Amirkabir University of Technology
(Tehran Polytechnic)

Applied Machine Learning Course By

Dr. Nazerfard

CE5501 | Spring 2023

Teaching Assistants

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

Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)

Ali Amirian (ali.amiryan@aut.ac.ir)

Ghazaleh Gholinejad (Ghazaleh.gholinejad@aut.ac.ir)

Assignment (3)

Outlines. In this assignment, Logistic Regression and Naïve Bayes are noticed (4 Questions).

Deadline. Please submit your answers before the end of May 12^{nd} 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:

```
AML_03_[std-number].zip
        Report
                AML_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 (17.5 pts)

#Theorical

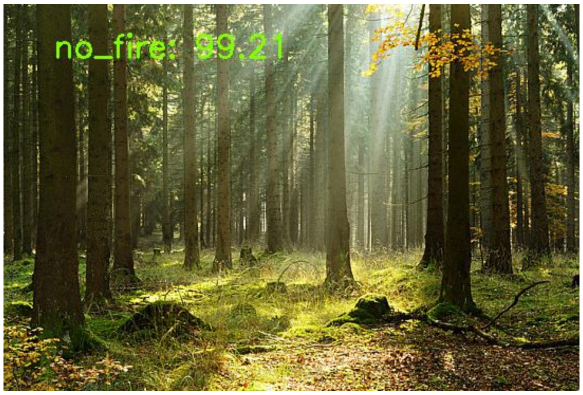
a) The logistic regression model can be generalized to support multiple classes directly without having to train and combine multiple binary classifiers. How?
b) One advantage of logistic regression is that it produces a model that can be scored to new data rapidly, without recomputation. Another is the relative ease of interpretation of the model, as compared with other classification methods. The key conceptual idea is understanding an odds ratio. what is an odds ratio?
c) The Bayesian classifier only works with categorical predictors. What should be done to apply naive Bayes to numerical predictors?
d) if the training data is small or if the dataset has a fewer number of observations and a higher number of features, can we use naive bayes? why?
e) Can logistic regression obtain a non-linear decision boundary? Explain.

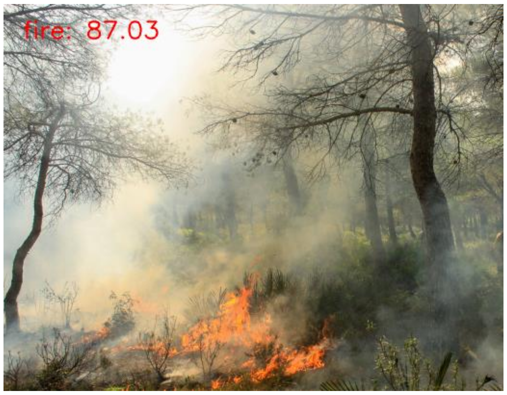
Problem 2: Forest Fire Classification (30 pts)

# Implementation

The dataset is designed for binary classification of Fire and No-Fire detection in the forests landscape. You should do this classification using logistic regression. Do the following steps:

a) Pre-processing: The size of the images may be different. Resize them. Then normalize the images. Determine the target of each image. (We suggest using glob and OpenCV libraries).
b) Split the data into train and test.
c) Run logistic regression on train data.
d) Report accuracy and confusion matrix for test data.
e) Find the best probability threshold for the training data and report the accuracy and confusion matrix again. Did the results improve?
f) Save the best model. import this model to another file (e.g., classifier.py). Download some forest images (fire and non-fire). feed this image to your classifier and make a prediction. Finally, show the label of each image with its probability in the photo. If the photo was fire, with red color, otherwise green color (similar to below).





Problem 3: Sentiment Analysis (30 + 10 pts)

# Implementation

Snappfood (an online food delivery company) user comments containing 70,000 comments with two labels: 1- Happy (Positive) 2- Sad (Negative). We want you to build naive bayes Classifier from scratch to perform sentiment analysis. (This link[1] can be very useful). Follow the steps below:

a) Before building the model, you must do the required text preprocessing. Explain all pre-processing steps. (You can use the hazm[2] library, which is used to process the Persian language. (Note that this section has the highest score).

b) Building the naive bayes classifier. Explain how naive bayes is used for this problem.

c) Fitting the model on training set and evaluating accuracies on the test set.

Problem 4: Remove noise from images (22.5 pts)

# Implementation

We want build a system that removes noise from digit images. It will take as input a noisy digit image, and it will output a clean digit image. Follow the steps below:

a) Load MNIST dataset from sklearn.

b) Build your training dataset by adding random noise to it.

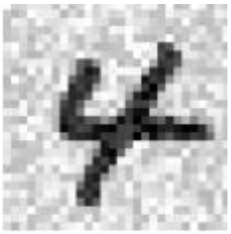
c) What is your target here?

d) Using whatever classification method, you want. (From those you have read so far), train a model to solve this problem. Explain your problem-solving method.

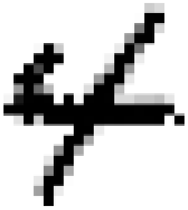
e) Plot like below for arbitrary samples:

f) Use grid search to find the best hyperparameters for the decision tree.

g) Evaluate the performance of the best decision tree on the test data. Can the decision tree determine the pass or fail status?



Before



After

[1] https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/

[2] https://www.roshan-ai.ir/hazm/