CIS600 Fundamentals of Data and knowledge Mining

HW2: Decision Tree and Artificial Neural Network

Due: 9pm EST, Thursday, May 9th, 2019

Problem Set 1: Authorship Attribution

For this problem, you are going to use Decision Tree induction algorithm to solve a mystery in history: who wrote the disputed essays, Hamilton or Madison?

1. About the Federalist Papers

Quote from the Library of Congress http://www.loc.gov/rr/program/bib/ourdocs/federalist.html

The Federalist Papers were a series of eighty-five essays urging the citizens of New York to ratify the new United States Constitution. Written by Alexander Hamilton, James Madison, and John Jay, the essays originally appeared anonymously in New York newspapers in 1787 and 1788 under the pen name "Publius." A bound edition of the essays was first published in 1788, but it was not until the 1818 edition published by the printer Jacob Gideon that the authors of each essay were identified by name. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

2. About the disputed authorship

The original essays can be downloaded from the Library of Congress. http://thomas.loc.gov/home/histdox/fedpapers.html

In the author column, you will find 74 essays with identified authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay. The remaining 11 essays, however, is authored by "Hamilton or Madison". These are the famous essays with disputed authorship. Hamilton wrote to claim the authorship before he was killed in a duel. Later Madison also claimed authorship. Historians were trying to find out which one was the real author.

3. Computational approach for authorship attribution

In 1960s, statistician Mosteller and Wallace analyzed the frequency distributions of common function words in the Federalist Papers, and drew their conclusions. This is a pioneering work on using mathematical approaches for authorship attribution.

Nowadays, authorship attribution has become a classic problem in the data mining field, with applications in forensics (e.g. deception detection), and information organization.

The Federalist Paper data set (Disputed_Essay_data.CSV) is provided. The features are a set of "function words", for example, "upon". The feature value is the percentage of the word occurrence in an essay. For example, for the essay "Hamilton_fed_31.txt", if the function word "upon" appeared 3 times, and the total number of words in this essay is 1000, the feature value is 3/1000=0.3%

Now you are going to try solving this mystery using decision tree induction algorithm. Document your analysis process and draw your conclusion on who wrote the disputed essays. You will need to separate the original data set to training and testing data for classification experiments. Describe how you create training and testing datasets. After building the classification model, apply it to the disputed papers to predict the authorship. Experiment with different decision tree model parameters in order to output the best performance model and discuss your fine tune process and choice of model performance evaluation methods and metrics.

Problem set 2: Educational Data Mining

As technology is now shaping the future of education, educational mining is a novel emerging field which concerns with developing algorithms to gather, analyze and discover hidden patterns in the educational data. There is a tremendous increase in the number of educational institutions adopting e-learning systems to leverage technology in classrooms. This streamlines the collection of educational data for analysis.

A kaggle dataset on student academic performance (available through https://www.kaggle.com/aljarah/xAPI-Edu-Data and also included as the attachment in this instruction) is gathered to identify the influential factors for students' performance. To predict the students' performance, the collected data was organized into four kinds of features: demographic, academic background, parents' participation on learning process and behavioral features. The demographic features consisted of demographic details of the students like gender, nationality, place of birth etc. The section, grades and semester details of the students were included under the academic features and behavioral features consisted of fields demonstrating students' engagement with the learning management system like viewed announcements, interaction with discussion groups, resources etc. To analyze the students' performance, the target "Class" attribute was discretized into orindal values based upon students' grades. Hence, we had three categories of student classes: High/H, Low/L and Medium/M.

A full list of attributes is provided as below:

Features Category	Feature	Description
Demographical Features	Nationality	Student nationality
	Gender	The gender of the student (female or male)
	Place of Birth	Place of birth for the student (Jordan, Kuwait, Lebanon, Saudi Arabia, Iran, USA)
	Parent responsible for student	Student's parent as (father or mum)
Academic Background Features	Educational Stages (school levels)	Stage student belongs such as (primary, middle and high school levels)
	Grade Levels	Grade student belongs as (G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12)
	Section ID	Classroom student belongs as (A, B, C)
	Semester	School year semester as (First or second)
	Topic	Course topic as (Math, English, IT, Arabic, Science, Quran)
	Student Absence Days	Student absence days (Above-7, Under-7)
Parents Participation on learning process	Parent Answering Survey	Parent is answering the surveys that provided from school or not.
	Parent School Satisfaction	This feature obtains the Degree of parent satisfaction from school as follow (Good, Bad)
Behavioral Features	Discussion groups	Student Behavior during interaction with Kalboard 360 e-learning system.
	Visited resources	
	Raised hand on class	
	Viewing announcements	

Discuss your data preprocessing steps and model the student classes using artificial neural network (ANN) algorithm. Experiment with different ANN architectural parameters (e.g. number of hidden layers, number of nodes within each layer) as well as model parameters (activation and loss functions, regularization, epoch/batch size, etc.). Evaluate and report the performance of your ANN models.

Grading rubrics:

1. Are the data preprocessing and preparation sufficient for modeling?
2. Are the models constructed correctly?
3. Is the result analysis conclusion convincing?
4. Is sufficient details provided for others to repeat the analysis?
5. Does the analysis include irrelevant content?

Please submit your rmarkdown document together with knitted report (in either PDF or html format) if you program in R or submit your Jupyter Notebook with exported report (in either PDF or html format) if program in Python.