Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2022

Teaching Assistants

Mohsen Ebadpour(head) (M.Ebadpour@aut.ac.ir)

Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)

Mohamadreza Jafaei (Mr.Jafaei@aut.ac.ir)

Assignment (3)

Outlines. In this short assignment, Gaussian Naïve Bayes and Logistic Regression are noticed. Be careful that this assignment is short and easier than two previous and the next; so, its deadline is short as well.

Deadline. Please submit your answers before the end of December 1st 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 7 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_03_[std-number].zip
    Report
        ML_03_[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 (30 pts)

#Theorical

a) Do a short study about setting a suitable threshold in Logistic Regression (LR) for binary classification. (Describe strategies)

b) How LR is used for multi-class classification task? Generalize your answer to any binary classifier.

c) Can LR obtain a non-linear decision boundary? Explain your answer.

d) Consider to below table, apply Naïve Bayes classifier for following probability: P(Target=1| F1=1 ∧ F2=1 ∧ F3=0)

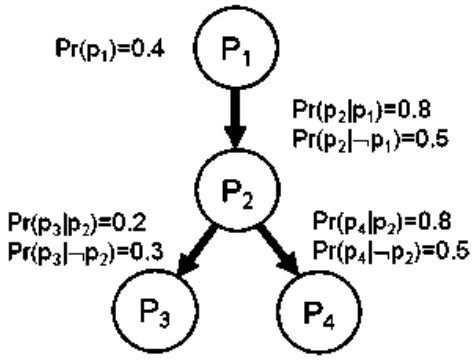
#ID	F1	F2	F3	Target
1	1	0	1	1
2	1	1	1	1
3	0	1	1	0
4	1	1	0	0
5	1	0	1	0
6	0	0	0	1
7	0	0	0	1
8	0	0	1	0

e) Consider The Table Below and Answer the Question:

ID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125k	No
2	No	Married	100k	No
3	No	Single	70k	No
4	Yes	Married	120k	No
5	No	Divorced	95k	Yes
6	No	Married	60k	No
7	Yes	Divorced	220k	No
8	No	Single	85k	Yes
9	No	Married	75k	No
10	No	Single	90k	Yes

e.1) $P(Cheat = 'No' \mid Marital\ Status = 'Married')$?
e.2) $P(Income = 120 \mid Cheat = 'Yes')$?
e.3) if X = (Refund = 'No', Marital Status = 'Married', Income = 120) then P(X | 'Yes') and P(X | 'No')?

f) Given the network below, calculate marginal and conditional probabilities $Pr(\sim p_3)$, $Pr(p_1 \mid p_2, \sim p_3)$



Problem 2: Breast cancer **(40 pts)**

Implementation

In this problem, LR and Naïve bayes classifiers will be used to detect the breast cancer. Load dataset by sklearn "load_breast_cancer" function.(split ratio: 2:8)
(https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer)

a) Use the sklearn`s built-in function and train a LR model to detect cancer; Also, explain the parameters that used in built-in model and their effects. Tune the model and get better result as you can; report your tries and confusion matrix as well as accuracy.

b) Repeat previous part for gaussian naïve bayes classifier and compare the results. Which one performs better? How much and why?

c) A classic A classic simple technique for feature selection as pre-processing is calculating the direct importance of each feature on target; then, keeping some important features and dropping the others. In this part, you are asked to simulate the mentioned strategy.

   First, train the LR model for each feature and obtain 30 related models (For each model only one feature with the target is used).

   Second, calculate the AUC of each model and select the top 20 related features.

   Third, Using the obtained 20 features, Train the best LR and naïve Bayes classifiers that you obtained in parts (a) and(b); then, compare the results. Also, Report the results in each step and model. Is the feature selection worth it? Why?

   Note: You are allowed to use any libraries or functions.

d) Another interesting technique is feature extraction using LR. In this strategy, the goal is to extract the normed features that are informative about the target. For this aim, use the 30 trained LR models from the previous part. For each feature, pass the feature`s value to its related model and obtain its probability (using predict_proba or predict_log_proba functions); now replace the feature`s value with its probability. Do the mentioned procedure for each feature and sample. Now, you have a probabilistic dataset. As you learned, the Bayes classifier gets along with probabilistic data. Use the new dataset to train Gaussian and Multinomial Naïve Bayes and detect cancer. Compare the results to previous parts and discuss them. Also, report AUC, and confusion matrix.

Problem 3: Email Spam Classifier **(30 pts)**

Implementation

In this exercise, you will train the Naïve bayes model to classify spam and non-spam emails. You should follow three main steps in this question:

a) Explain the most important methods of text data pre-processing and use them for your dataset (you can use a library like NLTK). Visualization can also be a part of pre-processing.

b) Train a Naïve bayes model that classifies the email as either spam or non-spam (Note: you cannot use the library). explain how this model works in theory.

c) Finally, test the model's performance. (Recall Score, Precision Score, and F1 Score, etc.)