Supervised Learning

Datasets

- https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer/
- http://archive.ics.uci.edu/ml/machine-learning-databases/chess/king-rook-vs-king-pawn/

Algorithms

- Random Forests
- Decision Tree

Description of Dataset

Attributes

- Features
  - Age
    - Data: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
  - Menopuase
    - Data:lt40, ge40, premeno.
  - Node Caps
    - Data: Yes,No
  - Deg-Malig
    - Data: 1,2,3
  - Breast
    - Data: Left, Right
  - Breast-Quad
    - Data: Left-Up,Left-Low,Right-Up,Right-Low,Central
  - Irradiat
    - Data: Yes, No
  - Inv-Node

    - Data: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39.
  - Tumor-Size
    - Data: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50-54, 55-59.

Target

- No-Recurrence Events
- Recurrence Events

Datapoints: 286

Sample Data Point

Patient 1 :

- Irradiat:No
- Age:30-39
- Menopause:premeno
- Tumor-Size:10-14
- Inv-Node:0-2
- Node-Caps:No
- Deg-Malig:1
- Breast:Right
- Breast-Quad:Left Low

Prediction:

The algorithms aim to predict whether a patient has recurring breast cancer. Their personal information, such as age and menopause, along with their medical information give a good guideline for teaching the algorithm to ascertain their status.

Description of the structure

Normalization

The algorithms implemented are CART algorithms, otherwise known as classification and regression tree -however, only the classification aspect as been used in this instance. This means that data is represented and normalized differently.

The dataset priorly contained missing values which largely offset the algorithm. There were two solutions to this problem, namely

1. Discard the values
2. Replace missing values with random/general information

Decision trees significally suffer under these conditions, meaning that the accuracy of the alogrithm suffers dramatically due to missing values. A slight change in the representation of the data leads to completely different models, which mean different predictions.

So we removed the missing values at a cost of accuracy.

Data Split

The training data contains 70% of the dataset and the testing data contains the remaining. Validation data was not accounted during the creation of the model. The reason being that creating Random Forest is computationally expensive, the algorithm essentially creates n decision trees. This means that if validation data was created to learn the hyperparameters then the computation needed would drastically increase.

Algorithms

- Decision Tree
    - With Pruning
- Random Forest
    - With Random Generation

General Implementation

The algorithms were implemented using a rigid method,meaning trained to account for its disadvantages. Earlier it was explained that a slight change in a row or column changes the model.

These algorithms randomize the dataset and train the dataset on the randomized dataset so that it can predict the class without worrying about the order of the rows or columns

The models were pruned by chosing the highest occuring child leaf and/or a random leaf to increase flexibility and spread of the model

Decision Trees

Decision trees were implented using the ID3 algorithm with information gain(as opposed to Gini). The model prefers width over depth. This means that the algorithm decides to classify the majority of the child nodes. Training the tree this way allowed for the model to avoid overfitting where necessary.

Information regarding the accuracy of the model can be found here:
https://github.com/whiterose-fsociety/machine-learning/blob/version_1/machine_learning_assignment/src/decision_tree/Decision_Tree.ipynb

Random Forests

Rationale

The combination of learning models increases the classification accuracy – to create a model of low variance.

What are Random Forrests

Tree models are known to be high variance, low bias models – they are prone to overfit data. ID3 or CART are also relatively unstableChanging one row of the initial table can change the values for the information gain calculation.
Random forrest has proved to be one of the most useful ways to address the issues of overfitting and instabillity.

The Random Forrest approach is based on two concepts, called bagging and subspace sampling.

Bagging

Bootstrap aggregration.

Technique that combines the predictions from multiple machine learning algoritms together to make more accurate predictions than any individual model.

Method

- Create datasets of the same length with replacement
- Train a model on each one
- Take majority prediction model for unseen query
- We take the mean or median for regression tree models

    * Mode – Classification trees

* Mean – Regression Trees

<u>Analysis</u>

<u>What worked best ?</u>

The random forest performed significantly better, and not necessarily for a higher accuracy but higher reliability. The random forest fixes the lack of flexibility that the decision tree provides. Whenever a slight change in the dataset occurs the decision tree creates a different model. The random tree takes the mode of a number of models that were created, thus allowing the decision trees freedom to create different models.

Improvements

The algorithms can be improved by handling missing values a lot better than how they were implemented here. The models prune leafs but very innefficiently. A method that could make use of validation data to learn the hyperparameters and fix the model whilst maintaining a fair computation would optimize the algorithm

Credits

- Python Course
  - https://python-course.eu/Random_Forests.php
  - https://python-course.eu/Decision_Trees.php
- Tofti:
    - https://github.com/tofti/python-id3-trees
- YouTube Video Content
- Channel: StatQuest
  - Title
    - StatQuest: Random Forests Part 1 - Building, Using and Evaluating
  - Link:
    - https://www.youtube.com/watch?v=J4Wdy0Wc_xQ
- Channel: RANJI RAJ
  - Title
    - Machine Learning | Random Forest
  - Link

- https://www.youtube.com/watch?v=p2arCEdJ8Fk&t=563s