Identificando a quantidade de espécies de pinguim baseado em medidas corporais

Enzo Putton Tortelli

2024-04-13

Atenção: esse é um trabalho de ficção. Consultar as referência para mais detalhes.

1. Introdução e objetivo

O continente Antartida se destaca por ser o principal habitat natural dos pinguins. A região _____ é de especial interesse, pois é habitada por um grande conglomerado de milhares desses animais[1]. Existe uma suspeita na comunidade científica local de que esse conglomerado é formado por mais de uma espécie da ave. Dessa forma, esse estudo tem como objetivo identificar a quantidade de espécies presentes nesse grupo aplicando uma técnica de clusterização nas medidas corporais de uma amostra dessa população.

2. Materiais e métodos

Para atingir o objetivo proposto, foram amostrados 167 penguins macho[ref] desse conglomerado e, juntamente com o seu peso, foram medidas 3 partes de seus corpos: o comprimento da nadadeira, e o comprimento e profundidade do cúlmen (medidas relacionada ao bico).

Depois, removemos, dos dados, valores absurdos e observações incompletas. Com intuito de dar peso iguais as quatro variáveis, realizamos um escalonamento (por valores escolhidos à mão) do peso e do comprimento da nadadeira, e uma centralização de todas elas, para, então, fazer uma análise de componentes principais (PCA).

A PCA é uma técnica para redução da dimensionalidade dos dados. Em poucas palavras, essa técnica faz um mapeamento linear dos dados para um espaço de menor dimensão de maneira que a variância dos dados representada nessa dimensão seja maximizada. [ref] https://en.wikipedia.org/wiki/Dimensionality_reduction

Ainda, estudando o resultado da PCA, criamos nosso modelo de cluster empregando o método *k-medoids* com apenas as componentes principais mais importantes, utilizando o método de *Elbow* para decidir o número de clusters apropriado.

Resumidamente, o método *k-medoids* divide os dados em grupos e tenta minimizar, dentro de cada grupo, a distância entre os pontos e o medoid. O medoid de cada cluster é definido, justamente, como a observação que minimiza a soma dessas distâncias. Além disso, esse método permite que sejam usadas diversas definições de distância; nesse trabalho, usei a distância euclidiana. [ref] https://en.wikipedia.org/wiki/K-medoids

O método de *Elbow*

- elbow plot
- k-medeoids
- silhueta

3. Resultados e conclusão

Uma primeira análise descritiva dos dados é dada pela **Figura 1**. Nela, é possível observar que as medidas de massa corporal e comprimento da nadadeira estão numa escala maior que as demais variáveis, e, como dito anteriormente, queremos dar peso iguais a elas. Assim, dividimos a massa corporal por 100 e a medida da nadadeira por 10. Depois disso, foi feita uma centralização de todas as variáveis. O resultado pode ser visto na **Figura 2**.

	Compri. do culmen (mm)	Prof. do culmen (mm)	Compri. da nadadeira	Massa corporal (g)
	Min. :35.10	Min. :14.10	Min. :178.0	Min. :3250
	1st Qu.:41.05	1st Qu.:16.05	1st Qu.:193.0	1st Qu.:3900
	Median :46.80	Median :18.40	Median :201.0	Median :4300
	Mean :45.92	Mean :17.87	Mean :204.5	Mean :4547
	3rd Qu.:50.35	3rd Qu.:19.20	3rd Qu.:219.0	3rd Qu.:5325
	Max. :59.60	Max. :21.50	Max. :231.0	Max. :6300

```
## \begin{table}[!htbp] \centering \renewcommand*{\arraystretch}{1.1}\caption{Summary Statistics}\resize
## \begin{tabular}{lrrr}
## \hline
## \hline
## Variable & Min & Mean & Max \\
## \hline
## Compri. do culmen (mm) & 35 & 46 & 60 \\
## Prof. do culmen (mm) & 14 & 18 & 22 \\
## Compri. da nadadeira & 178 & 205 & 231 \\
## Massa corporal (g) & 3250 & 4547 & 6300\\
## \hline
## \hline
## \end{tabular}
## }
## \end{table}
```

Feito isso, criamos nosso modelo de PCA. Podemos ver, pela **Figura 3**, que a componente principal 1 (PC1) é responsável por 78,4% da variância, e a componente principal 2 (PC2), por 19,7%, sendo ambas responsáveis por 98,1% de toda a variância contida nos dados. A **Figura 4** ilustra esse grande acúmulo de percentual da variância nas componentes principais 1 e 2. Portanto, elas foram escolhidas para compor nosso modelo de clusterização. A **Figura 5** nos mostra que altos valores da componente principal 1 estão associados, principalmente, com maior massa corporal, enquanto que maiores valores da componente principal 2 estão assoaciados com maior comprimento do cúlmen. As variáveis profundidade do cúlmen e tamanho da nadadeira possuem pouca relevância no cálculo de PC1 e PC2 quando comparada às outras duas.

Com as componentes principais em mãos, para determinar o número ideal de clusters, foi feito uma análise utilizando o Método Elbow, que pode ser visto na **Figura 6**:

Utilizando esse gráfico, chegamos a conclusão de que três clusters oferecem o melhor "custo-benefício" entre a diminuição da soma de quadrado interna total e o número de clusters.

Para finalizar, seguem os gráficos que ilustram o resultado da clusterização por *k-medoids*:

Na **Figura 7**, temos nossas observações inseridas no plano cartesiano. Observa-se que os clusters estão bem definidos.

O gráfico de silhueta, **Figura 8**, confirma o que foi dito: com uma silhueta média de 0.61, ALGUMA COISA

Conclusão

Apresentada todas as evidências, concluí-se que a suspeita dos ciêntistas sobre a existência de mais de uma espécie de pinguim no grande conglomerado da região de _____ se confirma.

4. Referências