Homework 4 - Solution

Task

Download data set GDS4168 from http://file.biolab.si/files/GDS4168.zip. (The documentation is at at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4168). Unzip before use.

1. Use classification tree and logistic regression to construct a model for this data. Estimate their performances.

2. Run k-means clustering. How well do the clusters correspond to classes? This data set has very few instances, so Sieve could be misleading (try it!) We recommend using the Distributions or Box plot widget instead.

3. Did both approaches work well?

   - If yes: what does it tell you about the data? What must the data look like if clustering is able to re-discover the classes?

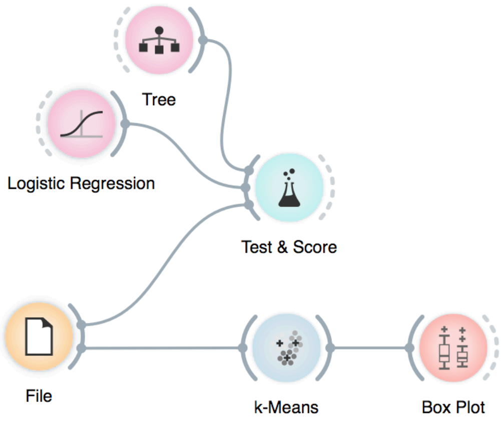
   - If not: what does it tell you about the data? How can you have excellent predictive models but clustering is not able to re-discover the two classes as subgroups in the data?

   Towards answering this question: can you paint some data in which the logistic regression has a perfect AUC and k-means discovers that there are two very distinct clusters - but the two clusters do not correspond to the two classes? (You can add this painting to the report if you wish.)

Solution

This was one of the more interesting homeworks. The aim was to teach you the difference between classification and clustering. But last year one student discovered something really interesting in this data. Read on!

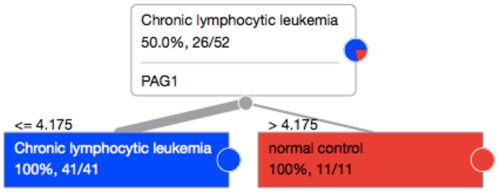
We need the schema on the right.



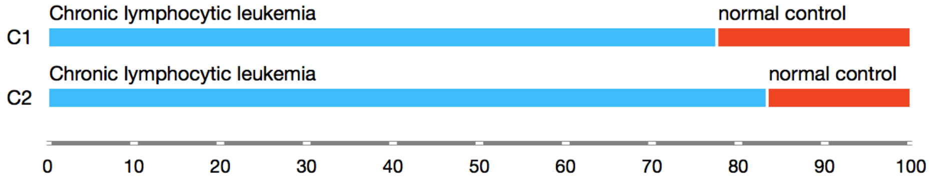
Logistic regression and classification trees give almost perfect results: their classification accuracy is 86.5 % and 98.1 %, respectively, and AUCs are 0.875 and 0.950.

These results are so good that we may suspect that problem may be trivial. Indeed, the classification tree on the entire data set looks as shown on the right. When the entire data set is considered for training, the classes can be perfectly separated using a single feature (expression of PAG1).

If the problem is trivial, then clustering shouldn't have any problems with it either, right? We should get two clusters, each one covering its own class…



… except that we don't. We get two clusters that contain roughly the same proportion of instances from both classes. The two clusters are unrelated to the two classes.



So, how come that clusters did not discover the two classes that are so trivial to discriminate between?

Difference between clustering and classification

Consider another example. Swedish are known to be tall, blonde, blue-eyed, and they speak Swedish amazingly well. Greeks are known to be the opposite: they are not tall, have dark hair and eyes and seldom utter a Swedish word. Two research institutions from Sweden and Greece who work on a joint project researching prostate cancer have collected data on 200 people. The data includes body height, hair and eye color, the person's native language, her or his gender and age, and whether she or he has prostate cancer or not.

If they build a classification tree, it's going to start with the gender, since the occurrence of prostate cancer among women is rather small in comparison with men. Among men, it's going to continue with age. Assuming the data is balanced, predicting no cancer for women and young men should correctly cover 75 % of data instances. This makes the problem rather simple and the classification accuracy should be good. The tree probably will not (and should not) include hair and eye color and the ability to distinguish between a, å, á and ä.

If they instead use clustering, they will find two clusters. The first will contain tall, blonde, blue-eyed Swedish speaking people, and the other will cover the shorter, dark-haired dark-eyed people that can't even say "God dag!", but have much less problems distinguishing between ϰ and χ — since these are the two actual groups of people in the data. The greatest distinction is not between those who have prostate cancer and those who don't, but between the two nationalities. The two clusters would not differ with respect to gender and age, since the gender and age distribution is roughly the same in both countries.

Classification models — and in particular classification trees — look for the few features that distinguish well between the *predefined groups* (that is, classes). Clustering observes values of all features to *define groups*.

The message of this homework is that

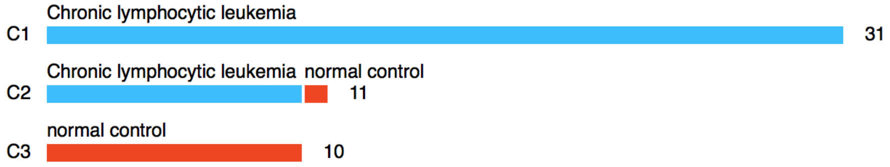
- classification is about discovering the rule(s) for discriminating between two predefined groups, and
- clustering is about finding groups in your data.

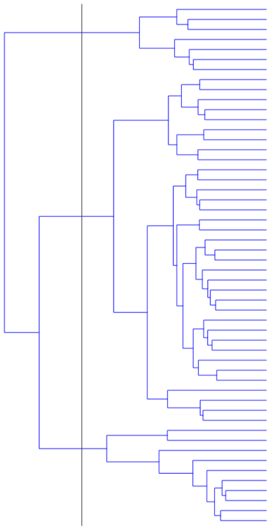
These are very different tasks. There may be data (like the Greko-Swedish study) in which both give sensible results. In general, you want to use one or the other, depending upon what kind of data you have and what the goals of your analysis.

So one student discovered this ...

Last year, one student in one of our classes discovered something interesting that we didn't know when we gave the homework. He writes: *A potential explanation for this poor assignment is that there are two types of chronic lymphocytic leukemia, not one. Recalculating k-means clustering with three classes, instead of two, yields results that are consistent with this hypothesis. This time all controls fall into their own cluster (again, except the one outlier – a sick patient misclassified as a healthy control), and, as before, the leukemia samples segregate into two separate clusters*.

It indeed works like magic:

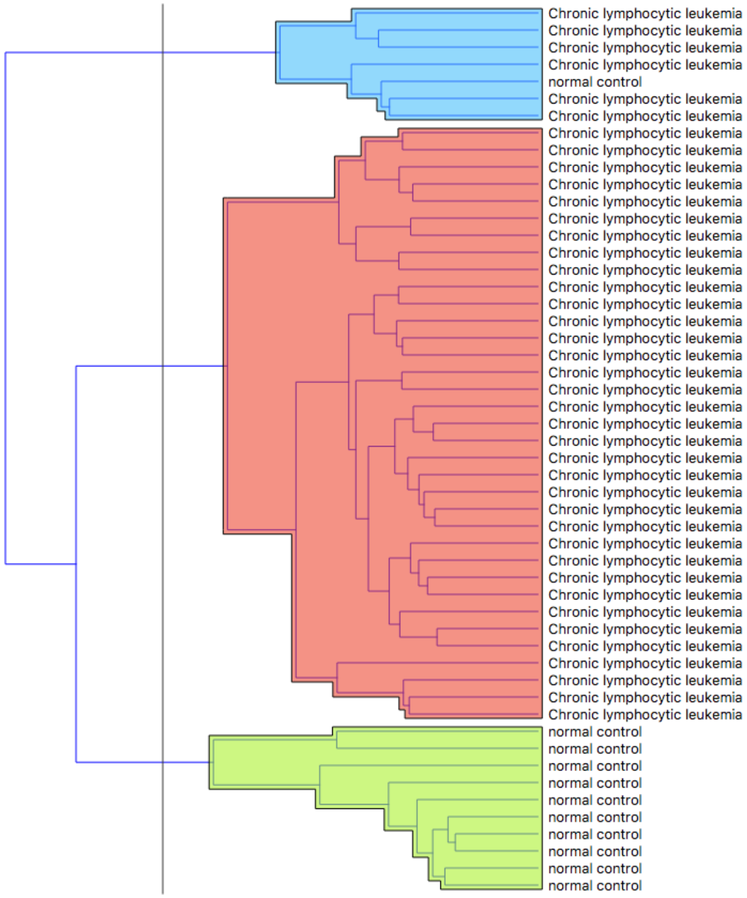




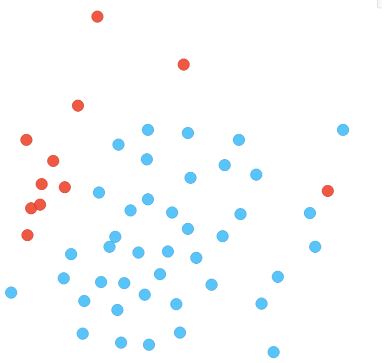
K-means thus discovers that there are three clusters, one with *normal control* and two with CLL.

This works even better in the hierarchical clustering (see the Figure on the right). The dendrogram indicates indicates that it may be a better to split the data into three clusters.

In this case, we know the classes of the data instances. Hence, let's continue with realistic analysis and add the class labels to the clustering.



We can see that if we create two clusters, one is going to contain the leukemic cells and the other a mixture of leukemic and control. If we create three clusters, the almost perfectly correspond to the two groups, except that the CLL cells are split into two groups - just as the student hypothesized.



There's a misclassified ("misclustered") data instance in the first cluster. It is indeed too similar to some CLL cells - see the MDS on the right.

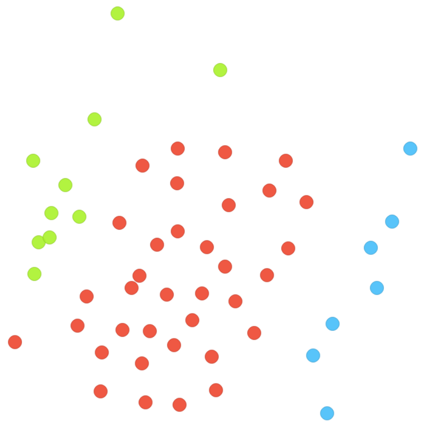
Then it gets even better. He continues: *Checking the literature, we see that our hypothesis of two disease subtypes is actually correct. There are two subtypes of chronic lymphocytic leukemia known to the field (10.1056/NEJMra041720):*

> *When chronic lymphocytic leukemia (CLL) was last reviewed in the Journal, it was considered a homogeneous disease of immature, immune-incompetent, minimally self-renewing B cells, which accumulate relentlessly because of a faulty apoptotic mechanism. In the past decade, these views have been transformed by a wealth of new information about the leukemic cells. CLL is now viewed as two related entities, both originating from antigen-stimulated mature B lymphocytes, which either*

> *avoid death through the intercession of external signals or die by apoptosis, only to be replenished by proliferating precursor cells.*

One thing that we still don't know is whether the two clusters indeed correspond to the two types of CLL, as described in the paper. If this is true, then clusters that you've got represent two subtypes of CLL instead of CLL and the control. Even more, it may be that one of the CLL subtypes is more similar to the control than to the other CLL group. On the other hand, this can indeed be related to the greater size and spread of the prevalent CLL group. We can check this in MDS - assuming it gives a correct impression about which cells are close to which in the data.

Indeed, coloring the cells by clusters shows this.



If we trust the MDS projection, the distinction between the CLL cells in C1 (blue) and the CLL cells in C2 (red) is larger than the distinction between the CLL cells in C2 (red) and the control in C3 (green).

A student thus provided a great example of how to use the example data we used in our class to make a real discovery that he could have potentially published (after finding another, not just visual confirmation), if he did this 10 years earlier!