Instituto Nacional de Telecomunicações - Inatel

AG2 – Engenharias de Computação e Software

Prof. Me. Marcelo Vinícius Cysneiros Aragão
Prof. Me. Renzo Mesquita Paranaíba

1 Introdução

Neste semestre a AG2 acontecerá na forma de um trabalho prático. Você deverá utilizar seus conhecimentos para, a partir do conjunto de dados proposto, treinar, avaliar e disponibilizar um modelo de aprendizado de máquina para classificar diferentes canais de vendas de produtos.

2 Conjunto de Dados

O conjunto de dados "Wholesale customers" [1], disponibilizado em 2014, se refere a clientes de um distribuidor atacadista, e inclui o gasto anual em unidades monetárias (u.m.) em diversas categorias de produtos. É composto por 440 amostras, que representam observações sobre diferentes vendas realizadas pelo distribuidor. Cada amostra do conjunto é descrita por:

- Sete atributos: *Region* (região de Portugal na qual a venda foi registrada), *Fresh* (gasto anual com produtos frescos em u.m.), *Milk* (gasto anual com laticínios em u.m.), *Grocery* (gasto anual com produtos de mercearia em u.m.), *Frozen* (gasto anual com produtos congelados em u.m.), *Detergents_Paper* (gasto anual com detergentes e produtos de papel em u.m.) e *Delicatessen* (gasto anual com guloseimas em u.m.);
- Um rótulo de classe (*Channel*), que indica o canal de vendas pelo qual a transação foi realizada, podendo ser "HoReCa" (hotel, restaurante ou café) ou "Retail" (varejo).

Neste trabalho será utilizada uma versão pré-processada do conjunto originalmente apresentado por Cardoso [1] em 2014. Os dados originais foram obtidos do UCI Machine Learning Repository.

3 Etapas para Realização

1. Baixar o conjunto de dados em formato CSV (*comma-separated-values*).
2. Fazer a leitura dos dados utilizando a biblioteca Pandas.
3. Converter os valores presentes no conjunto de dados para números inteiros, de acordo com este mapeamento:

Coluna	Tipo original	Valor original	Tipo após a substituição	Valor após a substituição
"Channel"	String (object)	"HoReCa"	Integer (int64)	0
		"Retail"		1
"Region"	String (object)	"Lisbon"	Integer (int64)	0
		"Oporto"		1
		"Other"		2

   Dica: função `replace`, presente na classe Series do Pandas.

4. Reordenar as colunas do conjunto de dados da seguinte forma:

Antes da ordenação	['Channel', 'Region', 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicatessen']
Depois da ordenação	['Region', 'Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents_Paper', 'Delicatessen', 'Channel']

   Dica: função `reindex` e atributo `columns`, presentes na classe DataFrame do Pandas.

5. Separar o conjunto de dados em duas partes: 80% para treinamento e 20% para testes. Dica: função `train_test_split`, presente no módulo Model Selection do scikit-learn.
6. Escolher um dos modelos de classificação a seguir:
   - Decision Tree: Wikipedia, KDnuggets e scikit-learn.
   - k-Nearest Neighbors: Wikipedia, Towards Data Science e scikit-learn.
   - Multilayer Perceptron: Wikipedia, KDnuggets e scikit-learn.
   - Naïve Bayes: Wikipedia, Towards Data Science e scikit-learn.
7. Treinar o modelo com o conjunto de treinamento e classificar as amostras do conjunto de teste. Dica: funções `fit` e `predict`, presentes nos classificadores.
8. Exibir métricas de avaliação, para que possa ser verificada a acurácia do modelo. Dica: função `classification_report`, que já inclui diversas métricas.
9. Criar uma opção que permita ao usuário inserir dados arbitrários que devem ser classificados pelo modelo. O modelo deverá imprimir, com base no conhecimento adquirido durante o treinamento, a qual canal de vendas os dados inseridos se referem ("HoReCa" ou "Retail"). Dica: funções `input` (para leitura dos dados) e `predict` (presente nos classificadores).

4 Orientações Adicionais

- O trabalho deverá ser feito em dupla;
- Qualquer linguagem de programação pode ser utilizada;
- A entrega deverá ser feita por meio de um arquivo zip com todo o conteúdo do projeto, ou o link de um repositório privado do GitHub;
- Para apresentação, o aluno deverá gravar um vídeo de no máximo 7min de duração, explicando em detalhes as etapas do projeto desenvolvido;
- O vídeo poderá ser feito gravando a própria tela do computador enquanto o aluno explica ou até mesmo ser usado o *smartphone*, desde que as explicações das etapas estejam nítidas;
- A entrega deve ser feita até o dia 27/11/2024. Disponibilize vídeo e arquivo zip (se for usar) no OneDrive ou GoogleDrive, com permissão de acesso para renzo@inatel.br. Se usar GitHub (em vez de arquivo zip), disponibilize o link também com acesso público.

Bom trabalhos a todos!

Referências

[1] Margarida Cardoso. *Wholesale customers*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5030X. 2014.