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)

(Project)

Outlines. Outlines. In this project, as your last task in this course, You will implement three problems related to course topics. You will have a presentation meeting and Q&A about your implementation and strategies. In that meeting, assignments will be noticed as well.

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

Project Manual

Delay policy. You have NO extra time to delay submitting your answers in this project. You can not use your remained extra delay time from assignments.

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.

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_Project_[std-number].zip
- Report
  - ML_ Project _[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: Regression in industry (30 pts)

Implementation

One of the critical tasks in the industry is predicting environmental measures such as pressure or temperature. You are given a dataset for bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea. This data consists of summer data from 2013 to 2017. The input data comprises the LDAPS model's next-day forecast data, in-situ maximum and minimum present-day temperatures, and auxiliary geographic variables. This data has two outputs (i.e., next-day maximum and minimum air temperatures). You will use 2013-2015 data for training and 2016-2017 for tests.

a) Load the dataset, apply the needed preprocessing steps(missing values, normalization, outlier detection using K-Means, removing unneeded columns, etc.), and split the dataset to train and test sets.

b) With reasonable logic, choose and train a regression model(gradient-based) to estimate the "Next_Tmax" column. Report the SSE metric, and plot the regression and truth lines in one graph. (the x-axis will be the "date" with station ID, and the y-axis will be Next_Tmax values[estimated and truth]). Explain the parameters and model complexity you used; How did you reach it?

c) Now, we want to estimate "Next_Tmax" and "Next_Tmin" at the same time. How can you address this problem? Describe your answer and strategy.

d) Based on your strategy in the previous part, implement a gradient-based regression function that takes feature vectors and training parameters(learning rate, etc.), and can estimate the two above columns. Evaluate your model, report the SSE metric, and plot the regression and truth lines in one graph(one graph for each column).

e) Compare results in parts (b) and (d). which one is better? is it suitable to train one model to estimate one column or train one model to multiple columns?

f) Try to implement, and report results if we use normal equation to estimate two columns as the same time.

g) How KNeighborsRegressor works? Implement and evaluate it. Compare the results.

Problem 2: Skin Detection(40 pts)

Implementation

In this problem, we want to segment the skin areas in images. You are given thress datasets; the first(D1) is a dataset with three features (RGB values) and a target(is it skin or not); You use this dataset for the train. The second one(D2) consists of multiple images that you should segment the skin areas and generate skin masks (You use this dataset for the test). The third one(pratheepan) is for training in e-g.

a) With sufficient tries, train a Gaussian mixture model for binary classification with the first dataset(D1). Discuss selected values for function parameters. (You are allowed to use libraries)

b) Load the second dataset(D2) and using obtained GMM in the part a, classify the images` pixels into skin/not skin. Plot your result as the same dimension images. You see an example of desired output in figure below(your result may have some noise). Based on given grand truth, calculate confusion matrix.


c) Now, using the Bayes classifiers, repeat the last part. Compare the results. Which one has better performance? Explain your reasons.

d) In probability-based classifiers, playing with probability instead of the predicted labels can outperform. Using the trained Bayes classifier, generate a skin probability map for each test image(D2). Skin probability map composed of the probability of each pixel that can be skin.


e) We want to apply to be locality to skin detection task. For this purpose, you should generate a new dataset using "pratheepan" images. First, load images and generate skin probability maps(with above Bayes classifier that you tranied), then create a 12-D feature vector with the original label for each pixel following as 3 RGB values + value of probability for that pixel + 8 probability values for neighbors(assume zero for border pixels that have fewer neighbor)

   In the end, you have a dataset with 12 columns for features, one for the target(0/1), and the count of all pixels in all training images as rows.

f) Calculate the correlation coefficient for the above dataset. Can you indicate the locality attribute in this result? Describe the correlation coefficient and its usage.

g) Train a logistic regression model for obtained dataset and evaluate it with D2. Plot your result as the same dimension images suach as part b. Based on given grad truth, calculate confusion matrix. Compare result with part (b).

Problem 3: Credit Approval (30 pts)

Implementation

In this problem, we were hoping you could study the given dataset related to credit card approval.

a) Load the dataset, apply the needed preprocessing steps(missing values, normalization, outlier detection using K-Means, etc.), and split the dataset to train and test sets(15:85).

b) Try to train a K-NN as the classifier. Evaluate your result using the confusion matrix and recall, precision, and accuracy. Find the best K hyperparameter with the Elbow technique

c) Try to train a Decision Tree as the classifier. Evaluate your result using the confusion matrix and recall, precision, and accuracy. Use cross-validation.

d) Try to train a Random Forest as the classifier. Evaluate your result using the confusion matrix and recall, precision, and accuracy. Use grid search. How much does it take to find and train a model? is it worth it?

e) Try to train a Gaussian Naive Bayes(GBN) as the classifier. Evaluate your result using the confusion matrix and recall, precision, and accuracy.

f) Try to train a Logistic Regression as the classifier. Evaluate your result using the confusion matrix and recall, precision, and accuracy.

g) Try to train an SVM as the classifier. Evaluate your result using the confusion matrix and recall, precision, and accuracy. Try to find the optimal kernels and parameters.

h) Try to train an AdaBoost as the classifier. Evaluate your result using the confusion matrix and recall, precision, and accuracy.

i) Try to train K-Means as the clustering algorithm, cluster the dataset into two part, and use it for classification. Evaluate your result using the confusion matrix and recall, precision, and accuracy.

j) Create a table for all results you obtained. Your columns will be: FN, FP, accuracy, precision, F1 score, recall, training time(ms), predicting time(ms-how much it takes that model to predict a single sample)

   Compare the results and discuss. Which one is the better choice for the given task and based on which metric? Why? Justify your answer.

k) Modify the best trained model in previous part to enhance your mentioned metric.