbor classifier) as well as an evaluation function (the leave-one-out validator) that you implemented and tested in Part II!

All you need to do for Part III is to replace the dummy evaluation function (random number generator) in your feature search algorithm with the leave-one-out validator. After that, your feature search algorithm will be complete: Given a data file, it should be able to search for the feature subset that results in the highest accuracy and report that feature subset along with the corresponding accuracy.

Please refer to the Project intro file again to review what the whole system is supposed to do:

https://docs.google.com/document/d/1UWfb-Twxxjb2smHPphPk76GdajF3Fcs6_IrXixNLjc0/edit?usp=sharing



Greedy Forward Section

Initial state: Empty Set: No features
Operators: Add a feature.
Evaluation Function: Leave-one-out

Highest-scoring node (feature subset)

2. Testing

Again, you can first test your system using the previous small and large datasets (**Note that your results can be slightly different than these**):

Small Dataset (Has 100 instances and 10 features)

- Your complete feature search algorithm should find features {3, 5, 7}, with an accuracy of about 0.89

Large Dataset (Has 1000 instances, and 40 features)

- Your complete feature search algorithm should find features {1, 15, 27}, with an accuracy of about 0.949

Once you debug your system and get results that are similar to above, you can proceed with your **personal datasets**, which are very similar to the above datasets.

2.1 Finding and downloading your personal datasets:

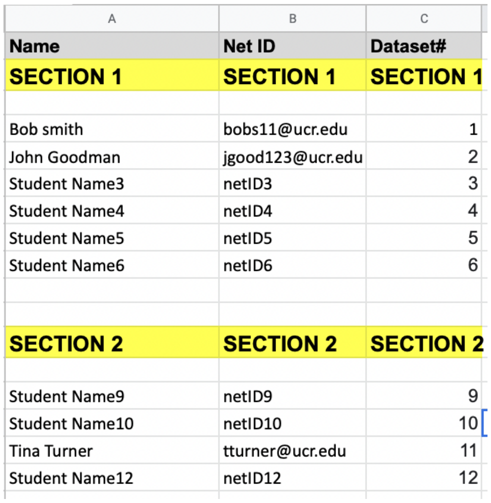
Please find your personal dataset# from the spreadsheet below (The first group of rows are students in section1 and the second group of rows are students in section2.):
https://docs.google.com/spreadsheets/d/1M0i9ithubhyE5y1_-dGpyYBUN6lJG0LoUeWFfmDEbzE/edit?usp=sharing

And then go to the following folders to download your small and large datasets:

1. Small dataset folder:
   https://drive.google.com/drive/folders/1OTFloME73RMmKzJTRVFRUiklOMevZcSV?usp=sharing

2. Large dataset Folder:
   https://drive.google.com/drive/folders/1YBbenjphIV6iRrh_BWv70NgKpTIw5GsX?usp=sharing

For example, according to the following sample datasheet, the student named **“Tina Turner”** in **section2**, needs to download CS170_Spring_2022_**Small_data__11**.txt and CS170_Spring_2022_**Large_data__11**.txt from the corresponding small and large dataset folders.



A	B	C
Name	**Net ID**	**Dataset#**
SECTION 1	**SECTION 1**	**SECTION 1**
Bob smith	bobs11@ucr.edu	1
John Goodman	jgood123@ucr.edu	2
Student Name3	netID3	3
Student Name4	netID4	4
Student Name5	netID5	5
Student Name6	netID6	6
SECTION 2	**SECTION 2**	**SECTION 2**
Student Name9	netID9	9
Student Name10	netID10	10
Tina Turner	tturner@ucr.edu	11
Student Name12	netID12	12

2.2. Reporting your results:

You are going to report your results **BOTH in the beginning of the report (filling out the provided table**) and **in the comments when submitting** your assignment (instructions at the end of this guide).
You need to report your results on your **personal small and large datasets for both forward-selection and backward-elimination algorithms.**

3. The Final Report (includes trace)

Please follow the report **TEMPLATE** provided here:
https://docs.google.com/document/d/1jLl2dgy1KZjD1q5FLoNkOPESj8lkXj9d/edit?usp=sharing&ouid=106145693636121867539&rtpof=true&sd=true

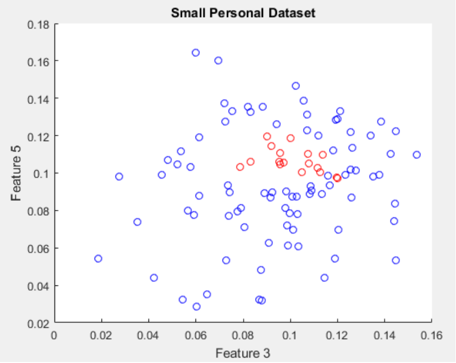
The first page **HAS** to be **EXACTLY** the same as the template with your **info and solutions filled in the table. The rest are suggested sections based on the previous submissions that have gotten the best grades. You can customize those sections.**

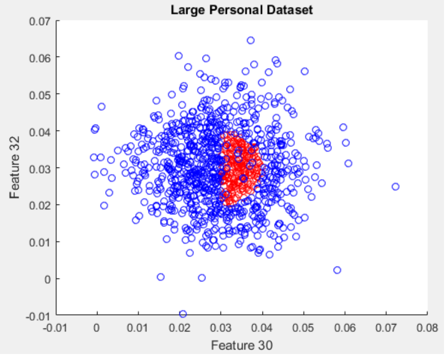
Your report should summarize **your findings**. You need to **compare the forward selection and backward elimination (and optionally your own) search algorithms** **on 4 datasets:**

a. The **initial small and large** datasets that I gave to everyone along with the correct answer (to test their code) and
b. **Your own small and large** datasets.

Here is a list of items you can add to your report. Of course you can add more items, if meaningful and informative.
- Challenges
- Your design (objects and methods)
- Did you try optimizing your code by using special data structures or algorithms to save time/memory?
- Plots for features that do separate the classes well and features that don't (see figures below); and their analysis
- Effect of normalizing the data (a table or chart that shows how it affects classification results/accuracy) and discussion
- comparison of different algorithms on different datasets and discussion (you might want to compare running times, memory usage, accuracy, etc)
- **If** you experimented with more than one nearest neighbor (e.g., 3 nearest neighbor, 5- nearest neighbors, etc, you can compare the results via charts/tables/plots/etc.) Note that using more than one neighbor is not required but some students prefer to do that.
- Your references (any material that you consulted or tools you used, etc.)
- **Trace on your personal small dataset (sample provided below)**

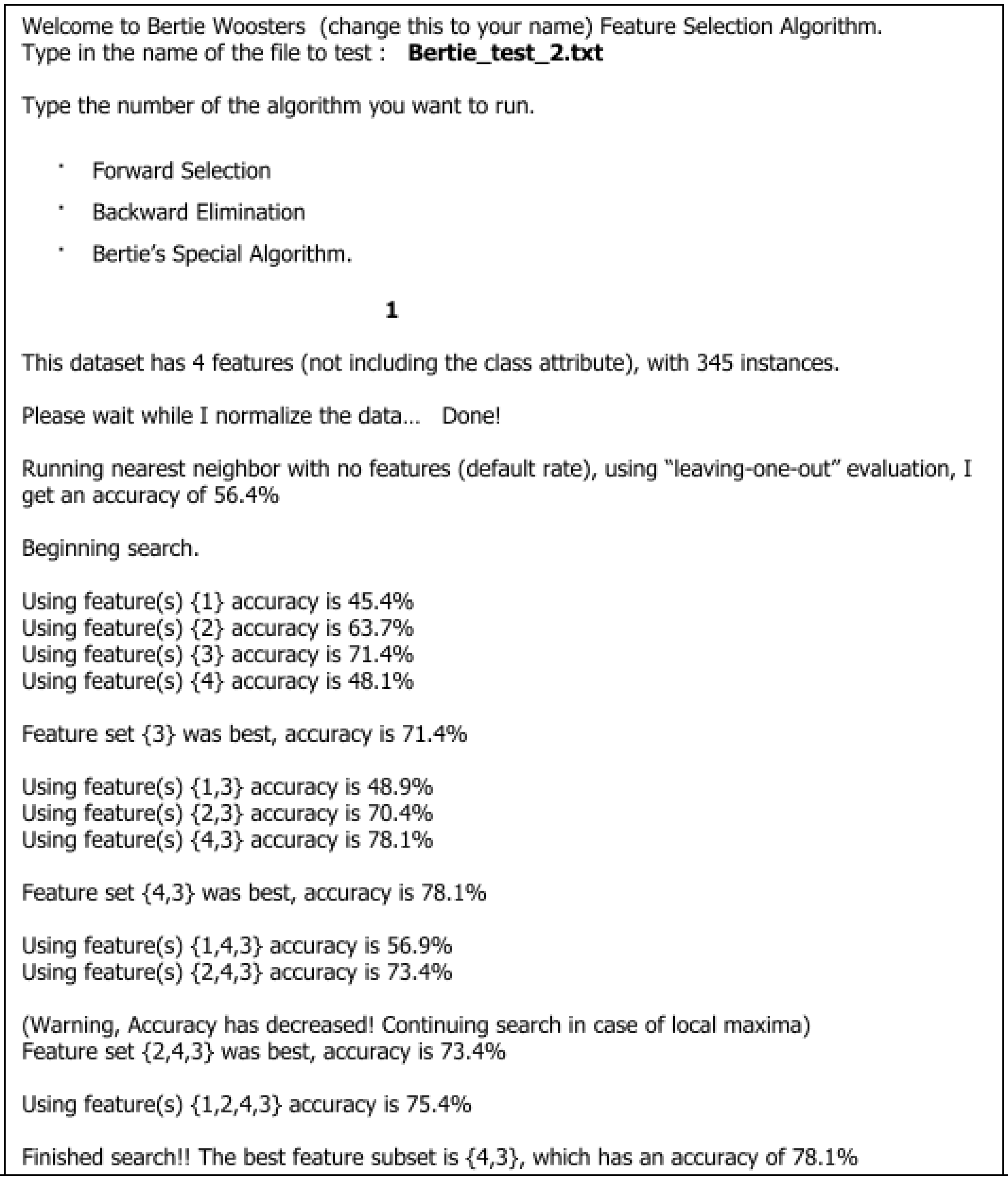
Note: Please have names and captions for your plots, figures and tables. Plots need to have labels for each axis and legend.





Sample Trace (To be added to the end of the report):

You will need to paste a **trace** like this at the end of your report (NO NEED FOR TIME ELAPSED):



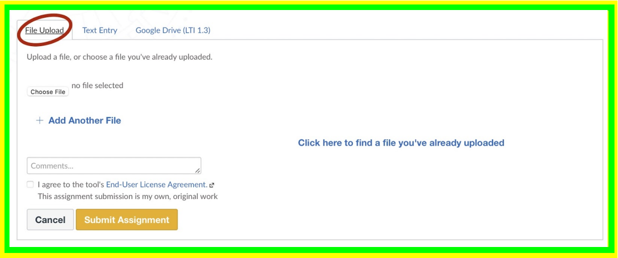
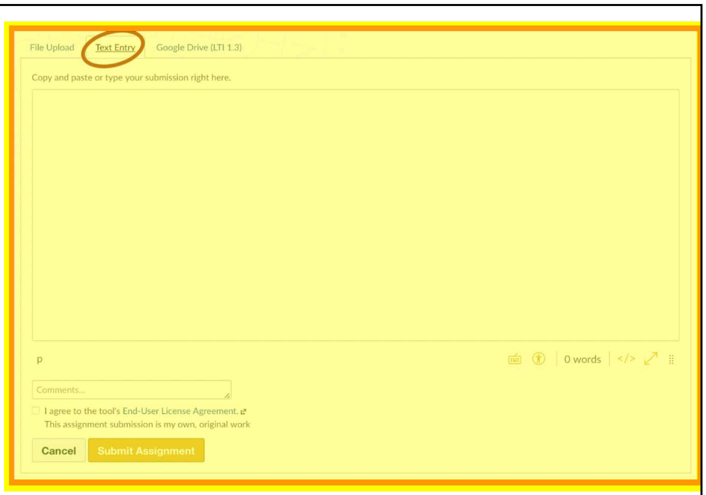
```
Welcome to Bertie Woosters  (change this to your name) Feature Selection Algorithm.
Type in the name of the file to test :   Bertie_test_2.txt

Type the number of the algorithm you want to run.

    ·   Forward Selection
    ·   Backward Elimination
    ·   Bertie's Special Algorithm.

                                        1

This dataset has 4 features (not including the class attribute), with 345 instances.

Please wait while I normalize the data...   Done!

Running nearest neighbor with no features (default rate), using "leaving-one-out" evaluation, I
get an accuracy of 56.4%

Beginning search.

Using feature(s) {1} accuracy is 45.4%
Using feature(s) {2} accuracy is 63.7%
Using feature(s) {3} accuracy is 71.4%
Using feature(s) {4} accuracy is 48.1%

Feature set {3} was best, accuracy is 71.4%

Using feature(s) {1,3} accuracy is 48.9%
Using feature(s) {2,3} accuracy is 70.4%
Using feature(s) {4,3} accuracy is 78.1%

Feature set {4,3} was best, accuracy is 78.1%

Using feature(s) {1,4,3} accuracy is 56.9%
Using feature(s) {2,4,3} accuracy is 73.4%

(Warning, Accuracy has decreased! Continuing search in case of local maxima)
Feature set {2,4,3} was best, accuracy is 73.4%

Using feature(s) {1,2,4,3} accuracy is 75.4%

Finished search!! The best feature subset is {4,3}, which has an accuracy of 78.1%
```

4. Submission

UPDATE: It turns out that Canvas doesn't allow both text entry and zip file submissions and it is an either or choice.

Please ignore the text entry (I cannot remove it at this point because students have attempted text submissions already)

>>>>Please submit the zip file and then report your results on the personal dataset in a COMMENT under your submission.<<<<<

 File Upload Text Entry Google Drive (LTI 1.3) Choose File no file selected + Add Another File Click here to find a file you've already uploaded Comments... Cancel Submit Assignment Upload a file, or choose a file you've already uploaded. □ I agree to the tool's End-User License Agreement. This assignment submission is my own, original work	 File Upload Text Entry Google Drive (LTI 1.3) p 0 words Comments... Cancel Submit Assignment Copy and paste or type your submission right here. □ I agree to the tool's End-User License Agreement. This assignment submission is my own, original work
Tab1: File upload (SUBMIT THIS AND ADD A COMMENT)	Tab2: Text Entry (**IGNORE**)

For reporting your results in the **COMMENTS**, please make sure you follow the format below:

- **DatasetID:** <your_dataset_ID>
- **Small Dataset Results:**
    - **Forward:** Feature Subset: <your best feature subset>, Acc: <your accuracy on that feature subset>
    - **Backward:** Feature Subset: <your best feature subset>, Acc:<your acc. on that feature subset>
- **Large Dataset Results:**
    - **Forward**: Feature Subset: <your best feature subset>, Acc: <your accuracy on that feature subset>
    - **Backward**: Feature Subset: <your best feature subset>, Acc: <your acc. on that feature subset>

Here is an example:

- **DatasetID:** 211

- **Small Dataset Results:**
    - **Forward:** Feature Subset: {1,2,4}, Acc: 0.86
    - **Backward:** Feature Subset: {1,5,4} Acc: 0.83
- **Large Dataset Results:**
    - **Forward**: Feature Subset: {23,56,12}, Acc: 0.95
- **Backward**: Feature Subset: {23,36,12}, Acc: 0.96