Due on March 16th 2017, beginning of class (Online students can email me a PDF/MS Word file, by that time).

You can either do this in two person teams, or as individuals.

If you do it as a team, I will expect a little more work. It is important both team members understand everything in the report, you cannot divide the labor like “*you do the coding, I will get the coffee*”.

-

You have a choice of A, B or C.

1. Do a project in data mining of your own choosing, it could overlap another classes/project. *Before* you do this, you must talk to me in person, to pitch your idea.
2. Do feature search using KNN
3. Clustering and visualization

----

**B) Feature search using KNN**

In this project you will need to do KNN (many times!) You can us the build-in MATLAB code, or adapt the code I gave you, or write your own.

1. Read 235\_Feature\_search\_hints.pptx
2. Implement a forward feature search algorithm, that attempts to find the best feature subset of a given dataset. Use just 1NN for simplicity.
3. Test your algorithm on these two datasets, you will need to z-normalize these datasets.

On dataset 1 the accuracy rate can be 0.89, when using only features 57 70 99

On dataset 2 the accuracy rate can be 0.93, when using only features 55 87 41

1. Your answers might be a *little* different, you might have gotten one or two spurious features, and you might have miss one of the true features. However, if you did not get very similar results, STOP! Check your code, if you still have problems, come see me. Plot the accuracy vs the number of features.
2. Now run your code on the two mystery datasets below, report the features you find and the accuracy you can achieve.

On dataset 3 the accuracy rate can be ----, when using only features --------------

On dataset 4 the accuracy rate can be ----, when using only features -------------

1. Go to the UCI data archive, and download two datasets (note, you may need to download more, “play” with them as discussed below, and report only two). If another team hands in the exact same pair of datasets as you do, I will make both groups do one more dataset.
2. Run standard KNN, classifying the *test* against the *train*, call this *Benchmark*.
3. Now run your search algorithm, on just the *train* data.
4. Using the **best** subset you found, classify the *test* against the *train*, did you beat the *Benchmark*?
5. Write an explanation as to why certain features were dropped from the dataset, and others were kept. For example.. *When classify GOODvsBAD\_Student, our search algorithm kept {GPA, SAT, GRE, ZIPCODE}, but dropped {EYECOLOR, SEX }. The inclusion of ZIPCODE was initially surprising, why should where you live effect how good student you are? However, in the US, persons coming from “rich” zip codes are more likely to have gone to a school with lots of resources (see Smith and Jones 2014)… … it makes sense that SEX was dropped, because we might expect males and females equality likely to be…*
6. Your report should be a self-contained narrative, with all appropriate tables, figures, citations.

----

1. **Clustering and visualization**

(note, this project is a little more opened ended than the above)

1. Go to the UCI data archive, and download three datasets. If another team hands in the exact same datasets as you do, I will make both groups do one more dataset. The datasets you grab should be mostly or all *numeric*, and have at least 20 features.
2. Write a function that can plot any pair of features, colored by class labels (you have seen dozens of such examples in my slides, these plots are sometimes called *scatter plots*).
3. Some pairs of features will give better visual “separation” than others. For example, in the below (left), **a** is best, **b** is pretty good to, but **c** and **d** are not very good (note, it is possible, but unlikely, that *no* pair is particularly good).
4. Try to find the best, or at least, very good pair to make a scatter plot. Try selecting the pairs by…
	1. Learning a decision tree, and using its two best features (recall Decision\_Tree\_Matlab.pptx)
	2. Measuring the correlation of data to class labels, and use the pair that has the highest correlation
	3. Something else you came up with (you must come up with your own idea, and explain it)
5. **From this point on, just work with your favorite example from the three datasets above.**
6. Randomly select a manageable number of instances, from one of your datasets (maybe 30 to 60). Create two dendrograms, one using *all* the features, one using just the two features that made the best scatter plot. Color the branches based on class labels (if you can’t do this in matlab, paste the figure in to powerpoint, ungroup it, and manually edit the colors line thicknesses etc.) Annotate the dendrogram with some interesting observation (as in my example below)
7. Your report should be a self-contained narrative, with all appropriate tables, figures, citations.

|  |  |
| --- | --- |
|  |  |
|  | **Figure X.** A clustering of random expamplars for the XYZ dataset, notice that instances 18 and 19 are cats, but the are grouped with pigs. We discover that they are *Maine Coons,* the largest… |