INF-0615 – Aprendizado de Máquina Supervisionado I

Exercício 10
Análise de Conjunto Aberto (*Open-Set*)
Economic Region Prediction

1 Descrição do Problema

O enunciado e as bases de dados são similares àquelas utilizadas no exercício 06. No entanto, agora iremos considerar que é possível que o exemplo de entrada seja um país que não pertence a nenhuma das Regiões Econômicas estudadas. Assim, como no exercício 06, temos um total de 34 atributos que serão utilizados para a predição da região, a qual está presente na coluna *continent*. Além disso, há 5 regiões (classes) de interesse: Central America, EasternAfrica, South-Eastern Asia, South America, Westen Europe, ou seja, é um problema multi-classe. Além disso, há países de outras regiões do mundo que também podem aparecer durante a operação do modelo e, para estes casos, precisamos ser capazes de dizer se eles pertencem ou não há uma das cinco classes de interesse. Se nosso modelo predizer que o país fornecido apresenta como *output* uma das cinco regiões estudadas, devemos ser capazes de dizer a qual destas cinco ele pertence. Senão, ele deve ser detectado como um país que não está em nenhuma das regiões de interesse, e assim deve ser classificado como um caso do conjunto aberto (*Open-Set Case*).

2 Tarefas

Neste exercício, pedimos que você:

1. Inspecione os dados, verifique a distribuição das classes e os tipos dos atributos.
2. Normalize os dados para que fiquem mais bem preparados para o treinamento.
3. Treine 5 regressões logísticas, uma para cada região, seguindo o protocolo One-vs-All.
4. Execute o modelo treinado sobre a base de validação com classes conhecidas (*Known_validation_set.csv*).
5. Considerando as bases de validação das classes conhecidas e desconhecidas (*Known_validation_set.csv* e *Unknown_validation_set.csv*), busque pelo melhor *threshold* para definir se o exemplo pertence ou não a uma das classes conhecidas.
6. Carregue o conjunto de teste final (*OpenSet_test_set.csv*)e para cada exemplo, classifique se ele pertence ou não ao conjunto das classes conhecidas. Para aqueles preditos como pertencentes, classifique ainda em qual das cinco regiões ele está presente.
7. Calcule a matriz de confusão relativa, a acurácia por classe e acurácia balanceada do modelo. O modelo conseguiu ter uma boa performance para predizer os casos conhecidos e desconhecidos?

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *Known_training_set.csv*: dados de treinamento com as classes conhecidas;
- *Known_validation_set.csv*: dados de validação com as classes conhecidas;
- *Unknown_validation_set.csv*: dados de validação com as classes desconhecidas;
- *OpenSet_test_set.csv*: dados de teste final com classes conhecidas e desconhecidas;
- *Ex10.R*: Código que implementa as soluções do exercício;

4 Referências

1. *Economic Freedom of the World*. Kaggle. https://www.kaggle.com/gsutters/economic-freedom.