Text Mining
Final Exam

02-october-2019

1 Building a simple spam classifier (Weka)

1. Using the file *spamDataTxt.csv* (described in the appendix A), create two *arff* files, one containing a training set and the other a test set.
2. Use the training set to create a decision tree spam classifier.
3. Apply your classifier to the test set.
4. Give the confusion matrix and the accuracy of the classifier.

2 Processing corpuses (R)

Let us go back to the file *spamDataTxt.csv*.

1. Create a corpus with the messages considered as spam.
2. Clean this corpus.
3. Show the corresponding wordcloud.
4. Repeat the previous questions with the set of all messages.
5. Can we notice differences between the two wordclouds ? If yes, can these differences help to give us rules characterizing spam messages ?
6. Some methods use TF-IDF values to build spam classifiers.
   (a) How can we use intermediate results of the previous questions to compute these values ? (you don't have to do that, you have just to explain how to do it).
   (b) What do you think about these methods ?

3 Text Clustering applied to spam detection (R)

We consider the list of messages described in the file *spamDataNum.csv* (see appendix B).

1. Explain why the messages' description used in this file can give more accurate spam classifier than those of the previous questions ?
2. Apply the algorithm kmeans with $k = 2$ to this data.
3. Compare the clusters and the classes. What 'accuracy' do we obtain ?
4. Normalize the data and apply again kMeans. How do you explain the difference between the two accuracies ?
5. How can we use your model to 'classify' a new message ?
   We define the importance of each variable Xi by the ratio $imp(Xi, CSpam)$ described in appendix C.
6. Compute and plot the importance values.
7. What are the most three important variables ?

A The file *spamDataTxt*

This file contains 200 messages. Each message is described by its content and its class (1 for spams and 0 for non spams).

B The file *spamDataNum*

1. Each variable $X1, ..., X48$ represents the frequency of a given word in the mail.
   — Exemples : $X23$ represents the frequency of the word '*000*', $X24$ represents the frequency of the word '*money*', ...
2. Each variable $X49, ..., X54$ represents the frequency of a given character in the mail.
   — Exemples : $X53$ represents the frequency of the character '*$*', $X52$ represents the frequency of the word '*!*', ...
3. $X55$ represents average length of uninterrupted sequences of capital letters.
4. $X56$ represents length of longest uninterrupted sequence of capital letters.
5. $X57$ represents total number of capital letters in the e-mail.
6. $CSpam$ represents the class : 1 for spam and 0 for non spam.

C Measuring the correlation between the class and an explanatory variable

Let us consider a dataset described by a numerical variable X and a categorical variable Y. Let us suppose that Y have two possible values (classes) C_1 and C_2.

Let us consider the following values :

1. n is the cadinality of the dataset.
2. $n1$ is the cadinality of the subset with $Y = C_1$.
3. $n2$ is the cadinality of the subset with $Y = C_2$.
4. m is the mean of X.
5. $m1$ is the mean of X for the subset corresponding to $Y = C_1$.
6. $m2$ is the mean of X for the subset corresponding to $Y = C_2$.
7. v is the variance of X.

8. *vint* is defined as follows :
   — $vint = \frac{1}{n}(n1 * (m1 - m)^2 + n2 * (m2 - m)^2)$
9. We measure the link between X and Y by the following ratio :
   — $imp(X, Y) = \frac{vint}{v}$.