06/10/15

DATA ANALYTICS

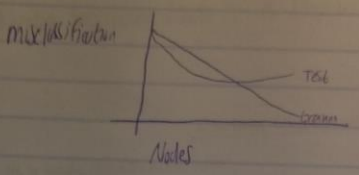
CLASSIFICATION AND REGRESSION TREES

- like Logistic regression

- Goal: how to derive splits
  grow the tree?
  What size tree?

- Do it by looking at all possible splits
- Decide the split
- Split with higher classification rate the "best"

13/10/15 DATA ANALYTICS



Drop in sample of data into each tree and get estimated of misclassification
Split data into say 10 parts and calculate misclassification rate, check lowest.

Tree opposite to Logistic reg. tree good for interaction
Logistic reg good for linear relationships.

COMPARISON
- Simple to capture liner relationship with logistic regression
- Tree not good at linear relationship
- Tree → Can use one variable many times.

INTERACTION
when level depend on effect of another variable - level of y from X_1 depends on X_2
will have term like $y = X_1\beta_1 + X_2\beta_2 + X_1X_2\beta_3$
↑ captured interaction

- Tree automatically include interactions
No interaction - just look at the marginal.
CASE 2 Example: No interaction no relationship
No relationship/interaction: tree will give just the root node

CASE 3: X_1 has an effect, doesn't depend on level of X_2 no interaction
tree: just one split

16/10/15 DATA ANALYTICS EXTRA NOTES

N_J Number in class J overall
$N(t)$ Total number of cases in node t
$N_J(t)$ Number of class J cases in node t
Proportion of class J cases in node t = $N_J(t)/N(t)$
$\pi(J)$ = prior probability $\Sigma \pi_T = 1$

Probability case will be both in class J and in node t $p(J,t) = \pi_J \frac{N_J(t)}{N_J}$

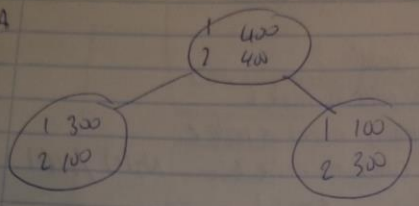
Probability of falling into class J given that a case is in node t
$p(J|t) = p(J,t)/p(t)$

Probability any case fall into node t = $p(t) = \sum_J p(J,t)$

Two common impurity funcs
when two categories in target variable:
Gini = $2 * p(1|t)p(2|t)$

How to choose Split?
- look at impurity of parent node t : $i(t)$
- each split gives 2 children t_L and t_R
- Measure impurity of the two children
$i(t) = p_L * i(t_L) + p_R * i(t_R)$
Smaller values of $i(t_L)$ and $i(t_R)$ are better.
- Choose split with largest $\Delta(t,s) \rightarrow$ R calls this improvement
R multiplies $\Delta(t,s)$ by N - total no. of cases

A



Gini = $2 * p(1|t) p(2|t)$

= $2 * (0.5)(0.5)$

= 0.5.

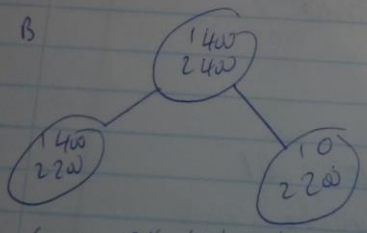
Impurity = $i(t) - P_L * i(t_L) - P_R * (i(t_R))$

$0.5 - 0.5 * (2 * 0.75 * 0.25) - 0.5 (2 * 0.25 * 0.75)$

$0.5 - .1875 - 0.1875$

$\Delta(t,s)$ = 0.125. CORRECT

B



Gini 1 = $2 * p(1|1) p(2|1)$ = $2 * 0.5 * 0.1 = 0.5$

Impurity $i(t) - P_L * i(t_L) - P_R * i(t_R)$

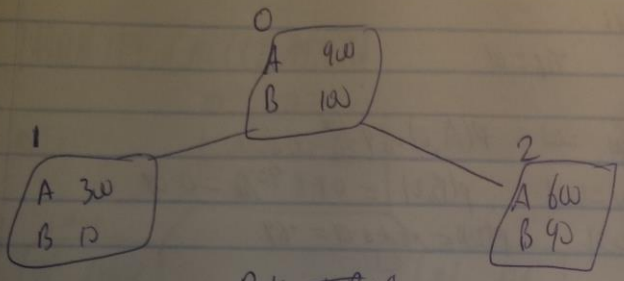
$= 0.5 - 0.75 (2 * \frac{4}{6} * \frac{2}{6}) - 0.25 (2 * 0 * 1)$

$0.5 - \frac{1}{3}$

$\Delta(t,s)$ = = 1/6 (0.1666)

Pick highest vote if $\Delta(t,s)$ | => Second split CORRECT

10/15 DATA ANALYTICS EXTRA NOTES



~~Data of Priors~~

$P(A|0) = 9/10$
$P(B|0) = 1/10$
$P(A|1) = 300/310$
$P(B|1) = 10/300$
$P(A|2) = 600/690$
$P(B|2) = 90/690$

Equal prior = multiply by 0.5 to make fair

$P(A|1) = 0.5 \times 300/900 = 0.17$ $P(B,2) = 0.5 \times 10/100 = 0.5$

$P(B,1) = 0.5 \times 600/900 = 0.33$ $P(B,2) = 0.5 \times 90/100 = 0.45$

$P(2) = 0.33 + 0.45 = 0.78$ $p(1) = 0.17 + 0.05 = 0.22$.

$P(A|1) = 0.17/.22 = .77$ $p(A|2) = .33/.78 = 0.42$

$p(B|1) = 0.05/.22 = 0.23$ $p(B|2) = .05/.78 = .58$

Assign A or B whichever has higher value

Data as given
$\pi_A = 0.9$ $\pi_B = 0.1$

$P(A,1) = 0.9 \times \frac{300}{900} = 0.3$ $P(A,2) = 0.9 \times \frac{600}{900} = 0.6$
$P(B,1) = 0.1 \times \frac{10}{100} = 0.01$ $P(B,2) = 0.1 \times \frac{90}{100} = 0.09$
$P(1) = 0.3 + 0.1 = 0.31$ $P(2) = 0.6 + 0.09 = 0.69$

$P(A|1) = 0.3/0.31 = 0.97$ $P(A|2) = 0.6/0.69 = 0.87$
$P(B|1) = 0.01/0.31 = 0.03$ $P(B|2) = 0.09/0.69 = 0.13$

Node 1 and 2 to A (higher certainty)

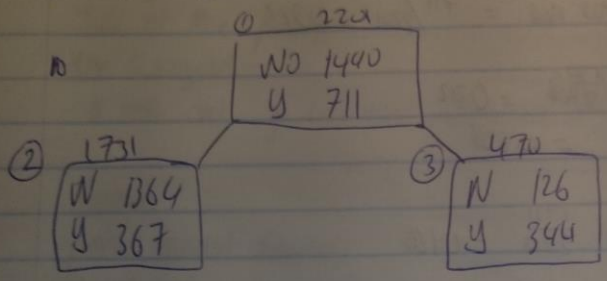
MISCLASSIFICATION
$r(t)$ prob of misclass at node $t = 1 - \max(p(i|t))$
$R(t) = r(t) * p(t)$
For the tree $R(T) = \sum r(t)p(t)$ T is terminal nodes

Same tree with data as given
Node 0: $r(t_0) = 0.1$
$R(t_0) = r(t_0) * p(t_0) = 1 * 0.1 = 0.1$
Node 1: $r(t_1) = 1 - 0.97 = 0.03$
$R(t_1) = r(t_1) * p(t_1) = 0.03 * 0.31 = 0.0093$
Node 2: $r(t_2) = 1 - 0.87 = 0.13$
$R(t_2) = r(t_2) * p(t_2) = 0.13 * 0.69 = 0.0897$

$R(T_1) = 0.0093 + 0.0897 = 0.099$

16/10/17. DATA ANALYTICS EXTRA NOTES

IMPURITY CALCULATIONS



Impurity root = $2 * p(N|root) * p(Y|root)$

$2 * 1490/2201 * 711/2201 = 0.437$

Node 2: $i(t) - P_L * i(t_L) - P_R * i(t_R)$

i node 2: $2 * (p(N|node2)) * (P(Y|node2))$

$2 * (1364/1730)(367/1730) = 0.33451$

Node 3 $= 2 * (126/470)(344/470) = 0.3424$

P(1) = 1

P(2) = $1731/2201 = 0.7860$

P(3) = $470/2201 = 0.2135$

Decrease in impurity = $i(root\ node) - [proportion\ left * i(left) + proportion\ right * i(R)]$

$0.437 - [0.786(0.334) + 0.214(0.392)]$

$= 0.091$

CALCULATIONS FOR CP

R(0) = misclassification for root node = $711/2201 = 0.323$

r(1) = $367/1731$ = 0.212

r(2) = $126/470$ = 0.268

R(1) = $r(1) * \frac{1731}{2201}$ = 0.166

R(2) = $r(2) * 470/2201$ = 0.057

R(Tree) = R(1) + R(2) 0.167 + 0.057 = 0.224

$cp = \frac{R(0) - R(Tree)}{\#nodes - 1}$ $\frac{0.323 - 0.224}{2 - 1}$ = 0.99

20/10/15 DATA ANALYTICS

Missing Data
- Talk at Mice package
- Use Surrogates -
- if most cases go left → send missing case to left.

Surrogate
- possible equivalent information
- A split that splits exactly like primary split
- Finding a split that splits in a similar way to primary split.
- Done at every node in the tree
- What is probability both split send data in same way - "agreement" in R
- Default rule - send down side with majority of data.

Large MSE - a factor being treated as quantative - error.
Example: improvement in order → tells # of missing cases
If missing the variable in normal split it will split by the extra variable in extra split.
Adjustment is same as association
Surrogate rule dependent.

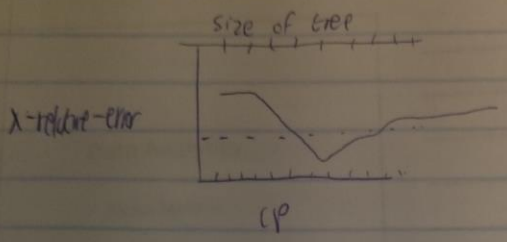
Be skeptical of variable importance - different packages have different values

4/5 DATA ANALYTICS

Sometimes you might want a high Sensitivity value and lower Specificity value → ie want more true ones than true zeros

May want to weight wrong classifications

23/10/15 DATA ANALYTICS



plot of complexity parameter CP - values going down
- dotted line - top end of standard deviation of smallest x-val-rel-error
- cross validation error - relative to the root node
- Put CP=0 for biggest value

ROC for random model is a 45° line

How to compare 2 ROC curves? - area under curve?
- Draw curves on top of each other

Why no Splits - may not be a good enough variable available
convert area to factor summary(defaults) will give frequencies
not min and maxes

27/10/15 DATA ANALYTICS 1(1)

Model Evaluation
looking at "operating condition" - prior probability etc.
ROC curve rarely actually curved - Assume we're taking area under jagged curve
p(+) prior probability of a pts cat

When overall class % is very small - model tends to over emphasise it.
Convex hull - looking for the slope of tangents being the same

Graph - Two models, A and C.
FN → FNR - bobe to save - create worry on slides
FP → FNP

New Graph - got an outline at bottom of graph with lowest possible error rates
How model behave for prior probabilities.

Costs important

		Predicu +	Predicu −
Actual	+	✓	x
	−	x	✓

So far we have treated X's the same → may want to weight it.

Want to know rates of x

Plot cost v cut off point for different cut off points

COSTS

- Total cost - numbers * Cost
- Can incorporate cost into growing or pruning trees.
- The two graphs - opposite of each other cost of false pos and false negative
- Only the ratio of costs is important.
- Gini/IG - measure impurity
- Cost is basically a weighting variable

Ecost - expected cost used in ROCR some"

MCA value - FN=1, FP=1

Telling us ranges in which model will work

10/15 DATA ANALYTICS

CABEL CABEL

Predicted values in two vectors [yes, no]

The graph of trees v regression
- Should expect something around a diagonal
- Tree model looks better, most of the dots are high

Decides - RPP in the R graph, different to my own way

What are you going to use for classification?
- Depends what question you ask / are answering
- Purpose of study
- Classification tool?
- What do we want to optimise?

03/11/15 DATA ANALYTICS

When fitting logistic regression - assuming a linear relationship

Seeing if there is a linear relationship in are variables

1 2 3 4 5 ← split data variable into 5 factor or level

yes
no

Carat Package Output

predicted v actual (reference) values

Accuracy: main diagonal over total

95% CI - for accuracy, "good idea" CI: $p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$ ← these should go here

roughly.

or bootstrap and take percentiles

wide b/c sample size is small

No information rate: how good do without model if I predict everything as yes

P-value: how better you did model than without it.

Kappa: (Ref and actual yes / total agreement) controlling for chance

Mcnemars Test p value: paired chi-squared test, suggests no difference in example probability of 1. looking at 6 and 5

Sensitivity:

Sensitivity Specificity: need to know which one is the event M or R or T or F or Y N

H_0: Acc $\leq$ NIR v H_1: Acc $>$ NIR

evidence against H_0, accept alternative hypothesis.

Pos pred value: percentage of predicted yes over all yes

Neg pred value: look at rows instead of columns

Prevalence: Proportion of event 27/51

better for medical use

03/11/15 DATA ANALYTICS

When fitting logistic regression - assuming a linear relationship

Seeing if there is a linear relationship in are variables

1 2 3 4 5 ← split data variable into 5 sections or 10 etc

yes
no

Carat Package Output

predicted v actual (reference) values

Accuracy - main diagonal over total

95% CI - for accuracy, 'good idea' CI: $p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$ ← where should go here roughly.

or bootstrap and take percentiles

wide b/c sample size is small

No information rate: how you'd do without model if I predict everything as yes

P-value: how better you did model than without it.

Kappa: (Factoring actual) yes/total agreement controlling for chance

Mcnemars Test P value: paired chi-squared test, suggests no difference in example probability of 1. looking at 6 and 7

Sensitivity:

Sensitivity Specificity: need to know which one is the event M or R or T or F or Y N

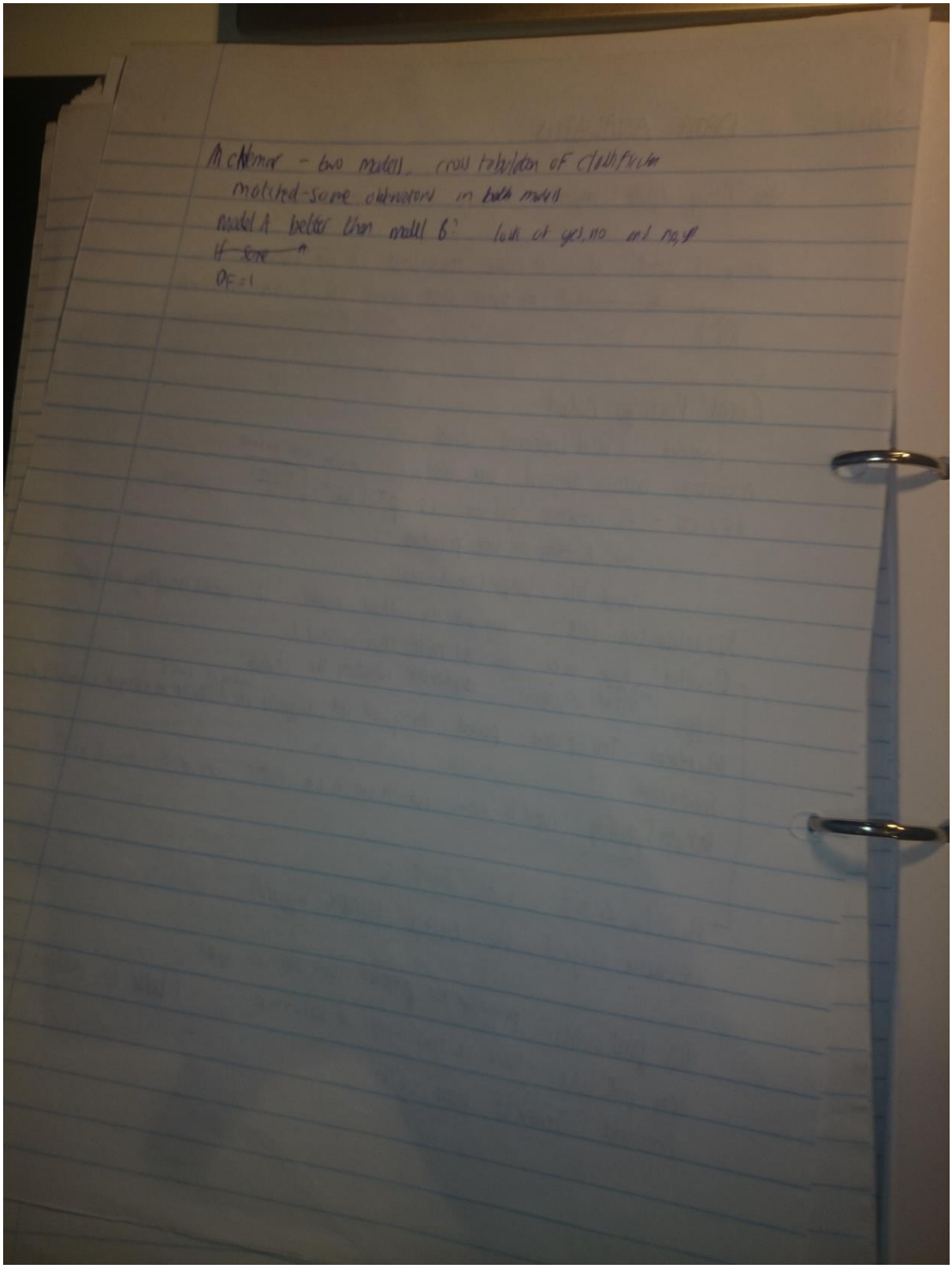
H_0: Acc $\leq$ NIR v H_1: Acc $>$ NIR

evidence against H_0, accept alternative hypothesis.

Pos pred value): percentage of predicted yes over all yes

Neg pred value: look at rows instead of columns / better for medical diagnosis

Prevalence: Prevalence of event 27/51


McNemar - two models, cross tabulation of classification matched-same observations in both models
model A better than model B? look at yes, no and no, yes
~~H one~~
$D_F = 1$

27/11/15 DATA ANALYTICS

Predicting probabilities from neural net

predict (fitnn, data.frame (class = "", sex = "", age = ""))

Ensemble Methods Foundations, algorithms Zhou, Zhuthua