Guía 8: Clasificación

Laboratorio de datos 2023 (comisión: G. Solovey)

Para resolver estos ejercicios, cargar la librería `palmerpenguins`, usar el dataset `penguins`, borrando las observaciones que tengan algún `NA`. En cierto punto van a necesitar estos paquetes de R:

```
require(palmerpenguins)
require(patchwork)
require(tidyverse)
require(rpart)
require(rpart.plot)
require(parttree)
require(class)
```

Parte 1

El objetivo de los ejercicios es trabajar con árboles de decisión para clasificar a pinguinos de la especie Gentoo en macho y hembra utilizando como variables a el largo del pico y el largo de la aleta.

Ejercicio 0

Realizar visualizaciones adecuada para explorar si es esperable que las variables largo del pico y largo de la aleta serán útiles para clasificar a los pinguinos en machos y hembras. Evaluar cada variable por separado y las dos juntas.

Ejercicio 1

Dividir el dataset en un conjunto de entrenamiento y uno de test (80%-20%) en forma aleatoria utilizando una semilla fija. Utilizar, para este ejercicio, la variable largo del pico del pinguino para predecir el sexo de los pinguinos.

a. Escribir el modelo de clasificación. Es decir, cómo se decide si un nuevo pinguino, del que se conoce su largo del pico, es macho o hembra. ¿Cuál es el “accuracy” mínimo esperable para un modelo de clasificación con estos datos?
b. Usando el conjunto de entrenamiento, encontrar el clasificador que minimiza el error. Para hacerlo, tomar una grilla de valores posibles de la “masa crítica” y calcular el error de clasificación para cada uno de estos valores. Hacer un gráfico del error de clasificación en función de la “masa crítica”.
c. Probar la capacidad predictiva de este clasificador en el conjunto de test. ¿Cuál es la matriz de confusión y la “accuracy” del modelo?

Ejercicio 2

Repetir los pasos del ejercicio 1 utilizando esta vez sólo la variable largo de la aleta. Comparar la capacidad predictiva de este modelo con el encontrado en el ejercicio 1. ¿Era previsible el resultado en base a lo visto exploratoriamente en el ejercicio 0?

Ejercicio 3

Crear un árbol de decisión para predecir el sexo de los pinguinos en función del peso y el largo de la aleta. Evaluar la perfomance del modelo usando la matriz de confusión y “accuracy”. Realizar visualizaciones pertinentes.

Ejercicio 4

Repetir el ejercicio 3 para otras 1000 particiones train-test. Para cada una guardar el “accuracy” en el grupo de test. Con esos datos hacer una visualización del “accuracy”. ¿Qué se puede decir del accuracy del modelo (por ejemplo, en qué rango se encuentra)?

Ejercicio 5

A partir de lo hecho en el ejercicio 4, explorar diferentes parámetros de los árboles de decisión para encontrar la mejor configuración en términos de mejorar la predicción en el grupo de test. Si se aumenta el número de particiones del árbol, ¿qué es esperable que ocurra con el error de predicción dentro del dataset de entrenamiento y en el de test?

Ejercicio 6

Implementar un clasificador de k-NN que prediga el sexo de los pinguinos utilizando como variables a el largo del pico y el largo de la aleta. Hacerlo para diferentes valores de k (impares) y evaluar el error de predicción en cada caso. ¿Cómo elegiría el valor de k óptimo? Comparar con el modelo de árboles de decisión de la parte 1.

Parte 2

Este ejercicio es más libre, tienen que montar un modelo de clasificiacón para la especie de los pinguinos. Pueden elegir las variables y el método (k-NN o árboles de decisión). Reporten los resultados con visualizaciones adecuadas.