D-Trees: Pre- and Post- Pruning

AW

Lecture Overview

′Chi-square′ Splits - the Idea

- We are still building D-Trees using Recursive partitioning scheme
- For now let's assume that
  - all features are nominal. We'll remove this assumption later
  - feature X_i has d_i values in its domain. After we remove first assumption it'll be d_i intervals of values, which does not change a thing
- So we must consider possible splits in each feature and a prospective split in a feature X_i generates d_i possible children of current node

Intuitively, if split makes sense, then data in the children nodes should be better classified than data in their parent.

- Treat both prospective child node (i.e. its number) and class as random variables on data in the parent.
- If split improves classification then class variable and node variable should not be independent
- Is observed data different from independence hypothesis? If there is no statistically significant difference, then no split needed.

Contingency tables

If a set of data points that meets certain conditions is classified according to two criteria of classification then independence of these classifications can be tested using χ^2-test on contingency table:

- If domain of classification variable X is entered in rows and domain of classification variable Y is entered in columns, then entry a_{ij} is the number of data points that are classified as $X = i$ and $Y = j$.

This two-way table is called a contingency table. The χ^2-test is applied to this table.

Contingency tables

The χ^2-test is applicable to:

- Two categorical random variables that are defined on the data such that
  - Domain of each variable is of size ≥ 2.
  - There is independence of observations in the sample:
    - Data is sampled from the same distribution independently (i.i.d.)
    - The random variables are not ”paired” in any way (e.g. pre-test/post-test observations).
  - Sample has relatively large size

In our case random variables are ’child number’ and ’class number’.
Then all of the above requirements are satisfied

Formal Setting for 'Chi-square' Splits

- Suppose that based on domain size of feature Y we consider r-way split of a parent node, that is child variable t has r values. Let also the class attribute c has p values. For a data point x we have $t(x) \in \{1, \ldots, r\}$ and $c(x) \in \{1, \ldots, p\}$.
- We test independence of t and c hypothesis H_0 against complement hypothesis H_1 that t and c are correlated
- After the split we get the pairwise statistic given by contingency table below that is the basis for testing independence hypothesis:

$$\begin{array}{c} \\ \text{Node t} \end{array} \begin{array}{c} \text{Class } c \\ \begin{pmatrix} n_{11} & n_{12} & \ldots & n_{1p} \\ n_{21} & n_{22} & \ldots & n_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r1} & n_{r2} & \ldots & n_{rp} \end{pmatrix} \end{array}$$

χ^2-test

- The Pearson χ^2-square test is a method for testing the association between the row and column variables in a two-way table. The null hypothesis H_0 assumes that there is no association between the variables (in other words, one variable does not vary according to the other variable), while the alternative hypothesis H_1 claims that some association does exist.
- The test statistic for the χ^2-square test of independence is:
  $\chi^2 = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
  where o_{ij} is observed value in the contingency table and e_{ij} is expected/theoretically computed value in the assumption that independence hypothesis holds.
- χ^2-statistic is distributed with χ^2-pdf with the number of degrees of freedom q which is how many many entries within the table can vary independently. We'll compute it next

Hypotheis Estimation

Given the contingency table of random variables Node Number vs. Class

- The sample probability estimate (frequency) of node taking value i is $\Pr(t = s) = \frac{\sum_{j=1}^{p} n_{sj}}{\sum_{i=1}^{r}\sum_{j=1}^{p} n_{ij}}$ i.e. the number of data points in child-node i over total number of data points in the parent
- The sample probability estimate (frequency) of class taking value j is $\Pr(c = k) = \frac{\sum_{j=1}^{r} n_{jk}}{\sum_{i=1}^{r}\sum_{j=1}^{p} n_{ij}}$ i.e. the number of data points in class k in the parent over total number of data points in the parent
- Under the assumption H_0 that node and class variables are independent the probability of class k occurring in child node s is $p_{sk} = \Pr(t = s)\Pr(c = k)$. If so then the expected number of data points of class k in node s must be $e_{sk} = p_{sk}n$ where $n = \sum_{i=1}^{r}\sum_{j=1}^{p} n_{ij}$ is the total number of datapoints in the parent node.

Degrees of Freedom

How many cell values could possible vary (= degrees of freedom) in contingency table, given our knowledge about the table?

- As many as we have table entries each entry is an independent variable, so $p \times r$
- But every row should sum up to known number of data points defined by split outcomes that cannot vary (n_j in node j). So r degrees of independence must be removed.
- Similarly each column should sum up to the total number of data points of a given of class (m_i for class i). Thus another p degrees of independence need to be removed.
- Total number of data points is also known and cannot vary within a split (i.e $\sum_{i=1}^{p} \sum_{j=1}^{r} n_{i,j} = |D_p|$), so p and r are related that we we counted one degree of freedom twice and we need to add it back, i.e. $q = p \times r - (p + r) + 1 = (p - 1) \times (r - 1)$

How Likely is Obtained Value of Statistic?

- For any value z of a random variable θ we have $p(z) = \Pr(\theta > z) = 1 - F_\theta(z)$ where $F_\theta(z)$ is cdf of θ
- It is known that χ^2 statistic with k degrees of freedom (sum of squares of k independent normally distributed standard variables) has pdf

$$f(x;k) = \begin{cases} \frac{x^{(k/2-1)}e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} \text{ if } x > 0 \\ 0 \text{ otherwise} \end{cases}$$

  where $\Gamma(q) = \int_0^\infty x^{q-1}e^{-x}dx$
- The p-value of a statistic is the probability of obtaining a value at least as extreme as the observed value. Say observed value is z, then as before $p(z) = \Pr(\theta > z) = 1 - F_\theta(z)$ where $F_\theta(z) = \int_{-\infty}^{z} f(\theta)d\theta$ and $f(\theta)$ is pdf of the statistic
- The lower the p-value, the more surprising the observed value is, and the more the grounds for rejecting the null hypothesis.

Accepting/Rejecting at Significance Level

- Significance level α is the level of p-value below which we are willing to reject null hypothesis. To compute acceptance use either tables or $\mathbf{R}$
- Tables: for known degrees of freedom k and significance level α find a value in a cell $t_{k,\alpha}$ in the χ^2-table. If observed value is below table value then accept null hypothesis

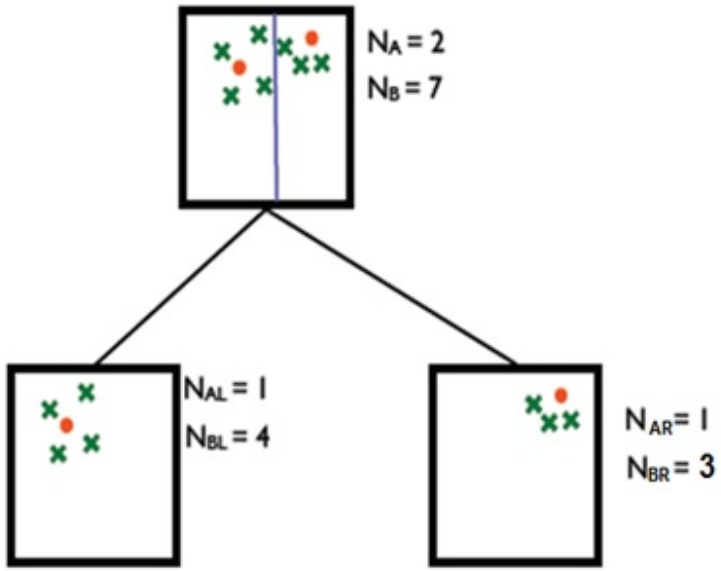
Degrees of freedom (df)	χ^2 value[19]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Example: let for $k = 3$ computed $\chi^2 = 10.55$, then at $\alpha = 0.01$ we should accept H_0 since $t_{3,0.01} = 11.34 > 10.55$

Same example in $\mathbf{R}$: `v<-10.55; z<-qchisq(1-0.01,3); t<-(v<z);t`

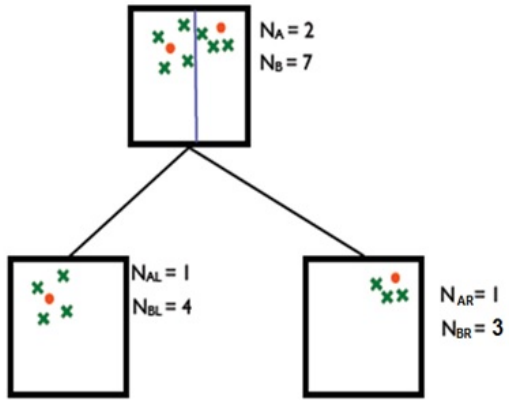
2-class Example - Binary Split

- Considering binary split of a node with $N = 9$ data points, with $N_A = 2$ data points of class A, and $N_B = 7$ datapoints of class B.
- If a split is allowed then left child gets $N_L = 5$ records, $N_{AL} = 1$ datapoints of class A and $N_{BL} = 4$ datapoints of class B, right child gets $N_R = 4$ datapoints with $N_{RA} = 1$ datapoints of class A, and $N_{RB} = 3$ datapoints of class B.
- We want independence hypothesis to be rejected at significance level $\alpha = 0.01$
- The contingency table is 2×2 so χ^2 has $k = (2-1)(2-1) = 1$ degrees of freedom



2-class Example continued

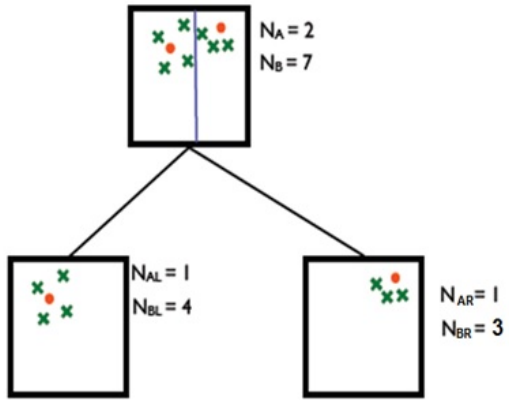
- In parent $N = 9$, we need $\alpha = 0.01$ and we have $k = 1$
- Frequency of children are $\Pr(L) = \frac{N_L}{N} = \frac{5}{9}$, and $\Pr(R) = \frac{N_R}{N} = \frac{4}{9}$. Class frequencies in parent node are $\Pr(A) = \frac{N_A}{N} = \frac{2}{9}$ and $\Pr(B) = \frac{N_B}{N} = \frac{7}{9}$.



  - So expected numbers of datapoints of each class in children are $e_{AL} = N \Pr(L) \Pr(A) = 9 \cdot \frac{5}{9} \cdot \frac{2}{9} = \frac{10}{9}$, $e_{AR} = N \Pr(R) \Pr(A) = 9 \cdot \frac{4}{9} \cdot \frac{2}{9} = \frac{8}{9}$, $e_{BL} = \frac{35}{9}$ and $e_{BR} = \frac{28}{9}$
  - $$\begin{aligned} \chi^2 &= \frac{(N_{AL}-e_{AL})^2}{e_{AL}} + \frac{(N_{BL}-e_{BL})^2}{e_{BL}} + \frac{(N_{AR}-e_{AR})^2}{e_{AR}} + \frac{(N_{BR}-e_{BR})^2}{e_{BR}} \\ &= \frac{(1-10/9)^2}{10/9} + \frac{(4-35/9)^2}{35/9} + \frac{(1-8/9)^2}{8/9} + \frac{(3-28/9)^2}{28/9} = 0.0321 \end{aligned}$$
  - For $k = 1$, at significance level $\alpha = 0.01$ value of z is 6.64 which is > 0.0321, so insignificant, i.e. independence hypothesis cannot be rejected, do not split!

2-class Example continued

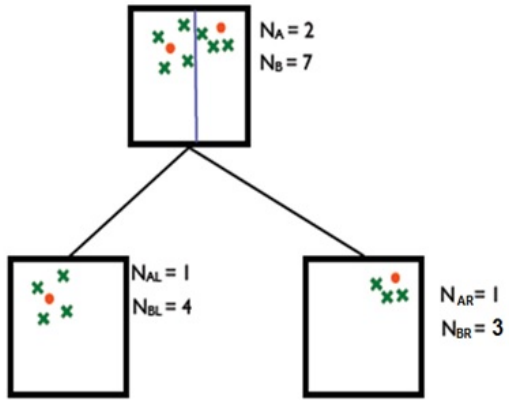
- In parent $N = 9$, we need $\alpha = 0.01$ and we have $k = 1$
- Frequency of children are $\Pr(L) = \frac{N_L}{N} = \frac{5}{9}$, and $\Pr(R) = \frac{N_R}{N} = \frac{4}{9}$. Class frequencies in parent node are $\Pr(A) = \frac{N_A}{N} = \frac{2}{9}$ and $\Pr(B) = \frac{N_B}{N} = \frac{7}{9}$.



- So expected numbers of datapoints of each class in children are $e_{AL} = N \Pr(L) \Pr(A) = 9 \cdot \frac{5}{9} \cdot \frac{2}{9} = \frac{10}{9}$, $e_{AR} = N \Pr(R) \Pr(A) = 9 \cdot \frac{4}{9} \cdot \frac{2}{9} = \frac{8}{9}$, $e_{BL} = \frac{35}{9}$ and $e_{BR} = \frac{28}{9}$
- $$\begin{aligned} \chi^2 &= \frac{(N_{AL}-e_{AL})^2}{e_{AL}} + \frac{(N_{BL}-e_{BL})^2}{e_{BL}} + \frac{(N_{AR}-e_{AR})^2}{e_{AR}} + \frac{(N_{BR}-e_{BR})^2}{e_{BR}} \\ &= \frac{(1-10/9)^2}{10/9} + \frac{(4-35/9)^2}{35/9} + \frac{(1-8/9)^2}{8/9} + \frac{(3-28/9)^2}{28/9} = 0.0321 \end{aligned}$$
- For $k = 1$, at significance level $\alpha = 0.01$ value of z is 6.64 which is > 0.0321, so insignificant, i.e. independence hypothesis cannot be rejected, do not split!

2-class Example continued

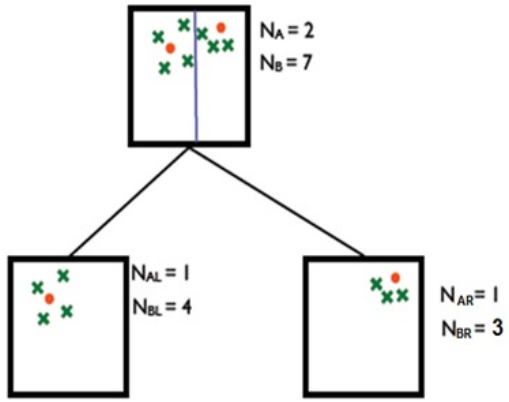
- In parent $N = 9$, we need $\alpha = 0.01$ and we have $k = 1$
- Frequency of children are $\Pr(L) = \frac{N_L}{N} = \frac{5}{9}$, and $\Pr(R) = \frac{N_R}{N} = \frac{4}{9}$. Class frequencies in parent node are $\Pr(A) = \frac{N_A}{N} = \frac{2}{9}$ and $\Pr(B) = \frac{N_B}{N} = \frac{7}{9}$.



- So expected numbers of datapoints of each class in children are $e_{AL} = N \Pr(L) \Pr(A) = 9 \cdot \frac{5}{9} \cdot \frac{2}{9} = \frac{10}{9}$, $e_{AR} = N \Pr(R) \Pr(A) = 9 \cdot \frac{4}{9} \cdot \frac{2}{9} = \frac{8}{9}$, $e_{BL} = \frac{35}{9}$ and $e_{BR} = \frac{28}{9}$
- $$\chi^2 = \frac{(N_{AL}-e_{AL})^2}{e_{AL}} + \frac{(N_{BL}-e_{BL})^2}{e_{BL}} + \frac{(N_{AR}-e_{AR})^2}{e_{AR}} + \frac{(N_{BR}-e_{BR})^2}{e_{BR}}$$
$$= \frac{(1-10/9)^2}{10/9} + \frac{(4-35/9)^2}{35/9} + \frac{(1-8/9)^2}{8/9} + \frac{(3-28/9)^2}{28/9} = 0.0321$$
- For $k = 1$, at significance level $\alpha = 0.01$ value of z is 6.64 which is > 0.0321, so insignificant, i.e. independence hypothesis cannot be rejected, do not split!

2-class Example continued

- In parent $N = 9$, we need $\alpha = 0.01$ and we have $k = 1$
- Frequency of children are $\Pr(L) = \frac{N_L}{N} = \frac{5}{9}$, and $\Pr(R) = \frac{N_R}{N} = \frac{4}{9}$. Class frequencies in parent node are $\Pr(A) = \frac{N_A}{N} = \frac{2}{9}$ and $\Pr(B) = \frac{N_B}{N} = \frac{7}{9}$.



  - So expected numbers of datapoints of each class in children are $e_{AL} = N \Pr(L) \Pr(A) = 9 \cdot \frac{5}{9} \cdot \frac{2}{9} = \frac{10}{9}$, $e_{AR} = N \Pr(R) \Pr(A) = 9 \cdot \frac{4}{9} \cdot \frac{2}{9} = \frac{8}{9}$, $e_{BL} = \frac{35}{9}$ and $e_{BR} = \frac{28}{9}$
  -

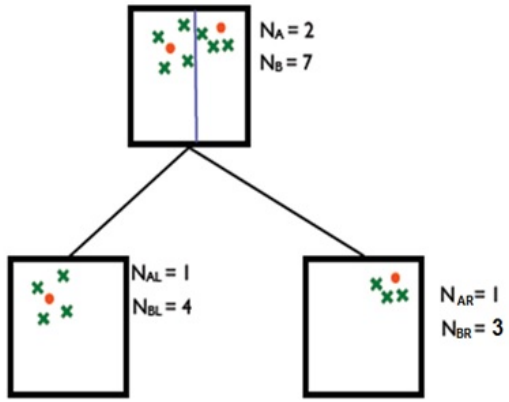
$$\begin{aligned} \chi^2 &= \frac{(N_{AL}-e_{AL})^2}{e_{AL}} + \frac{(N_{BL}-e_{BL})^2}{e_{BL}} + \frac{(N_{AR}-e_{AR})^2}{e_{AR}} + \frac{(N_{BR}-e_{BR})^2}{e_{BR}} \\ &= \frac{(1-10/9)^2}{10/9} + \frac{(4-35/9)^2}{35/9} + \frac{(1-8/9)^2}{8/9} + \frac{(3-28/9)^2}{28/9} = 0.0321 \end{aligned}$$

Should we reject Null hypothesis?

- For $k = 1$, at significance level $\alpha = 0.01$ value of z is 6.64 which is > 0.0321, so insignificant, i.e. independence hypothesis cannot be rejected, do not split!

2-class Example continued

- In parent $N = 9$, we need $\alpha = 0.01$ and we have $k = 1$
- Frequency of children are $\Pr(L) = \frac{N_L}{N} = \frac{5}{9}$, and $\Pr(R) = \frac{N_R}{N} = \frac{4}{9}$. Class frequencies in parent node are $\Pr(A) = \frac{N_A}{N} = \frac{2}{9}$ and $\Pr(B) = \frac{N_B}{N} = \frac{7}{9}$.



  - So expected numbers of datapoints of each class in children are $e_{AL} = N \Pr(L) \Pr(A) = 9 \cdot \frac{5}{9} \cdot \frac{2}{9} = \frac{10}{9}$, $e_{AR} = N \Pr(R) \Pr(A) = 9 \cdot \frac{4}{9} \cdot \frac{2}{9} = \frac{8}{9}$, $e_{BL} = \frac{35}{9}$ and $e_{BR} = \frac{28}{9}$
  -

$$\begin{aligned} \chi^2 &= \frac{(N_{AL}-e_{AL})^2}{e_{AL}} + \frac{(N_{BL}-e_{BL})^2}{e_{BL}} + \frac{(N_{AR}-e_{AR})^2}{e_{AR}} + \frac{(N_{BR}-e_{BR})^2}{e_{BR}} \\ &= \frac{(1-10/9)^2}{10/9} + \frac{(4-35/9)^2}{35/9} + \frac{(1-8/9)^2}{8/9} + \frac{(3-28/9)^2}{28/9} = 0.0321 \end{aligned}$$

  - For $k = 1$, at significance level $\alpha = 0.01$ value of z is 6.64 which is > 0.0321, so insignificant, i.e. independence hypothesis cannot be rejected, do not split!

Lecture Overview

CHAID - Discretizing Features

The idea is

- To convert features to have finite domain (discretize).
- Then the choice is made between multi-outcome splits: each feature is evaluated complete split (i.e. on all values of new domain).
- Each split is evaluated using χ^2-test. Only splits that reject independence hypothesis at the specified level of significance are retained. Among these splits the one that has highest p-value is chosen for tree expansion.

CHAID - Discretizing Features

CHAID (Chi-square Automatic Interaction Detection) decision tree learning uses the following simple algorithm to discretize features. Note that the algorithm has initial domain size d as a parameter:

- Partition data set D into d subsets of equal size by ordering data w.r.t to the discretized feature and assigning to data point x in position i domain value k where $(k-1) \cdot \frac{|D|}{d} \leq i \leq k \cdot \frac{|D|}{d}$
- Starting bottom up take two adjacent categories an from contingency table between the domain values and classes
- Apply χ^2 test to see if class variable is independent of domain variable. If it is, merge domains assigning all datapoints in both lower domain value and upper domain value to merged domain.
- This process is repeated until we can reject independence hypothesis for all pairs of categories. Take midpoints of feature values on both ends between categories as ends of discretization intervals.

Mock Example of Discretization

Suppose we have the following feature-class pairs of 17 data points:

(0.5,N), (4.3,N), (5,Y), (5,Y), (5,Y), (8,N), (12,N), (13,N), (15.01,N), (16,N), (16,N), (18,N), (24,Y), (26,Y), (26,Y), (28,Y), (32,Y).
The domain parameter with which we start is 4, so new domain is $\{1, 2, 3, \}$, and 4 or 5 data points have the same value (since $\lceil 17/4 \rceil = 5$). Then since the list is already ordered w.r.t. numeric feature new domain values are given in the format (new value, old value, class):

(1,0.5,N), (1,4.3,N), (1,5,Y), (1,5,Y), (2,5,Y), (2,8,N), (2,12,N), (2,13,N), (3,15.01,N), (3,16,N), (3,16,N), (,18,N), (4,24,Y), (4,26,Y), (4,26,Y), (4,28,Y), (32,Y).
Contingency table for new values 1 and 2 are:

	N	Y
1	2	2
2	3	1

We are ready to apply χ^2-test these categories. If the partitioning is not rejected then interval for domain value 1 is $(-\infty, 5]$

Example of (modified) CHAID in R

```
library(sets); library(partykit); library(TH.data)
data("GlaucomaM", package = "TH.data")
a<-dim(GlaucomaM)
set.seed(2)
x<-sample(1:a[1],a[1]/3,F) #randomly select record numbers
y<-as.integer(as.set(1:a[1])-as.set(x)) # take a complement
Train.GlaucomaM<-GlaucomaM[y,];Test.GlaucomaM<-GlaucomaM[x,]
    # training set; testing set
GlaucomaM.tree <- ctree(Class   ., data =Train.GlaucomaM)
    #build a tree model w training set
GlaucomaM.tree # show the tree
plot(GlaucomaM.tree) # plot the tree
pred.test.GlaucomaM <-
predict(GlaucomaM.tree,newdata=Test.GlaucomaM[-11])
    # classify test set
table(Test.GlaucomaM$Class,pred.test.GlaucomaM, dnn = c("Actual
class", "Predicted class"))
acc.GlaucomaM.tree <- 100*sum(pred.test.GlaucomaM==
Test.GlaucomaM$Class)/dim(Test.GlaucomaM)[1];acc.GlaucomaM.tree
```

Lecture Overview

1-level vs 2-level Decision Trees

- Fact: decision trees (ϵ, δ)-agnostically learnable by uniform convergence as long as the domains of features are finite.
- Limit class of decision trees even further: let $\mathscr{H}$ be functions expressible by 1 node majority tree or two level binary trees with root-node-associated split based on one feature.
- Let m be the number of features and $d_1, \ldots, d_m$ be domain sizes of these features. Then $d = \max_{i=1}^{m} |\mathcal{D}(d_i)|$ is the size of the largest domain $\mathcal{D}(d_i)$, and we have size bound $|\mathcal{H}| \leq 2^{md}$ for class $\mathscr{H}$. The sample size necessary for uniform convergence is at most $\frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \leq \frac{\log(2 \cdot 2^{md}/\delta)}{2\epsilon^2}$
- So suppose we are given training data S sampled from D with $|S|$ larger than necessary for uniform convergence. But we cannot afford to do complete enumeration of trees to find out the best one, so we used approximation method (like recursive partitioning) and learned some 2-level tree T. How good is our tree in terms of approximation of optimal tree?

Parameters for approximate classifiers

- When given (ϵ, δ) agnostic PAC learning algorithm returns a DTree that with confidence probability δ has accuracy $1 - \epsilon$
- We have an approximate algorithm (not agnostic PAC learning), so we want to know with the same confidence what would be the accuracy (or error) of the decision tree that it returned.

The problem setting: What we know is that

- Single node tree classifier has associated Bernoulli distribution of correct classification: e if incorrect and $1 - e$ otherwise, where e is unknown error
- A sample (estimated) error of this classifier is Y/N where Y is the number of incorrect classifications, and N is sample size. So estimated error approaches e as N approaches infinity
- In a two level classifier each leaf is a single node classifier, so everything above applies to leafs

What we need is

1. for a single node classifier bound e with a given confidence δ ;
2. For a two level binary tree classifier bound combined leaf error (some combinations of e_1 and e_2 that we need to define) with a given confidence δ.

Reading

Textbook 3.5.4, 10.4.1 pp. 792-794 SSBD Sections 3.2, 18.1
You can skip proofs if you are not interested.