Introduction to Data Science, WMCS16002, semester 1a 2020

Assignment Classification

Due October 8, 2020 12:00:00 (midday)
Submit by creating a pull request on GitHub

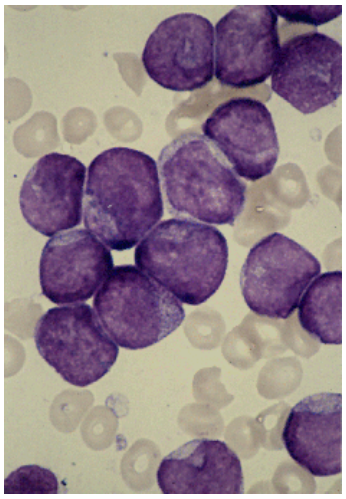
1 Pen&Paper (Example Exam Question) (15 P):

Consider the following set of training examples for a binary classification problem and potential binary splits on the attributes:

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

(a) What is the entropy of this collection of training examples with respect to the positive class?
(b) What are the information gains (using Entropy as test condition) of a_1 and a_2 relative to these training examples?
(c) For a_3, which is a continuous attribute, compute the information gain for every possible split.
(d) What is the best split (between a_1 and a_2) according to the classification error rate?
(e) What is the best split (between a_1 and a_2) according to the Gini index?

2 Classification in Practice



In this assignment, we are using the Flow cap cytometry dataset. For information about Flow cytometry please refer to `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003365`. The dataset is from DREAM6/FlowCAP2 Molecular Classification of Acute Myeloid Leukemia (AML) Challenge, 2011.
The file `featuresFlowCapAnalysis.csv` contains the extracted features of 359 subjects. Labels of the first 179 records can be found in `labelsFlowCapAnalysis.csv`. There are healthy and AML patients (class 1 and 2). Your task is to predict the labels of the latter 180 unlabelled subjects.

Tip: As disclosed in the original challenge, there are 20 AML patients in the test set, so see if you can catch them all.

2.1 Descriptive and Exploratory Analysis (20 P)

Apply the knowledge you gained in the lectures about visualization, preprocessing, descriptive and exploratory analysis.

(a) **Investigate the features** (5/20P): Use Eigen value decomposition as well as perform a univariate analysis (ANOVA) (minimum requirement). How many features explained a certain percentage of the variance in the dataset? Find out whether all the features are important and if not which features might be more promising to solve the problem. Use boxplots to visualize the top features and a selection of less significant ones.

(b) **Get a holistic view on the data** (5/20P): Embed the data using the first 2-3 principal components from PCA and a state-of-the-art non-linear method like tSNE to get an idea about the complexity of the problem. Do you observe certain characteristics which could become a problem for classification? Do you find considerable differences of the distribution of the training data and test data?

(c) **General impact of preprocessing** (10/20P): Use the non-linear embedding (e.g. tSNE) to investigate the impact of different preprocessing to the distribution of the classes. Do you observe positive or negative effects when you perform preprocessing (feature selection, z-score transform, PCA with leading x eigenvectors) on the data?

Please refer to and provide explanation of your plots in the report. And the scripts you submit should be such that when we execute your code we are able to obtain the same plots which you have put in the report (e.g. fix the seed of the random generator before tSNE so you get the same plot).

2.2 Experimentation to find the best prediction (65 P)

By now you know that the dataset is imbalanced and not absolutely trivial. Welcome to the world of real world biomedical data! Now experiment with at least 2 classification methods (namely kNN and DT), strategies to handle the imbalance and different preprocessing in a cross validation setting. Investigate if an ensemble improves the single classifier models and

explain the 5-10 best performing experiments and their configuration. (You can mention other experiments you tried, which had less performance as well) The minimal requirements for experimentation are listed below:

(a) **Base-line experiments** (20/65 points): Perform N-fold cross validation experiments (that includes preprocessing if needed for a method) with kNN and Decision Trees. Evaluate classwise training and validation errors for different preprocessing and different settings (k for kNN and pruning, kind of splits, etc. for DT). Use knowledge from your descriptive exploratory analysis to improve the performance. Do you observe overfitting? How do you tackle it? KNN as well as Decision Trees are very intuitive interpretable classifiers. Can you confirm certain findings from the exploratory analysis?

(b) **Ensemble** (20/65 points): Using your most promising experimental settings from the above to build an ensemble classifer and perform classification. Explain which models (combination of which classifier and which experimental settings) you selected. Compare to random forests, an ensemble build with decision trees on random subsets of the features. (Of course you can still limit the number of features before-hand removing the least promising ones to increase performance)

(c) **Summarization of your experiments** (25/65 points): Explain the experiments you performed and your conclusions. In addition, summarize the experiments you have performed, settings and combinations tried, in an hierarchical tree structure (see for example Fig. 1)[2], a flowchart, or tabularize them. You are supposed to try out different experimentals settings (for instance, how many features, which value of neighbors in knn, etc). Remember to tabularize the top 5 classifier experimental settings in your report.

(d) **Submit** your predicted labels in a separate csv file named `Team_x_prediction.csv`. Include in the summary which experiment those prediciton are stemming from.

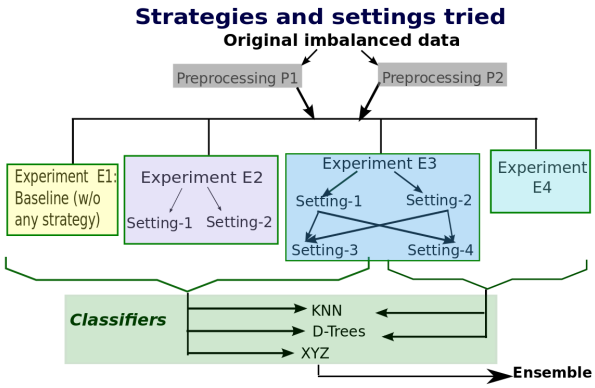


Figure 1: Tip: a schematic of experiments performed might help you to structure your report.

2.3 Bonus (10 points)

The team(s) with the best performance on the 20 AML patients of the 180 unlabelled subjects will be awarded full bonus point. If you think that the classifiers mentioned in this assignment cannot predict all the 20 AML subjects easily feel free to test other techniques and explain your choices. The teams with lesser performance will get a fraction of the possible points based on the

[2]Fig 1 is just a suggestion. If you have better ideas to summarize your work you are welcome to do so.

effort and argumentation of their experimentation. You are free to use any classification method as long as you have also used the ones mentioned in class.

Tip: The paper published for this challenge leads you to the best performers in case you are interested (uploaded in Nestor). Our team was one of the best performers of that challenge and the approach has been published in PLOS ONE (2013).