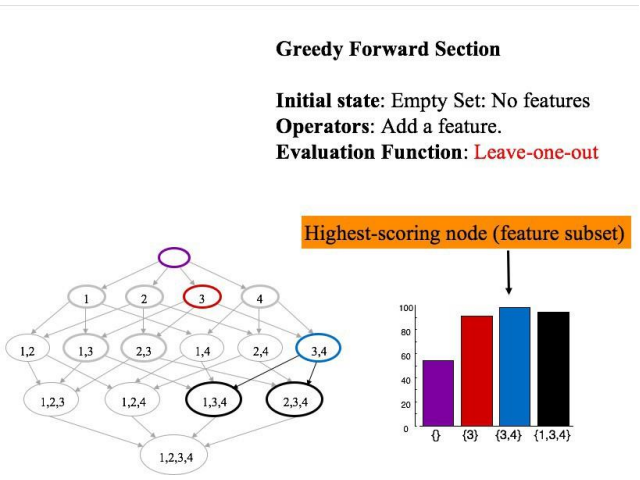
PARTS III, IV: Putting it together and writing the report

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!

In Part III, you will need 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.



Testing

Again, you can first test your system using the previous (from Part II) 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 like above, you can proceed with the **Titanic Dataset**.

2.1 Titanic Dataset

The Titanic dataset is a real dataset containing the information of a subset of passengers in the Titanic as well as their survival outcome (https://www.kaggle.com/competitions/titanic). However, there is a problem: The dataset comes in a different format. You have two options; you can change the code to accept the new format or you can change the file to fit what your code expects.

In this guide I will show you how to do the latter:

1. Identify the target class (Survival)
2. Remove the PassengerID, Name, Ticket, and Embarked columns (if you are using Google Sheets, you can right click the column and select "Delete Column"
3. Find any row that is missing a value (some passengers don't have an Age entry) and remove that entry
4. Under the Sex column, replace male and female with a number for each (e.g. 1 and 2).

5. Convert all numbers to scientific notation (On Google Sheets, you can select your data, then go to the Format menu, choose Number, and then Scientific).
6. Download as CSV, replace commas with two spaces, convert the file to TXT.

(Too much? No worries, you can use the clean version that is ready for your code, but it is important to get familiarized with the sanitization process)

Now it is the time to try it out, can you identify which features are most predictive of survival?

2.2. Reporting your results:

You need to report your results on the Titanic dataset for both forward-selection and backward elimination algorithms.

3. NN to KNN (K-Nearest Neighbors)

This is the last step of your project: extend your NN code to KNN, to work with values of k > 1.
Repeat all tests performed on the Titanic dataset, this time using k = 1, 3, 5, and 7, and compare the results.

PART IV

The Final Report (includes trace)

Please follow the report TEMPLATE provided.

Your report should summarize your findings. You need to compare the forward selection and backward elimination (and optionally your own) search algorithms on 3 datasets:

a) The **initial small and large** datasets that I gave to everyone (PART II) along with the correct answer (to test their code) and

b) The Titanic dataset.

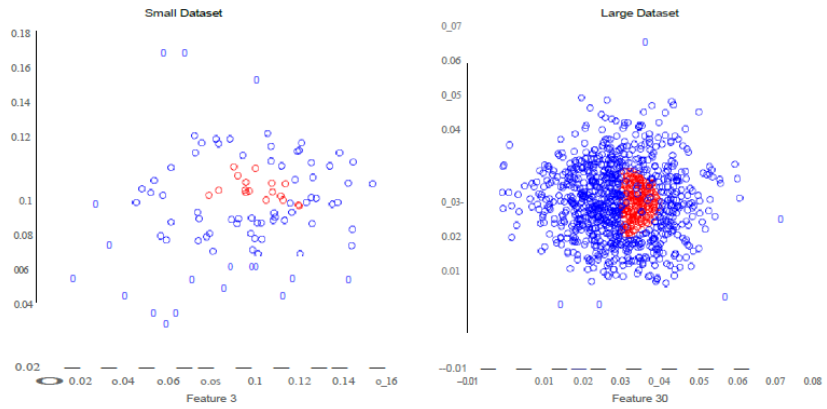
Here is a list of items you need to add to your report. Of course, you can add more items, if meaningful and informative.

- Group members and contribution of each student in the group.

- 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.)
    - Provide the results and a comparison of KNN performance for various kkk values (e.g., using charts, tables, or plots). It's sufficient to compare accuracy only.
    - Your references (any material that you consulted or tools you used, etc.)

    - Trace on the Titanic 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:

(change this to your name) Feature Selection Algorithm.
Type in the name of the file to test : **John_test_2.txt**

Type the number of the algorithm you want to run.

Forward Selection
Backward Elimination

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%

- **Group: For each student in the group list Students Name, NetID, and**
- **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:

…

Titanic Dataset Results, k = 1:

…

Titanic Dataset Results, k = 1, 3, 5, 7:

…

Here is an example:

Group: Name1 – NETID1 – Session 1, Name2 – StudentID2 – Session 2

- **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