PRACTICAL ASSIGNMENT
of
Artificial Intelligence

INSTITUTO POLITÉCNICO DE BRAGANÇA
Escola Superior de Tecnologia e Gestão

Computer Engineering
December 16, 2024

Delivery deadline: 23/12/2024

HEALTH INSURANCE

Develop a Machine Learning model based on the Support Vector Machines (SVM) algorithm to predict future costs associated with new customers of a health insurance company. The goal is to create a tool that helps the company set insurance premiums more accurately and fairly. The dataset provided to induce the model contains the insurer's historical costs for each customer, along with characteristics defining the customer's profile, such as gender, marital status, area of residence, body mass index, whether they smoke and age class.

Additional task to increase your mark (2 points): Identify which features most influence a customer's healthcare costs. Hint: Although the SVM algorithm does not directly provide feature importance, specific techniques allow this estimation. One of these techniques consists of evaluating the influence of each feature by removing it from the dataset and measuring the impact on the model's performance. Scikit-learn provides a specific function for applying this approach.

Tasks to complete

- Implement in a Jupyter Notebook document the solution to the problem enunciated, developing, with the data made available in the `dataset.csv` file, a high-performance ML model which fully meets the described specifications. [generate grupo#_solucao.ipynb, replacing the # character with the working group number]
- With the developed model, estimate customer costs with the characteristics indicated in the file just_features.csv, for which the real cost is unknown. [generate grupo#_custos_estimados.csv]
- Prepare a brief presentation of the work developed. [`generate` grupo#_apresentacao.pdf]

Implementation considerations

- To develop the ML model requested, Scikit-Learn and other Python support packages introduced during the course should be used.
- The dataset.csv and just_features.csv files, containing, respectively, the dataset for inducing the model (with 2,215 instances - each instance contains the characteristics of an insured customer and the cost that this customer represented for the company) and the dataset that will put it to the test (with 550 instances without cost indication), can be downloaded from the course unit area of the ipb.virtual platform, at Resources/avaliacao/trabPratico.
- The presentation, after being elaborated in MS PowerPoint or another suitable tool of your choice, must be converted to pdf, contain between 3 and 7 slides (not counting the 1st), and the font size of the body text should be between 16 and 20 pt. Don't forget to include the number of the working group and the name and number of each group member on the first slide, and avoid including the implementation code in the presentation.
- The file to be submitted group#estimated_costs.csv should only contain the cost column and should, therefore, look like this:

```
custo
1111.1
999.9
1234.5
...
```

- The evaluation of the practical assignment will consider, in particular, the correctness, according to the R^2 metric, of the estimates contained in the file group#estimated_costs.csv, generated by the model for clients with undisclosed costs.

General considerations

- Groups of 3 students should carry out this practical work, and it is compulsory for passing the course (work by two students will incur a penalty of 0.5 marks, and individual work will incur a penalty of 1 mark).
  (Suggestion for cooperation in the tasks to be carried out: each group member can start by developing their regression model and then jointly prepare a single model for submission, which takes advantage of the best ideas and options considered in the three proposals.)
- It is expressly forbidden to copy all or part of code from sources other than the documentation provided by the teachers of the curricular unit.
- The work must only be submitted by one of the group members within the established deadline, obligatorily on the e-learning portal (at http://virtual.ipb.pt/, choose <Trabalho Pratico> on the <Assignments> tab, within the IA area), and under no circumstances may it be sent by e-mail.
- The 3 requested files (grupo#_solucao.ipynb, grupo#_custos_estimados.csv and grupo#_apresentacao.pdf) must be submitted in separate attachments (3 attachments) and not compressed. Failure to comply with this rule will result in a penalty of 1 point.
- Work can be submitted up to 5 days late, subject to a daily penalty on the final mark. The 24th and 25th are not counted for penalty purposes: those who submit their work on the 24th, 25th or 26th will only lose 1 mark; those who submit on the 27th will lose 2 marks; on the 28th, they will lose 3 marks; and so on.
- Resubmissions will not be allowed (when submitting, make sure it is the final version).
- Students may be called upon to defend their work if deemed necessary.