Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2023

Teaching Assistants

Zahra Zanjani

Atiyeh Moghadam

Zahra Akhlaghi

Assignment (3)

Outlines. In this assignment, Naïve bayes and logistic regression are noticed.

Deadline. Please submit your answers before the end of November 29th in courses.aut.ac.ir. Other methods like sending via email or in social networks are not accepted and will not be considered.

Assignment Manual

Delay policy. During the semester, you have extra 10 days for submitting your answers with delay. Mentioned time is for all assignments. After that, for each day of delay you loss 20% points of that assignment. After 4 days you miss all points and any submit doesn`t acceptable. Remember that saving this time doesn`t have any extra point.

Sharing is not caring. Students are free to discuss and share their ideas about problems with others. But sharing source codes, solutions, answers and other results is not allowed and based on university`s rule, both sides will be graded zero.

Problems are waiting you. Some problems are required to be implemented within a programming language and obtain some charts, images, results, etc; then discuss about it. These types of questions are tagged by #Implementation. Some other problems are required to be solved or computed by hand or research about them. These types of questions are tagged by #Theorical. You are not allowed to use programming language or other technical tools to answer theorical problems.

Report is the key. All students' explanations, solutions, results, discuss and answers must be compacted into a single pdf report. A clean and explicit report is expected and may followed by extra pts; so, you may need to write any related detail or experience during the solving problems. Report file should started within a cover page that it includes course and assignment information as well as identical details like name, student number and email address. Second page should be table of contents that indicates student`s answer to each question. Please repeat your name and student number in left side of footer in other pages. Also, you are free to write in Persian or English. If typing is bothering you, so write in a paper and put its picture with acceptable readability quality in report file.

Organize the upload items. Students should upload their implementation source codes as well as results and report. You should upload a single .zip file with the following structure:

ML_01_[std-number].zip
- Report
  - ML_01_[std-number].pdf
  - [other material and results]
- Source codes
  - P[problem-number]_[a-z].py
  - P[problem-number]_[a-z].ipynb
  - …

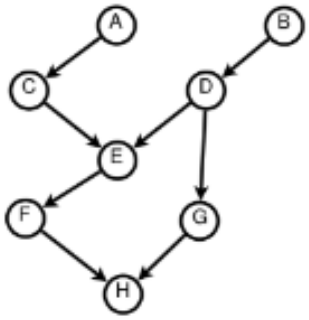
Python is the power. Students are free to use any programming language like python, matlab, C++ , etc. However it is recommended strongly to use python in jupyter notebook environment; so, you may need to upload your .py or .ipynb sources.

Feel free to contact. If you have any question or suggestion, need guide or any comment be comfortable to ask via email as well as Telegram group.

Problem 1: Why and how (15 pts)

#Theorical

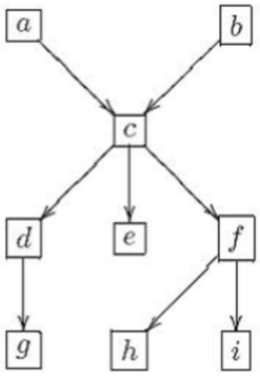
a) Is Naive Bayes exclusively limited to discrete data, or can it be utilized with other types of data as well? If it can be used with other data types, which ones does Naive Bayes have the ability to handle?
b) What are the reasons behind Naive Bayes having a higher bias yet lower variance in contrast to logistic regression?
c) Under what circumstances does logistic regression demonstrate superior performance over Naive Bayes, and what are the underlying reasons for this? Please refer to the paper authored by Professor Andrew Ng and Professor Michael I Jordan to substantiate your explanation.
d) provide an overview of the pros and cons associated with employing a Bayesian-based model for spam filtering.
e) Consider the Bayesian network given below:



Answer the following statements about conditional independence. For each one, give a brief justification involving paths.

i. Given {} (i.e., the empty set), is C conditionally independent of G? (Because the set is empty. we say "unconditionally independent," or just "independent.")
ii. Given D. is B conditionally independent of F?
iii. Given F, is C conditionally independent of G?

f) Consider the following Bayesian network:



at node a:

Prob(a)
0.7

at node b:

Prob(b)
0.8

at node c:

a	b	Prob(c)
0	0	0.1
0	1	0.6
1	0	0.7
1	1	0.9

at node d:

c	Prob(d)
0	0.6
1	0.8

at node e:

c	Prob(e)
0	0.1
1	0.7

at node f:

c	Prob(f)
0	0.8
1	0.3

at node g:

d	Prob(g)
0	0.15
1	0.75

at node h:

f	Prob(h)
0	0.25
1	0.65

at node i:

f	Prob(i)
0	0.35
1	0.55

i. Compute P(f=1|a = 1).
ii. Compute P(a = 1 | f = 1).
iii. Compute P(c = 1 | g = 1, h = 1).

Problem 2: Detection of Hate Speech in Tweets(25 pts)

#Implementation

In this task, you will be identifying hate speech within tweets. For simplicity, we define a tweet as containing hate speech if it exhibits a racist sentiment. Thus, the objective is to distinguish racist tweets from non-racist ones.

Formally, you are provided with a training set of labeled tweets, where label '1' indicates a racist tweet and label '0' signifies a non-racist tweet. Your goal is to make predictions on the test dataset utilizing a Naive Bayes classifier.

a) Begin by loading the dataset. Is the dataset balanced? How does the imbalance impact Naive Bayes classification? Proceed to partition the dataset into training and testing sets.
b) What are the most important preprocessing steps for text data? Apply these steps to your dataset. [You are free to utilize the NLTK library.]
c) **Implement** a Naive Bayes classifier to classify tweets into hate speech or non-hate speech. Explain the underlying theory of the model and the steps involved in classification. Is Laplace smoothing necessary for this particular problem?
d) Classify the test tweets and evaluate the performance of your model using recall, precision, and F1 score. In this context, which metric, precision or recall, holds greater significance?

Problem 3: Football Match Result Prediction (35 pts)

#Implementation

Predicting the outcome of football matches is an exciting task that garners attention from fans and analysts alike. However, accurately predicting match results is challenging and requires a comprehensive understanding of historical match data, team statistics, and the application of advanced machine learning models. In this assignment, we explore the captivating domain of football match score prediction.

a) Load the match.csv dataset and create a home team result column based on the home team goal and away team goal columns. The result of a game can be categorized as a loss, draw, or win.
b) Apply proper feature engineering techniques by dropping irrelevant features and creating new ones that may enhance the predictive power of the model.
c) In each season, three new teams join the Premier League, and their historical data is limited or nonexistent. How can we predict the results for these teams? Consider strategies to address this challenge.
d) Split the dataset to train and test and reserve the last season for test.
e) Utilize the Naive Bayes and Logistic Regression algorithms to predict match results.
f) Provide an explanation of Bernoulli Naive Bayes and Multinomial Naive Bayes. Discuss the types of problems where these algorithms can be applied. Can they be used to predict match results? If applicable, apply them.
g) Compare all the models used and report the accuracy achieved by each of them.

Allowed Libraries:

Feel free to use any preferred machine learning library to complete this assignment.

Ideas to Improve Results:

You are encouraged to perform any feature engineering techniques and utilize features from the team and player attribute tables. Here is one basic idea:

Extract useful features such as each team's performance probabilities (e.g., probability of winning, drawing, or losing at home or away). Consider the influence of a team's performance from a decade ago to be less significant compared to recent seasons.

Dataset Description:

The dataset used in this assignment contains tables with information about football matches, teams, and players. It includes the following tables:

i. **Match:** Contains details about individual matches, such as team IDs for the home and away teams, and the number of goals scored by each team. It also includes player IDs for the starting lineups.
ii. **New Teams:** Lists the details of new teams that have joined the league, along with the seasons they joined.

iii. **Player:** Contains information about individual players, including unique identifiers and attributes such as the player's name, birthdate, height, and weight
iv. **Player Attributes:** Includes attributes and statistics associated with players, such as overall rating, potential, preferred foot, and work rates.
v. **Team:** Provides information about the participating teams, including API and FIFA IDs, as well as their long and short names.
vi. **Team Attributes:** Offers additional attributes and characteristics related to teams. These attributes provide insights into the playing style and strategic approach of the teams.

Problem 4: Classify rooms as messy or clean (25 pts)

#Implementation

The dataset provided is tailored for the binary classification task of distinguishing between messy and clean rooms. In this context, logistic regression serves as the chosen classification algorithm to efficiently categorize rooms based on their cleanliness status. Your goal is to do this classification using logistic regression.

a) The size of the images may be different. Resize them. Then normalize the images. Determine the target of each image.
b) Split the data into train and test.
c) **Implement** logistic regression and compare the results and runtime with logistic regression implemented in scikit-learn.
d) Find the best probability threshold for the training data on your model and report the accuracy and confusion matrix.
e) Preprocess the images in the part5 folder, then feed this image into your classifier and make predictions. Finally, label each image with its probability in the photo.