ENFIN761

Business Intelligence para las Finanzas

Ayudantía 8

Profesor: *David Díaz S.*
Ayudantes: *Gabriel Cabrera G.*[1]
03 noviembre 2019

Information Entropy

1. Cargue a su espacio de trabajo (*workspace*) la base de datos `iris` desde la librería `sklearn`. Separe la variables *target* y *features*.

2. Construya los siguientes gráficos:

   a. Un gráfico 2d del tipo *scatter* entre los *features*, *sepal width (cm)* y *sepal length (cm)*.

   b. Un gráfico 3d del tipo *scatter* entre los *features*, *sepal width (cm)*, *sepal length (cm)* y *petal length (cm)*.

3. Genere una función que permita calcular la entropía (*entropy*) existente en un conjunto de datos. Recuerde que la formula de la entropía es la siguiente:

$$H(p_1, ..., p_n) = \sum_{i=1}^{n} p_i \cdot log_2(p_i)$$

   Donde p_i es la probabilidad del valor i y n el número de valores posibles.

4. Utilizando la función creada en (3), calcule la entropía cuando:

   a. La variable *target* es igual a la especie *setosa*.

   b. La variable *target* es igual a las tres especies (*setosa*, *versicolor*, *virginica*).

Information Gain

1. Genere una función que permita calcular la *Information Gain* existente en un conjunto de datos. Recuerde que la formula de la *Information Gain* es la siguiente:

$$IG = H_p - \sum_{i=1}^{n} p_{ci} \cdot H_{ci}$$

[1] :gcabrerag@fen.uchile.cl

Donde H_p es la entropía de los *parent* (muestra completa, sin realizar ninguna segmentación), n es el número de valores de la variables *target* (*childs*), p_{ci} es la probabilidad de que una observación se encuentre en el *child* i y H_{ci} es la entropía del *child* (segmento) i. La función debe tener los siguientes parámetros:

a. `data`, contiene tanto la variable *target* como *features*.

b. `feature`, string que contiene el nombre de la variable *feature*.

c. `target`, string que contiene el nombre de la variable *target*.

d. `bins`, número de segmentación, por default debe ser 4.

2. Utilizando la función creada en (1), calcule la *Information Gain* de cada *feature* (*sepal length*, *sepal width*, *petal length* y *petal width*).

Propuesto

1. Simule la probabilidad de obtener cara o sello en una moneda no cargada, luego calcule su entropía. Utilice la librería `random` y repita el ejercicio 1.000.000 veces, debe guardar cada resultado en una lista o `DataFrame`.