CIS6060: Final Project Report
Lisa Hoeg
Due: Fri. April 9

Building a successful decision tree classifier of disease state for application to medical testing

Background

While machine learning has allowed for the integration of many covariates into highly successful predictive models in many fields, many medical diagnoses instead depend on expert interpretation of fewer attributes. At present, Hepatitis C is diagnosed by an antibody panel, with current infection status being assessed by a test for HCV RNA (Schillie et al., 2020s). While overall diagnosis has become a matter of routine blood work, the extent of damage that infection has caused to the liver still requires surgical biopsy to assess with accuracy (Hoffmann et al., 2018). As surgical biopsy is physically and psychologically taxing to the patient, an ideal diagnostic model would be able to use the integrative strengths of machine learning models to predict disease status with blood tests alone (Hoffmann et al. 2018; Lichtinghagen et al., 2013). A professionally useful model would realistically take into account prohibitive costs of testing on the patient, and minimize resources required for discriminating between different levels of severity of liver damage. Lichtinghagen et al. (2013) and Hoffmann et al. (2018) have used a series of blood panels from donors and from patients with confirmed state of liver damage through biopsy to construct decision tree style algorithms that can support diagnosis of Hepatitis C and liver function.

Using the "HCV data" available on the UCI website (Dua & Graff, 2019), in this project I have built on the decision tree concept described by Hoffmann et al. (2018) to construct a tree that discriminates severity of liver damage without biopsy. Although the classification is not perfect, and falls somewhat short of the 95% accuracy of the less intuitive SVC model, the parallels between the decision tree and the ordered testing procedure already inherent to the medical profession demonstrate its value. The tree designed by this project could be used to support medical professionals in diagnosis of liver damage from Hepatitis C, guide treatment to best care for patients, and help prevent unnecessary surgery.

Methods

The data set used for this project was uploaded to UCI's Machine Learning Repository June 2020, and cites two research studies on machine learning for medical application as the source of data (Hoffmann et al. 2018; Lichtinghagen et al., 2013). The data set has 589 complete samples, of which 533 are from healthy donors, 20 have confirmed Hepatitis C, 12 have confirmed fibrosis of the liver, and 24 have confirmed liver cirrhosis. Of the 14 attributes included, the UCI description specifies patient ID, diagnosis category, age, sex, and ten blood panel laboratory values by abbreviation only (see appendix). The laboratory values are not described with a medically relevant interpretation: the precise test performed and range of expected values are not given. Although the missing information about medical relevancy will not impact the machine learning algorithms, applying the results to a practical decision structure will require additional research. The two journal articles referenced and other medical resources (Hoffmann et al. 2018; Lichtinghagen et al., 2013, Slack et al., 2010) have filled in the

remaining details. The data set included 26 samples with missing values. As human health factors vary in complex ways, it was decided to discard samples with missing data rather than impute their values.

The first algorithm used to classify samples according to disease state was modified from the "evaluate_svn.py" script provided for assignment 2 using SVC classifier from scikit-learn. As the authors of the original paper indicated, machine learning algorithms that can make use of all available data have high discriminatory success (Hoffmann et al., 2018). Following this line of reasoning, the support vector network algorithm will be used to create a target accuracy for subsequent decision-tree based analysis.

The support vector method will be performed four times, both with and without principal components analysis, for each of binary classification and multi-class. The binary classification only looks to distinguish donor from disease samples, and the multi-class procedure attempts to classify all disease subclasses of extent of liver damage. While it would be ideal to fully predict the extent of liver damage on the basis of blood work alone, the most important practical distinction is whether or not the sample is expected to have hepatitis C, because that distinction may be sufficient to encourage the patient to get a diagnostic biopsy. As suggested by Hoffmann et al. (2018), due to the small number of samples labelled with each disease classification (n[HepC]=20, n[Fibrosis]=12, n[Cirrhosis]=24) a Leave-One-Out Cross Validation (LOOCV) will be used to determine the predictive model used with each sample.

The second algorithm being explored in this project is the DecisionTreeClassifier from scikit-learn. As described above, not all attributes provided in the data set are expected to be influential in any classification algorithm applied, and in particular for a practical outcome to a decision tree, only a portion of tests should be considered. The first decision will use binary mapping as described above, a max depth of 1, and class weighting of 1:100 to select the attribute and threshold most useful in distinguishing whether or not the sample is likely to have hepatitis C, with preference being given to allowing false positives over false negatives. As with the SVC() algorithm and recommended by the source paper, LOOCV was used to determine the binary classification of the first split. Since LOOCV constructs a unique tree for each sample, the feature and threshold were saved and inspected, with the most frequently identified feature being extracted to use as the node feature, and the mean of the corresponding thresholds used as the overall tree's threshold.

The remainder of the decision tree was constructed using the multiclass labelling for training, and samples that pass the threshold of the binary classification. This procedure is expected to produce a tree that can reasonably predict severity of disease state with ease of practical interpretation by humans. Since decision trees have a tendency to overfit, but the attributes must remain distinct and unmapped for ease of interpretation, instead of PCA, feature selection will be performed. Using scikit-learn's feature_selection function, I will choose k=4 to select the top four attributes based on chi-squared statistics as pre-processing of the data, and max depth of 3, then aggregate the resulting LOOCV trees to produce a single diagnostic tree.

Results and Discussion

Principal components analysis, as conducted for the SVC pipeline, reduced the dimensionality of the attributes from ten to eight, while retaining 90% of the variance explained (figure 1). As the support vector machine was being used as an ideal of classification success, and the data

set was relatively small, reduction in variables was primarily for comparative purposes. Since minimizing computational resources required was not a goal of PCA, only decreasing by two variables was acceptable to continue downstream analysis.

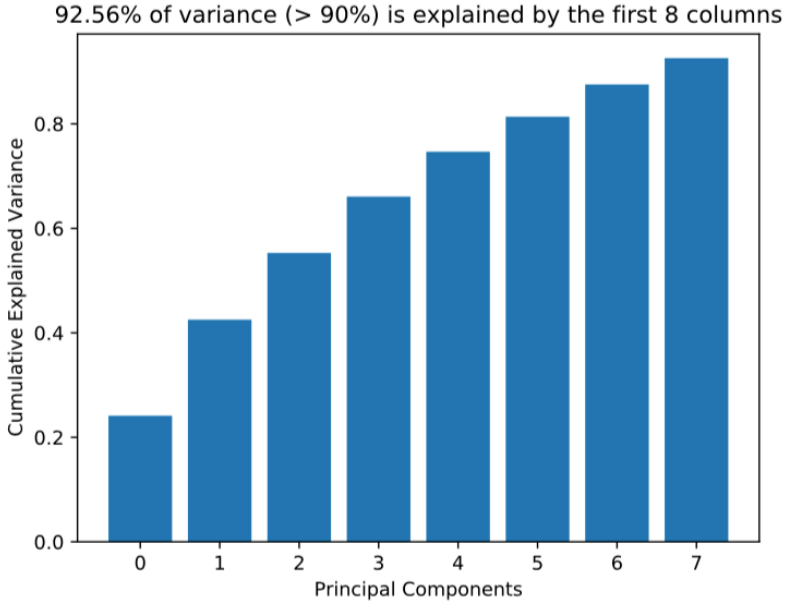
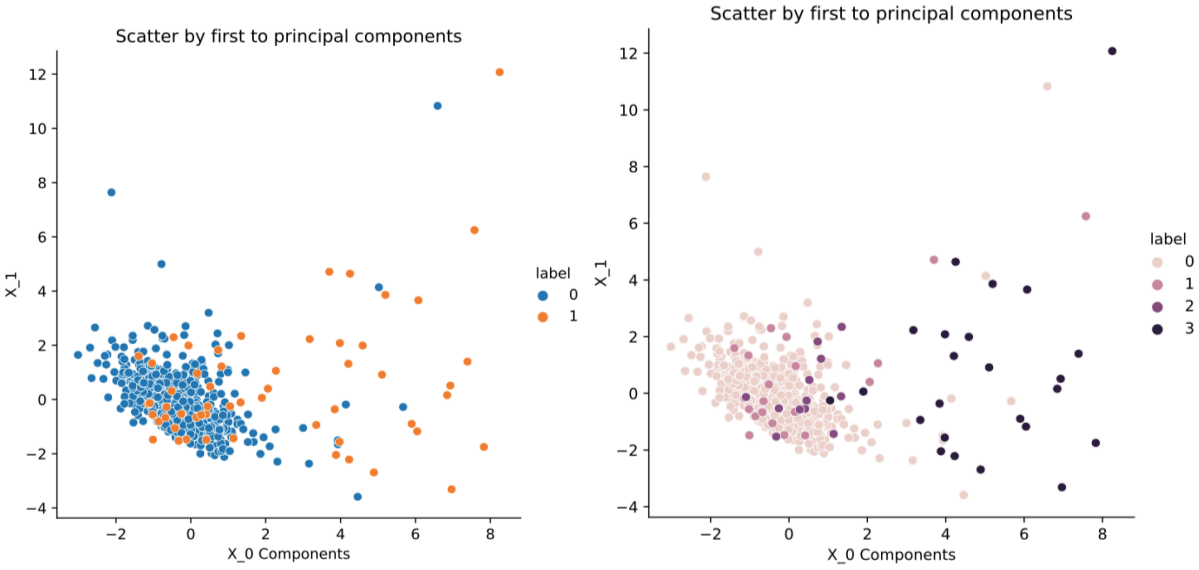


Figure 1: Cumulative variance explained by principal components analysis of HCV data set with 589 samples and 10 attributes in the original set. Dimensionality have been reduced to eight while retaining greater than 90% variance.

Projecting the samples into the dimensional space of the first two principal components, the binary and multiclass mappings of disease-state category are shown below in figure 2.



A **B**

Figure 2: Projection to 2-dimensional principal components using (**A**)binary mapping (0=donor; 1=disease) and (**B**) multiclass mapping (0=donor; 1=hepatitis only; 2=fibrosis; 3=cirrhosis). Donor samples well clustered, but less severe liver damage classifications are not well distinguished. Most distinctly separated samples appear to be category 3 (Cirrhosis).

Resulting confusion matrices of the SVC pipeline were as follows:

Binary, data scaling only:

```
Performance metrics:  accuracy 0.98, precision 0.90, recall 0.93
         .->    Donor Disease
  Donor  :       527       6
Disease  :         4      52
```

Binary, data scaling and PCA:

```
Performance metrics:  accuracy 0.98, precision 0.89, recall 0.89
         .->    Donor Disease
  Donor  :       527       6
Disease  :         6      50
```

Multiclass, data scaling only:

```
Performance metrics:  accuracy 0.95, precision 0.54, recall 0.52
           .->      Donor Hepatitis  Fibrosis Cirrhosis
    Donor  :         532         1         0         0
Hepatitis  :          11         5         2         2
 Fibrosis  :           2        10         0         0
Cirrhosis  :           3         1         0        20
```

Multiclass, data scaling and PCA:

```
Performance metrics:  accuracy 0.95, precision 0.55, recall 0.52
           .->      Donor Hepatitis  Fibrosis Cirrhosis
    Donor  :         532         1         0         0
Hepatitis  :           9         5         4         2
 Fibrosis  :           2        10         0         0
Cirrhosis  :           2         0         2        20
```

Within the binary classification problem, using support vectors to predict blood sample category was more accurate without PCA. This is not surprising as dimensionality reduction is primarily done to minimize computational requirements, and does include a small loss in percent of variance explained by the model. The multiclass problem was very similar with and without PCA. The iteration with data scaling only had more “Donor” category false predictions, and the PCA iteration had more “Fibrosis” category false predictions, but neither iteration of the algorithm was correctly able to identify any of the samples truly to the “Fibrosis” category. This could be because it was the category with the fewest number of samples in the set at 12, and the sample data was unable to train the model effectively.

The binary classification task as implemented with the decision tree model was designed to favour false positives on the assumption that a medically useful test would more likely encourage a patient to get additional testing for diagnosis confirmation rather than miss relevant conditions. As described in the introduction, when a patient is suspected of having Hepatitis C, the routine diagnostic procedure utilizes an antibody test, which is much more accurate than the tests described in this data set. The advantage of using the attributes as given in this data set to suggest likely diagnosis is that the blood tests used are very general and might be given to someone with no expectation that they should be testing for Hepatitis C in the first place. The discriminatory power of the decision tree being developed would be a warning sign to the health care professional to conduct a more targeted test.

The results of the binary classification are expressed as a confusion matrix below:

```
Performance metrics:  accuracy 0.78, precision 0.29, recall 0.95
         .->    Donor Disease
  Donor  :       405     128
Disease  :         3      53
```

Although the overall accuracy is significantly lower compared to the SVC binary classification (78% vs. 98%), the recall statistic indicating the percent of positive cases that were identified was slightly improved (95% vs. 93%). Limiting the tree's node depth to 1 and using LOOCV produced decision thresholds using feature AST in all cases, with a mean threshold of 30.35. The high recall comes at the expense of precision, which describes the percent of cases identified as positive which are true positives. In this case the precision and recall can be interpreted as 29% of cases labelled as positive are known to have Hepatitis C, and 95% of all cases with Hepatitis C were identified. The "positive" case based on this data might best be interpreted as an indicator for further medical testing rather than a diagnosis.

For the subsequent steps of the decision tree analysis, only cases with a passing threshold from the first step were considered, that is, cases with AST levels above 30.35. Dimensionality was reduced prior to implementing the classification algorithm using k-best features, and choosing k=4. The retained features were ALT, AST, BIL, and GGT.

The confusion matrix of the subsequent LOOCV decision tree analysis with a node depth of three is as follows:

```
Performance metrics:  accuracy 0.85, precision 0.57, recall 0.60
          .->      Donor Hepatitis  Fibrosis Cirrhosis
    Donor  :         123         4         0         0
Hepatitis  :           3        15         0         1
 Fibrosis  :           2         9         0         1
Cirrhosis  :           1         6         1        15
```

Running the same analysis of multiclass decision tree, but limiting maximum node depth to two produced the following confusion matrix:

```
Performance metrics:  accuracy 0.82, precision 0.56, recall 0.60
          .->      Donor Hepatitis  Fibrosis Cirrhosis
    Donor  :         119         8         0         0
Hepatitis  :           2        16         0         1
 Fibrosis  :           0        11         0         1
Cirrhosis  :           0         9         0        14
```

Although restricting the maximum node depth to two calculates a slight decrease in accuracy, the loss comes from misclassifying additional donor samples, as the disease state categories have modestly increased accuracy. As with the SVC method, Fibrosis was the worst performing category, with no correct classifications. Although the accuracy of the decision tree classifiers is decreased from the SVC algorithm, there are improvements to precision and recall.

The trees constructed from the classifier run with max node depth of three is shown below in figure 3, and the decision tree with node depth of two is shown in figure 4. Inspecting the trees, neither one used the GGT feature presented through the k-best selection method. The tree with node depth of 2 split the 0-level node with ALT, and used BIL and AST respectively at the subsequent nodes. The algorithm creating the tree with node depth of three found it more successful to further refine the three features ALT, AST, and BIL, rather than incorporating the fourth available feature, GGT.

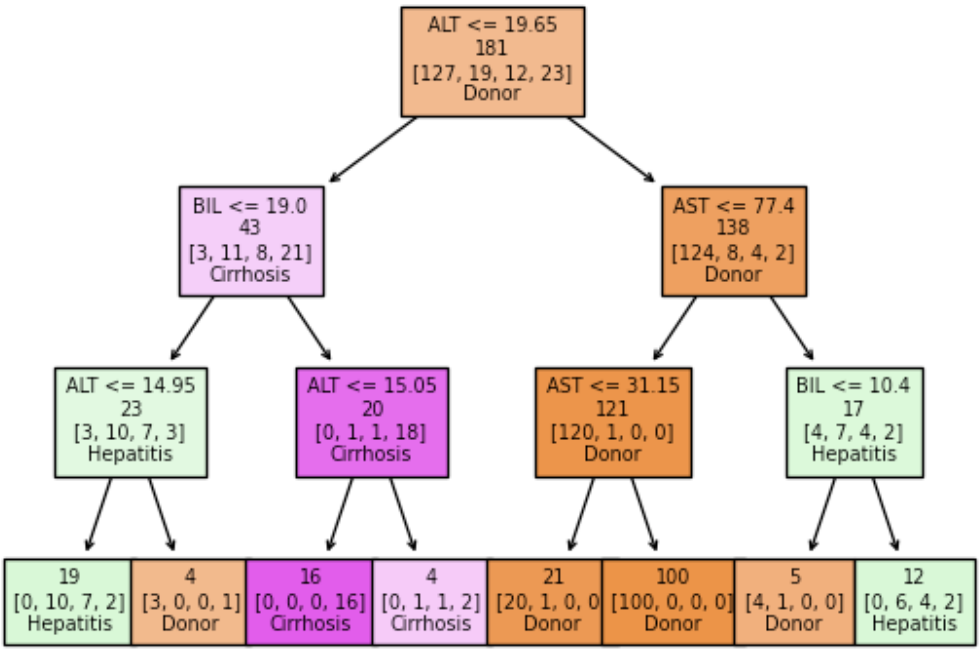


Figure 3: Decision tree produced by the scikit-learn tool set, with a node-depth of three. Data reflects predictions of extent of liver damage of Hepatitis C patients based on biochemical blood work. Interpretation of biochemical abbreviations can be found in the appendix, and represent traditional liver tests.

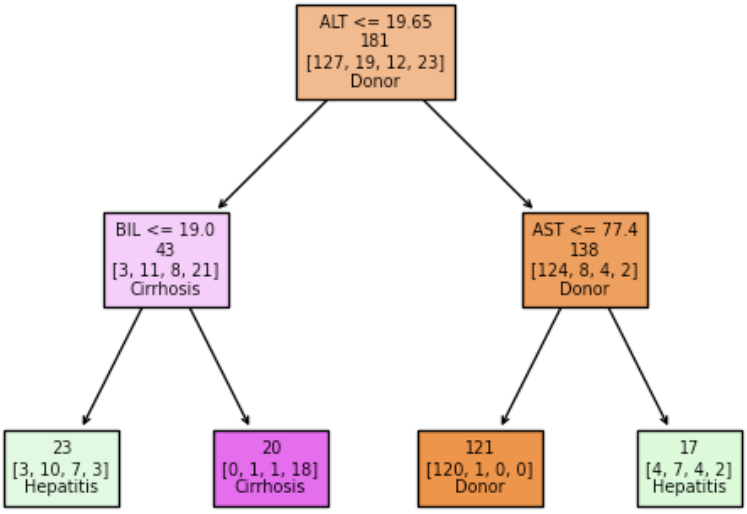


Figure 4: Decision tree produced by the scikit-learn tool set, with a node-depth of two. Data reflects predictions of extent of liver damage of Hepatitis C patients based on biochemical blood work. Interpretation of biochemical abbreviations can be found in the appendix, and represent traditional liver tests.

Conclusion

Since there is very little accuracy lost from node depth of 2 to node depth of 3, if the blood tests are to be performed sequentially I would recommend using the simpler tree of node

depth 2 to guide medical practitioners. Although the AST test is expected to be completed in the first split, if following the smaller tree in figure 4 and the ALT test result is greater than 19.65, then there will be no need to conduct the BIL test in order to make a best prediction between “Donor” and “Hepatitis” classification. If the tests are expected to be run simultaneously, the slight benefit to refinement of using the tree in figure 3 of node depth three might be useful to the professional conducting the tests.

The order of tests and thresholds recommended by the scikit-learn DecisionTreeClassification algorithm according to the procedure that I chose produced different trees than indicated by the Hoffmann et al. paper (2019). The tree produced by Hoffmann et al. (2019) had a node depth of 4, and used ALB, BIL, CHE, GGT, AST, ALT as their features of interest; see figure5 below. One key reason for the different results are that the data Hoffmann et al. used to construct their trees included 73 patients with known Hepatitis C diagnoses of varying severity of liver damage, and no control donor samples (2019). Hoffman et al. described their study as being more proof-of-concept to applying machine learning to improving non-invasive medical diagnostics, rather than a refined model for use (2019).

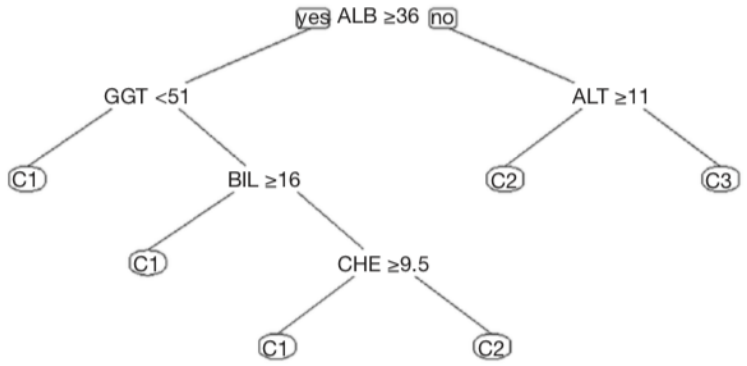


Figure 5: Decision tree presented by Hoffmann et al. (2019) as created with rpart Rpackage. Interpretation of biochemical abbreviations can be found in the appendix, and represent traditional liver tests.

In their decision tree analysis through rpart and ctree packages in R, Hoffmann et al. (2019) cite accuracy levels of 57.5% and 72.6% respectively using LOOVC. In this project, multiplying the accuracy of step 1 and 2 produces an overall accuracy of 64.0% (step1: 0.78; step 2: 0.82). Although, the metric of interest in step one was considered to be recall, not accuracy, so considering the product of the probabilities of interest, this project can be interpreted as having a success rate of 78.0% (step1: 0.95; step2: 0.82). In the first case of a strict measure of accuracy, this project produced comparable results to the original paper, but if considering the particular goals of each step, this project can be considered successful in designing a medically applicable, human interpretable, and simple decision tree algorithm for predicting status of liver health in Hepatitis C patients.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository: HCV data Data Set. Irvine, CA: University of California, School of Information and Computer Science. https://archive.ics.uci.edu/ml/datasets/HCV+data

Gordon, E. (2016). Children Exposed To Hepatitis C May Be Missing Out On Treatment. *NPR: Public Health*. https://www.npr.org/sections/health-shots/2016/07/26/468139416/children-exposed-to-hepatitis-c-may-be-missing-out-on-treatment

Hoffmann, G., Bietenbeck, A., Lichtinghagen, R., & Klawonn, F. (2018). Using machine learning techniques to generate laboratory diagnostic pathways—a case study. J Lab Precis Med, 3, 58.

Lichtinghagen, R., Pietsch, D., Bantel, H., Manns, M., Brand, K., & Bahr, M. (2013). The Enhanced Liver Fibrosis (ELF) score: Normal values, influence factors and proposed cut-off values. Journal of Hepatology, 59(2), 236–242. https://doi.org/10.1016/j.jhep.2013.03.016

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, B., Blondel, M. et al. (2011). Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Schillie S, Wester C, Osborne M, Wesolowski L, Ryerson AB (2020). CDC Recommendations for Hepatitis C Screening Among Adults — United States, 2020. *MMWR Recomm Rep* 2020;69(No. RR-2):1–17. DOI: http://dx.doi.org/10.15585/mmwr.rr6902a1

Slack, A., Yeoman, A., & Wendon, J. (2010). Renal dysfunction in chronic liver disease. Critical care (London, England), 14(2), 214. https://doi.org/10.1186/cc8855

Appendix

Attributes in the data are as follows:

1. X (Patient ID/No.) (Dua & Graff, 2019)
2. Category (diagnosis) (values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis') (Dua & Graff, 2019)
3. Age (in years) (Dua & Graff, 2019)
4. Sex (f,m) (Dua & Graff, 2019)
5. ALB = albumin (Hoffmann et al., 2018)
6. ALP = alkaline phosphatase (Lichtinghagen et al., 2013)
7. ALT = alanine amino-transferase (Hoffmann et al., 2018)
8. AST = aspartate amino-transferase (Hoffmann et al., 2018)
9. BIL = bilirubin (Hoffmann et al., 2018)
10. CHE = choline esterase (Hoffmann et al., 2018)
11. CHOL = cholesterol (Lichtinghagen et al., 2013)
12. CREA = creatinine (Slack et al., 2010)
13. GGT = γ-glutamyl-transferase (Hoffmann et al., 2018)
14. PROT = protein (Lichtinghagen et al., 2013)