Naive Bayes

Program 2

Dataset

You can download the dataset and notes describing it from the link provided in the assignment “info” section in CSCADE.

These data were obtained from the UCI Machine Learning repository, the “Lymphography Data Set” was originally obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

Raw dataset available at https://archive.ics.uci.edu/ml/datasets/Lymphography

This version has the data files renamed (with “.csv” and “.txt” extensions) and the main data file has column names added to be more user-friendly.

- `lymphography.csv`: This is the data matrix. Rows are samples, columns are features. The first column is the class. The first row contains column names.
- `lymphography_info.txt`: Information about the data, including some descriptions of the variables and related references.

Original Owners:

University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia

Assignment

Using the programming language of your choice (Python 3 with Pandas and Numpy is highly recommended), write a program that will *learn* a Naive Bayes classifier to classify the lymphography dataset. there are 3 classes (NOTE: The class numbering starts at 2 due to a removed class with low support.)

2. metastases: 81 3. malign lymph: 61 4. fibrosis: 4

Details

- You may use a library for reading the CSV format data. Python can read CSV with the built-in `csv` library, or from `numpy`, or `pandas`.
- Split the data into a training and validation set before proceeding. The first 30 rows in the CSV file should be used as the *validation* set.
- All of the data is categorical - treat the integers as individual levels of categorical variables (see the "lymphography_info.txt" file).
- If you use any random number generation, be sure to lock down the PRNG seed so that it is reproducible (i.e. use a fixed seed so that the program produces the same results each time).

- Be sure your program reports its performance on the validation set. Report the number of correct classifications, number of incorrect classifications and accuracy ($acc = \frac{correct}{correct+incorrect} \times 100\%$).
- For each of the three classes, report the accuracy for that class individually (correct predictions for the class divided by the total number of samples of that class in the validation set). Note: For the *fibrosis* class there are only 2 samples in the validation set, so your accuracy on that class will be either 0%, 50%, or 100%.

Deliverables

Zip up all source code required for your project along with a document (plain text, Word, or PDF) detailing your choice of language and architectural design decisions, as well as a report of the performance of your decision tree on the validation set.

Submit your ZIP archive through CSCADE.