Data Mining Lab Assignment-5

CSE 6th semester

(Language/Platform: Python)

Note: Feature Selection- Select k useful features out of n futures in a dataset, where k<n.

Objective 1:

Load page_block.csv numeric (Beginner's Guide Page-block Classification (kaggle.com)) dataset (make appropriate preprocessing if required). Use **Fisher's Score** (f) ranking method for assigning an appropriate rank to each feature and select the top k ranked features.

$$f = \frac{\mu_1^2 - \mu_2^2}{\sigma_1^2 + \sigma_2^2}$$

Where $\mu_1, \mu_2, \sigma_1, \sigma_2$ are means and standard deviations for negative and positive class respectively.

Write the equivalent function in python for the following:

1. Compute the mean of attribute values against both class labels.
2. Compute the standard deviation of attribute values against both class labels.
3. ComputeFisher's score (f) the for each attribute.
4. Assigned rank for each attribute (High f value has a high ranking).

Objective 2:

Load buys_computer.csv nominal (Buy Computer (kaggle.com)) dataset (make appropriate preprocessing if required). Use the information gain formula.

Let dataset D with two class labels c_1 and c_2 then expected information (entropy) info (D) can be computed by the formula as given below:

$$Info(D) = -\sum_{i=1}^{label} p_i log_2 p_i$$

Where p_i is the probability of class c_i and label is the number of classes(2 in the case of the binary class dataset).

The entropy of an attribute A with v nominal labels can be computed by the formula as given below:

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

Where $|D_j|$ is count(v_j) and $|D|$ is count (D).

The information gain of attribute A is given by the formula as follows:

$$Gain\ (A) = Info(D) - Info_A\ (D)$$

Write the equivalent function in python for the following:

1. Compute the entropy (info (dataset)) of the entire dataset.
2. Compute $Info_A$(Dataset) for each attribute A in the dataset.

3. Compute Information Gain (A) for each attribute A in the dataset.
4. Select k attributes with the highest information gain.