11/08/2015

COMP4650/6490 Document Analysis Assignment 2 – ML

Marian-Andrei RIZOIU

CECS, ANU, Canberra, Australia.

Abstract: *This document will provide the details concerning the 2nd* ***Document Analysis*** *assignment.*

Main details

WHAT?	TYPE? HOW MUCH? WHERE? WHEN?
Maximum marks	10
Programming language	Java (only)
Assignment questions	Post to the Wattle Discussion forum only. No emails will be answered
Deadline	Q1: Lab on **Mon/Tue/Thu 22/23/25 August**, depending on your enrolled lab. No late days allowed, automatic 0 for Q1 if fail to attend/demonstrate. Q2-Q4: **Thu 25 Aug**, 11:59pm (online via Wattle)

Marking scheme:

- *Written*: Full marks given for a formulation that provides a well-reasoned and succinct response to the question that addresses all requested points. There may be more than one answer for each question that achieves full marks.
- *Code*: Full marks given for working, readable, efficient, commented code that performs well on the test case given in lab.
- *Academic Misconduct Policy*: All submitted written work and code must be your own (except for any provided Java starter code, of course) – submitting work other than your own will lead to both a failure on the assignment and a

referral of the case to the ANU academic misconduct review procedures: ANU Academic Misconduct Procedures

Electronic submission (only):

All written questions should be in a file `ANSWERS_yourname.pdf`. MS Word or other document formats are not accepted.
Please submit `ANSWERS_yourname.pdf` and Q1 source code (only the code you have written – no other code or data) zipped into a single file `assignML_yourname.zip` with a `README.txt` stating what each file does.

ML Programming

Q1 [5 pts]. Clustering (checked in lab).

Your task is to code up K-means clustering as discussed in lecture and apply it to the *Assignment 2 Q1*Data posted to Wattle (do not unarchive it, use it just like during the lab). Most design decisions are yours but you should pay careful attention to the notes provided below in order to receive full marks.

In lab you will be asked to run your code on a *new* data set, similar to the *Assignment 2 Q1 Data*, but containing different documents. In lab, you will be told what to set K to (*i.e.*, the number of clusters) and the code should display the *top 5 document filenames* for each of the K clusters found. The grader will both inspect the quality (purity) of the clustering (2.5 pts), as well as the *efficiency* of the code you write and your explanation of your design choices (2.5 pts).

Notes:

- The warm-up ML lab exercise using Naive Bayes will not be graded, it is intended to reinforce general principles from the machine learning lectures covering training, testing, feature selection, and overfitting. You need not understand or reuse the Naive Bayes code, but feel free to reuse any parts of it you find useful for this assignment. Note that the Naive Bayes code does not use an efficient, sparse document representation that is required for full credit on this assignment (see next point).

- Be careful when representing your documents – you should use a sparse (*e.g.*, Java HashMap) representation of term frequencies to save space. Points will be deducted for inefficient storage approaches.
- Be careful when computing similarity metrics between two term frequency (TF) vectors to do it efficiently. For example, if iterating over entries in two TF vectors represented as HashMaps to compute their inner product, you can just iterate over the smaller key set of the two HashMaps… why? Points will be deducted for inefficient calculations with TF vectors.
- If your algorithm randomizes its initialization, we will ask you to run a few times and we will grade the best clustering found.
- In the past, students have achieved good results using the standard centroid recomputation step of K-means but replacing *Euclidean distance* with *cosine similarity* for the reassignment step (note: minimizing distance is the same as maximizing similarity). This variant is known as spherical K-means.
- You might experiment with stopword and/or other frequency-based term selection methods to see what provides optimal performance.

ML Written

Q2 [1.5 pts]. Clustering with Naive Bayes.

(1.0 pts) In pseudocode, modify the K-Means algorithm given in the lecture slides (flat clustering, slide 30) to replace its implicit centroid-based classifier (i.e. Rocchio) with Naive Bayes as follows:

1. *Recomputation:* Rather than representing each cluster with a centroid (as in Rocchio), train a Naive Bayes classifier instead by calling `TrainNB(D)`. Be clear to show how the data $D=\{(y,\vec{x})\}D=\{(y,x\rightarrow)\}$ is constructed, *i.e.*, for each $(y,\vec{x})\in D(y,x\rightarrow)\in D$, what is the label yy and the corresponding feature vector $\vec{x}\ x\rightarrow$?
2. *Reassignment:* Assign each data point to its highest probability cluster according to the Naive Bayes classifier. This assignment should involve an arg max.

Your pseudocode should be mathematically precise – do not use English to describe computations. For simplicity, you may assume K=2K=2.

(0.5 pts) Considering that both centroid-based classification (i.e. Rocchio) and Naive Bayes are linear classifiers, would you expect their clustering counterparts – the original K-means and this Naive Bayes variant – to *always* converge to the same answers (given the same cluster initialization)? Why or why not? Answer in one sentence.

Q3 [1.5 pts]. Logistic Regression.

An experiment is conducted to find the price that users of online social network put on their privacy. A set of 20 male users are offered a sum of money to allow a third party full access to their Facebook, Whatsapp, Tinder and email account. We vary the sum of money and we count how many users would "take the deal".

The data are given below:

Offered sum of money (in thousands)	1	2	4	8	16	32
log2(log2(offered sum money))	0	1	2	3	4	5
Males taking the offer	1	4	9	13	18	20

Answer the following questions. All answers should be motivated, the inner calculations shown and*precise*. Long unrelated answers will not be graded.

1. In general, describe the conclusions you can draw about how males value privacy, based on the observed data.
2. What is the observed proportion of males selling their privacy when offered 16,000$?
3. What are the observed odds of males selling their privacy when offered 16,000$?

4. A logistic regression with the following equation was trained: $\hat{y}_i = \log\left(\frac{p_i}{1-p_i}\right) = -2.81856 + 1.25895 \times \log_2(\text{Offered\_sum})$
   Use this equation to predict the log-odds for a sum of 16. What are the predicted odds? What is the predicted proportion of males taking the deal?

5. If we increase the sum from 16 to 32, by what multiple will the predicted odds increase?

   The experiment is run again with batches of 20 females. The number of females who take the deal is again recorded and a logistic regression is run. Below is a plot of the fitted logistic regressions for males and females.

6. According to the plot, is there much of a difference between how males and females value their privacy? Explain briefly.
7. According to the plot, if you wish to obtain the data of 50% of the males what sum should you propose?
8. According to the plot, if you wish to obtain the data of 50% of the females what sum should you propose?

9. When modeling the relationship between a numerical predictor variable and a binary response, why is the logit transformation a good idea?

Q4 [2 pts]. Hierarchical Classification.

Imagine that classes are arranged in a known tree hierarchy such that every document is classified into one node (internal or leaf) of this tree hierarchy. (See the Yahoo! and MESH examples in the clustering slides for examples of such hierarchies.)

(1.0 pts) Given training documents that have been previously assigned to their correct node in the tree hierarchy, briefly describe one way you could use a multiclass classifier (for kk mutually exclusive classes) for supervised training for this hierarchical classification task.

Also briefly describe how a new (previously unseen) test document could be classified into this hierarchy once the hierarchical classifier has been trained.

(1.0 pt) Assume the classification tree has nn nodes (combined internal and leaf nodes) with a maximum branching factor of bb at each node, a maximum tree depth of dd, and a total of mm labeled training examples. Assume the term vocabulary size is vv (number of unique terms observed in the data) and that your multiclass classifier is Rocchio. Answer the following questions:

- What is the training time complexity of your algorithm over all of the training data? Provide brief justification for your answer.
- What is the testing time complexity of your algorithm for a single document? Provide brief justification for your answer.
- Would your algorithm be computationally feasible for large-scale training and testing in practice? Why or why not? Answer as succinctly as possible. (Recall that large-scale computational feasibility typically requires time complexity that grows at most linearly in *each* of the quantities.)