Instituto Nacional de Telecomunicações - Inatel

AG002 – 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 espécies de pinguins.

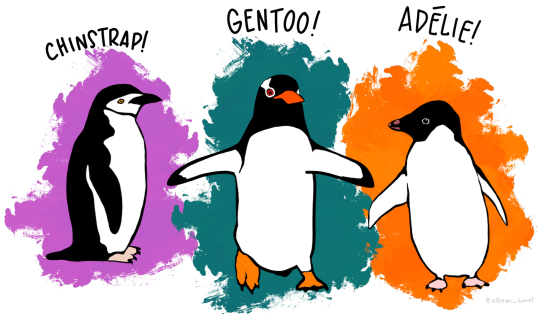

Figura 1: Os pinguins do Arquipélago Palmer. Arte por Allison Horst.

2 Conjunto de Dados

O conjunto de dados "palmerpenguins" [2] contêm medidas de tamanho para três espécies de pinguins observadas em três ilhas do Arquipélago Palmer, na Antártica. Esses dados foram coletados de 2007 a 2009 pela Dra. Kristen Gorman com o Programa de Pesquisa Ecológica de Longo Prazo da Estação Palmer, parte da Rede de Pesquisa Ecológica de Longo Prazo dos EUA.

Mais especificamente, o conjunto apresenta 333 amostras, que representam observações sobre diferentes pinguins analisados durante a pesquisa. Cada amostra do conjunto é descrita por:

- Seis atributos: *culmen_length_mm* (comprimento do cúlmen[1] em milímetros), *culmen_depth_mm* (profundidade do cúlmen em milímetros), *flipper_length_mm* (comprimento da nadadeira em milímetros), *body_mass_g* (massa corporal em gramas), *island* (nome da ilha no Arquipélago Palmer na qual foi feita a observação – Biscoe, Dream ou Torgersen) e *sex* (sexo do pinguim – Female ou Male);
- Um rótulo de classe (*species*), que representa a espécie do pinguim em questão (Adélie, Chinstrap ou Gentoo).

Neste trabalho será utilizada uma versão pré-processada do conjunto originalmente apresentado por Gorman, Williams e Fraser [1] em 2014. Os dados originais foram obtidos do Kaggle.

[1] Crista superior do bico de um pássaro.

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
island	String (object)	"Biscoe"	Integer (int64)	0
		"Dream"		1
		"Torgersen"		2
sex	String (object)	"FEMALE"	Integer (int64)	0
		"MALE"		1
species	String (object)	"Adeline"	Integer (int64)	0
		"Chinstrap"		1
		"Gentoo"		2

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

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

Antes da ordenação	['species', 'island', 'culmen_length_mm', 'culmen_depth_mm', 'flipper_length_mm', 'body_mass_g', 'sex']
Depois da ordenação	['island', 'sex', 'culmen_length_mm', 'culmen_depth_mm', 'flipper_length_mm', 'body_mass_g', 'species']

   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 espécie de pinguim os dados inseridos se referem ("Adeline", "Chinstrap" ou "Gentoo"). 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 11/06/2024. 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 permissão de acesso.

Bom trabalhos a todos!

Referências

[1] Kristen B. Gorman, Tony D. Williams e William R. Fraser. "Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (Genus Pygoscelis)". Em: *PLoS ONE* 9.3 (2014), e90081. ISSN: 19326203. DOI: 10.1371/journal.pone.0090081.

[2] Allison Marie Horst, Alison Presmanes Hill e Kristen B Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. 2020. DOI: 10.5281/zenodo.3960218. URL: https://allisonhorst.github.io/palmerpenguins/.