(This file IS updated on June 4, 9:50pm)

PART III: Putting it together and writing the report

1. Code

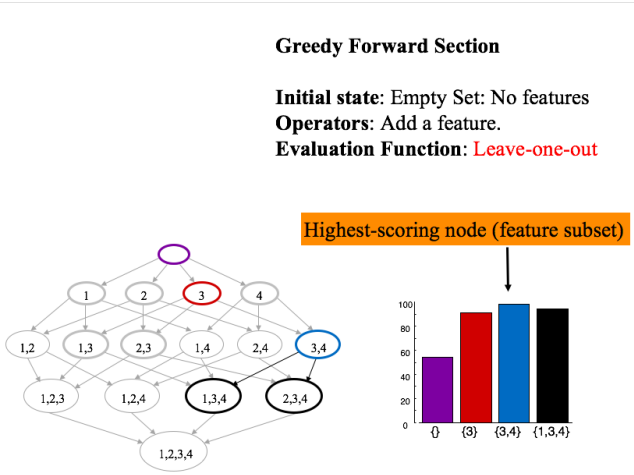
Remember that your feature search algorithm in part I didn't use real data, a real classifier, or a real evaluation function. Instead it only worked with feature numbers and assigned random accuracies to feature subsets.

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



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.

Finding and downloading your personal datasets (see team instructions below):

1) Lookup your name and user ID in this spreadsheet and find the file number associated with you (e.g., 34):
   https://docs.google.com/spreadsheets/d/1akU5ffOhrUbyzl4qUwwHkeUpqtUEQArtoYgJGoPnveI/edit?usp=sharing

2) Go to the following folders and retrieve the file ending with the number assigned to you (e.g., cs_170_small34.txt , cs_170_large34.txt):
   Small: **https://drive.google.com/drive/folders/0B8a6g1ZBMmUtWEhQRzBUT3h5QlE?usp=sharing**
   Large: **https://drive.google.com/drive/folders/0B8a6g1ZBMmUtb1NueDVHd0Q5YlU?usp=sharing**

Reporting your results to verify them with us (This replaces project DEMOs):

You will need to report your results by sending a **PRIVATE POST on PIAZZA** and getting confirmation about its correctness. **PLEASE DO NOT EMAIL THE INSTRUCTORS.**

I have made a post named **"TODO: Post your Personal Dataset Results Here Privately"** on Piazza: https://piazza.com/class/kmty756lfnl545?cid=191 **Please post your results privately under that post.**

For Team projects:

If you work as a team (with 2 members Student1 and Student2), please **choose ONLY one** of the datasets (either the one assigned to student1 or the one assigned to student2).

To confirm your results with the instructors (via piazza post), please do it only once on behalf of both team members (DON'T send two separate posts).

Example: Rutuja and Nikola are teammates. The spreadsheet says Rutuja is assigned datasets #23 (i.e., small#23 and large #23) and Nikola is assigned datasets #11 (i.e., small #11 and large #11).

1) Rutuja and Nikola will simply decide whose datasets to use (either #23 or #11, doesn't matter which one of the two). They run the code and find the best feature subsets and accuracies for their search algorithms (forward selection, backward elimination and optionally, custom algorithm)

2) Either Rutuja or Nikola will send a private post on Piazza (under the specified Piazza post that I have created). The post will be something like:

Dataset#23 - small: best feature subsets:
Forward selection: {2,6, 9}, Acc: 0.98
Backward Elimination: {1,6, 9}, Acc: 0.97

Dataset#23 - large: best feature subsets:
Forward selection: {12, 36, 42}, Acc: 0.87
Backward Elimination: {10, 36, 42}, Acc: 0.90

The instructors will check your results and let you know if you got them right.

3. The Final Report (includes trace)

Please follow the report **TEMPLATE** your TA has provided:
https://drive.google.com/file/d/1wyZQJNDVcnTE2bcFBzVQAat3bcYMWMjG/view?usp=sharing
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.

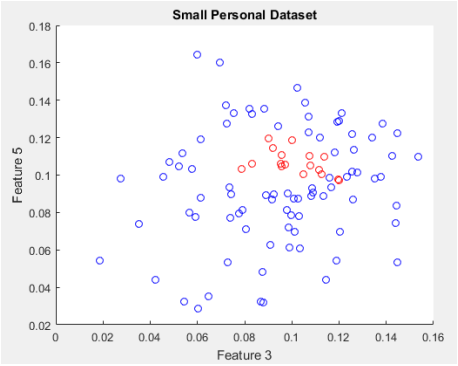
Your report should summarize **your findings**. You need to **compare the forward selection and backward elimination (and optionally your own) search algorithms** on 4 datasets:

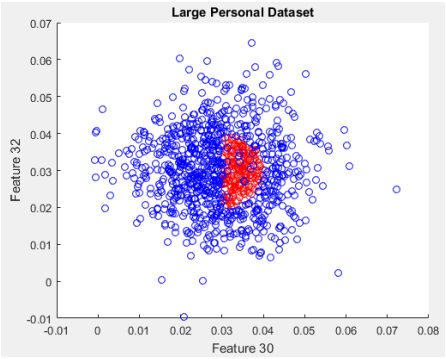
a. The initial small and large datasets that I gave to everyone along with the correct answer (to test their code) and
b. Your own small and large datasets.

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 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. Submissions

1. code.zip
2. Report.pdf (you paste the trace at the end of the report)