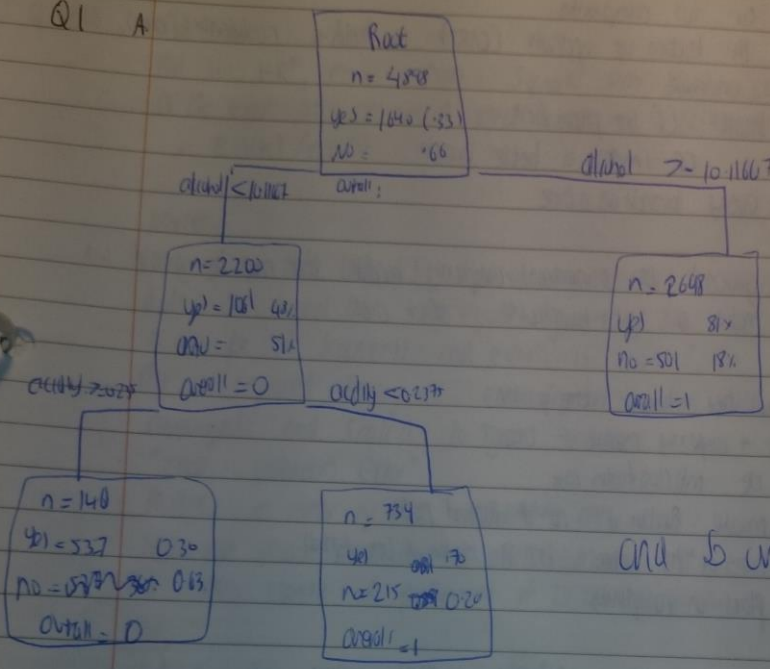
-

DA 2015 EXAM PAPER

Q1 A



and so on

First node - record the node number used for pruning in the tree
Split - highlight the rule used for each split at each node
n = number of cases in the particular node
loss = number of misclassified cases in each node (similar category)
yval = Given overall classification attribute for that particular node
yprob = Given probability of being assigned to class 0 and 1 respectively
* = denotes terminal nodes of the CART tree

B Explain CP, rel error, xerror and xstd

CP

- CP stands for complexity parameter
- CP measures our complexity which is a penalty placed on

each node for it's complexity
- It is used in the bottom up approach (CART) to building classification trees, to prune branches or nodes
- Nodes with smaller CP are pruned first
- Bigger values of CP implies a better branch
- Calculated at every branch in a tree

- The user gives the minimum complexity benefit that must be gained in order to make a split worthwhile called CP.

- Grow maximal tree - largest tree possible
- R_α = cost + complexity measure of tree T
- Here cost is the misclassification rate
- Complexity measure: function of no of terminal nodes.
- $R_\alpha = R(T) + \alpha * |T|$ where $|T|$ is number of terminal nodes
- α penalty placed on complexity

- For a single node t: $R(t)_\alpha = R(t) + \alpha$
- For a subtree T_t: $R(T_t)_\alpha = R(T_t) + \alpha |T_t|$
  When α increases
- Both $R(t)_\alpha$ and $R(T_t)_\alpha$ increase but $R(T_t)_\alpha$ increases faster.
- Value of α when $R(t)_\alpha = R(T_t)_\alpha$ price paid for complexity - ending on a tree

$$\alpha = \frac{R(t) - R(T_t)}{|T_t| - 1}$$

- Prune tree with lowest value of α and recalculate
- CP is α except it is divided by $R(0)$ - misclassification rate for the root node
- Set a stopping size on CP, possibly 0.01 to determine final tree

29/04/16

DA 2015 EXAM

Q 10.

Rel Error

This is $1-R^2$, root mean Squared error, Similar to linear regression this is the error on observations used to estimate the model

$= RSS(k) / RSS(1)$

Xerror

- Each of these trees is examined using 10-fold cross-validation, in which the data are divided into 10 equal segments; the tree is built using 9 of the 10 segments and error is assessed on the 10th segment. This is repeated leaving off each segment in turn and errors are then averaged and scaled to give xerror
- "cross validation error"
- Multiply rel error by our misclassification rate
- Xerror rate of 0.9 means that the misclassification rate is 0.9 times misclassification of root node
- More realistic estimate of performance of the tree on new samples of data.

xstd

- xstd is the variation between the 10 subsample estimates
- Ideally we pick the tree with the lowest xerror ± 1 s.d.

- rel error is estimated with the training data - the sample used for estimating the tree - and thus it decreases as the tree increases, because the tree branches more and more adjusted to the data.
- This apparently better performance should not be taken for "real" when predicting for a new sample of data because larger trees tend to overfit on training sample and will hardly generalise well on new data samples!

- Chose tree with 4 splits
- CP at 0.01
- Xerror and rel error at 0.11
- xstd small value 0.02

1c. Primary Split
- A primary split at a node is the rule which determines where the data should flow → i.e. to left or right subtree
- This rule is used on all rules where the variable missing to split on is available.
- A split is evaluated and determined as primary based on the concept of impurity
- A perfectly 'pure' split will assign all cases to one class, while a 'impure' node will assign one in a 50:50 split between class
- The goodness of fit is defined to be the decrease in impurity
- Commonly used impurity functions are Gini or entropy
- It looks at the largest class in a dataset and serves to isolate it from the other classes

- look at impurity of parent node
- Each split gives two children
- Measure impurity of the two children
- Choose the split with the largest decrease in impurity - R calls this improvement.

Surrogate Split
- For each split node, the "primary splitter" is the variable that best splits the node minus the party of the resulting child nodes
- When the primary splitting variable is missing for an individual observation, that observation is not discarded, instead a surrogate splitting variable is sought.
- A surrogate splitter is a variable that partners within the dataset, relative to the primary variable, is similar to the primary splitter.
- Thus, the program can be best available information in the face of missing values.
- The surrogate may have a different cut off point from the primary splitter but the number of cases the surrogate split sends into left and right nodes should be close to that of the primary split.
- By default, CART analysis produces 5 surrogate variables as part of its standard output.

20/04/16

2015 EXAM - DA

2 C. Show How ROC Curve is constructed

This is a plot of the true positive rate vs false positive rate for different possible cut off points of a diagnostic test.

- Need to calculate the TPR and FPR (Sensitivity vs 1-Specificity)
- Draw up a confusion table, using X as a cut off point
- Using the confusion table, derive Sensitivity = TP/(TP+FN) and Specificity = TN/(TN+FP)
- The TPR = Sensitivity and FPR = 1-Specificity
- Repeat for multiple values of X and draw the curve

Shows:

- Trade off between sensitivity and specificity (Sensitivity increases, specificity will decrease)
- Closer the curve follows the left and upper border, means a more accurate test
- Area under the curve is a measure of test accuracy

D Error rate vs P(+)

- This is a plot of Error versus P(+)
- P(+) is the proportion of + in population
- TPR and TNR are not dependent on prior probabilities but accuracy is:

$$Acc = p(+) * TPR + p(-) * TNR$$

$$Err = 1 - Acc = 1 - (p(+)TPR + p(-)(1-FPR))$$

$$= p(+)FNR + p(-)FPR$$

$$= (FN-FP)p(+) + FP$$

- Point in ROC space are going to be mapped onto line in "error/p(+)" space
- Allows us to see very clearly how error rate varies according to p(+)
- Diagonal line represents curve when test always positive and go to (1,1) line (test was always negative)

Rule $Err = FP$, $p(+) \to 0$ $\quad Err = 1 - TP$, $p(+) = 1$ (FN)

- Cost curve plots how error rate varies as a function of the prevalence of positive examples
- Generally the ideal when misclassification costs are not equal

- There is a point/duality between ROC space and cost space meaning that a point in ROC space is represented by a line in cost space and vice versa
- A classifiers operating range can be immediately read off a cost curve - it is defined by the PDF formed when the cost curve intersects the diagonal line representing trivial classifiers
- Classifier performance where curves are middle band

Q3 A What is an ensemble? Advantages?

- Ensembles (or committees) are machine learning methods that use or combine multiple models to achieve better prediction accuracy than any individual model can on its own
- Eg 100 classifiers with an individual error rate of 0.4 have 0.03 error rate overall
- Made up of 1000+ etc low accuracy models to boost predictive power
- The output of an ensemble is determined by using system. Each model in an ensemble propels an output and the most popular answer is the overall output of the ensemble

Advantages

- Higher accuracy
- Many types of variables accepted
- More outputs than just multiclassification rate (calculations etc)
- Can use proximities for MDS
- less overfitting
- Unlikely that all classifiers will make same mistake
- So long as each error is made by a minority of classifiers, optimal classification will be achieved
- Greater predictive power compared with individual net
- Random forest, Boosting, Bagging, Rotation

30/04/16 EXAM 2015 Q1 9

Q3 B Difference between Boosting and bagging

Bagging - Bootstrap aggregation
- Generate B bootstrap samples of the training data: random sampling with replacement
- Train a CART tree using each bootstrap sample
- For classification - majority vote
- For regression - average of predicted value
- Bagging minimise error for me, bias and variance
- Way to decrease the variance of your prediction by generating additional data
- By increasing size of your training set, you can't improve the model predictive but can decrease variance, narrowly tuning the prediction to expected value
- PARALLEL - Aim to decrease variance not bias

Boosting
- Iteratively learning weak classifiers
- Technique for combining multiple base classifiers whose combined performance is significantly better than that of any of base classifiers
- Sequential training of weak learners
- Each base classifier is trained on data that is weighted based on the performance of the previous classifier
- Each classifier votes to obtain final outcome
- SEQUENTIAL - Aim to decrease bias, not variance

Boosting Adv/Disadv
- Powerful classification algorithm
- Can achieve similar classification results with much less tendency to overfit parameter or setting
- Can be sensitive to noisy data and outliers
- less susceptible to overfitting problem
- Feature selection results in relatively simple classifier
- Instead of re-sampling, uses training set re-weighting
- Possibly a sub optimal solution

- lose the simple interpretability of classification trees
- Computation more difficult, slower
- No prior knowledge needed about weak learners
- No parameters to be except t - number of trees
- From empirical evidence, Adaboost is particularly vulnerable to random noise

- Boosting performs an exhaustive search for best predictor to split on, whereas RF must only search a subset of data
- Boosting grows trees in series, with later trees dependent on the results of previous trees, RF grows in parallel, independent of one another

3 C Explain How Random Forests are constructed

- We assume that we know about the construction of single classification trees
- RF is an ensemble method based on classification trees
- Numerous trees are built using different training sets so as to create different results
- The ensemble can predict according to majority vote

Each tree grown as follows:

1 - If # of cases in training set is N, sample N cases at random but with replacement, from original dataset. This sample will be dataset for growing the tree
2 - If there are M input variables, a number $m << M$ is specified such that at each node, m variables are selected at random out of M and the best split on these m is used to split the node. The value of m is held constant during the forest growing
3 - Each tree is grown to the largest possible extent - No pruning

- Reducing m reduces both the correlation and the strength
- Using oob error rate to get value of m
- OOB for each tree, around 36% of data not used - used to calculate error rate

DA - 2015 EXAM

1 C.

Improvement:
- Relates to CP
- R call to decrease in impurity improvement
- How good be additional split is correctly dividing the data
- Measure sum of Gini or Entropy

Adj:
- Called Association
- (default misrouth - surrogate misrouth) / default misrouth
- Simply a measure of similarity
- Reliable reduction in error obtained by using surrogate splits to predict cat instead of using data % for left and right
- Percentage reduction in error obtained by using surrogate to predict the primary variable

Agreement
- The amount that both splitter sent cases in to same direction
- Used for evaluating quality/accuracy of surrogate splitter against primary splitter.

2 A Confusion Table
- Also known as an error matrix, is a specific layout that allows visualisation of the performance of an algorithm
- Each column of the matrix represents the instance in an actual class
- Name stems from the fact that it makes it easy to see if the system is confusing two classes
- Given multi-class case of a given model
- Constructed by running the test data through the model then compare the predicted class to the actual class, creating a 4 by 4
- often used to describe the performance of a classification model on a set of test data for which the true values are known.

	Predicted Yes	Predicted No
Actual Yes	TP	FN
Actual No	FP	TN

2.8 Assessing A misclassification Table

- A list of Statistics which can be calculated from the confusion matrix:
- Accuracy: Overall, how often is the classifier correct?
  = (TP+TN) / (TP+TN+FP+FN)
- Misclassification rate/Error rate: Overall how often is it wrong?
  = (FP+FN) / (TP+TN+FP+FN)
  = 1 - accuracy
- True Positive Rate: When it's an actual yes, how often does it predict a yes?
  = TP / (TP+FN) True Positive / Actual yes's
  • Also known as Sensitivity or recall
- False Positive rate: When it's actually no, how often does it predict yes?
  = FP / FP+TN FP/Actual no's
- Specificity - When it's actually no, how often does it predict a no?
  = TN / FP+TN TN / Actual no's
  • equivalent to 1 - FPR
- Precision - When it predicts a yes, how often is it correct?
  TP / TP+FP • TP/Predicted yes's
- Prevalence - How often does the yes condition actually occur in our sample?
  Actual yes/total = ~~TP+FP~~ (TP+FN) / (TP+TN+FP+FN)
- Balanced Accuracy
  (Sensitivity + specificity) / 2
- Run a McNemar test which follows a Chi-Square distribution.
- Kappa Statistic - agreement between raters/models
  - how at agreement taking into account the accuracy that would be generated by chance
  Kappa = (O-E) / (1-E) O: observed accuracy
  E: expected accuracy
  Values between 0 and 1, larger are better

3D PARTIAL DEPENDENCY PLOT

- Shows how each predictor is related to response, holding other variables constant
- Let x be the initial predictor of interest with r distinct values
- Construct r datasets for each of the r values of x leaving all other variables unchanged
- For the r datasets, predict the response using random forest
- Calculate a single value averaged over all observations for each dataset
- Average these predictions over trees
- Enable us to visualise interaction between 2 variables
- If each explaining variable contributed additively to the target y, by some unknown function, this method is great to show that estimated hidden function
- For 2 categories - plots will be mirror images of each other

3E. RF vs Rulefit

- A mixture of regression and trees. RF only uses trees
- RF are/is good at picking up linear relationships, but rulefit uses (trees) for non-linear relationships and linear regression for delebey linear relationships
- Rulefit grows a series of trees like RF but can use indicators from the variables to form a logistic regression model.
- The important means that rulefit uses includes a combination of the importance from trees and linear regression RF just the trees

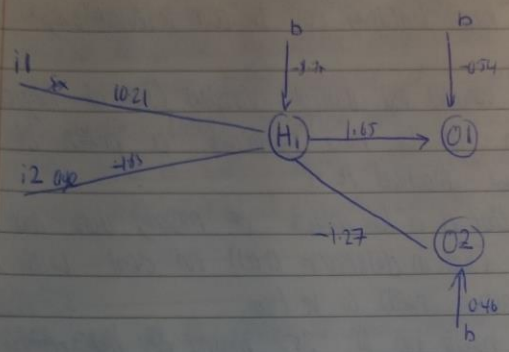
- First component produces 'rules' and the second component fits a linear model with rules as inputs
- Highly interpretable, decision rules have an easily understandable format

01/05/16

DA 2015 EXAM

12

Q4 A. Draw neural net, probability of adult male Lincoln



sex: no 0, yes 1 — gender = 1

Age = 0

ADULT MALE

$A1 = -8.75 + 10.21(1) + (-1.81)(0) = 1.46$

$F(a_1) = \frac{e^{1.46}}{1+e^{1.46}} = 0.8115$

$O1 \Rightarrow 1.65(0.811) - 0.54 = 0.799$

$F(O1) = \frac{e^{0.79}}{1-e^{0.79}} = 0.689$

$O2 \Rightarrow -1.27(0.8115) + 0.46 = -0.57605$

$F(O2) = \frac{e^{-0.57}}{1-e^{-0.57}} = 0.361$

- Calculate NB node with [illegible] value $\Rightarrow$ 0.799

ADULT FEMALE

$A1 = -8.75$

$F(a_1) = \frac{e^{-8.75}}{1+e^{-8.75}} = 0.000158436$

$O1 \Rightarrow 1.65(0.000158) - 0.54$

$F(O1) = \frac{e^{-0.54}}{1+e^{-0.54}} = 0.3662$

$O2 \Rightarrow -1.27(0.000158) + 0.46 = 0.454$

$F(O2) = \frac{e^{0.454}}{1+e^{0.454}} = 0.6129$

4.3 Missing Data

Raw Data will almost always need to be cleaned before analysis to find and deal with erroneous and missing data. To find data errors will be potentially damaging to model building can be done in several ways:

- Counts of NA's, blank variables and illogical variables (such as negative one etc) can give an indication as to how significant a problem of missing data is and also how localised it is
- Could be the case that there is a low amount of missing data and the model being constructed (such as classification trees) can deal with such a problem so no direct action needs to be taken
- Trees can theoretically handle up to 25% missing data and remain effective Neural nets cannot and the data must be dealt with
- One way of dealing with missing data is to impute a value to fill in the gap. Replacing a missing value with a mean or median value can neutralise the effects of missing data but also can bias the results towards the center. If there is a lot of data that has been imputed to the mean, the standard deviation of the data will be falsely low and the variable may no longer be valuable
- If there are a large number of cases of missing the same variable, it may be possible to remove variable entirely
- It may be preferable to use a surrogate variable, or in the case of CART a Surrogate Split. These are splits that give similar results and should be used when the information is difficult or too expensive to gain from the original variable (when and explain why there is so much missing).
- If a case is missing data, the entire case could be deleted - listwise deletion with sufficiently large data sets, this will have a negligible impact on the ultimate aim. Removing too many observations (and reducing sample size) will reduce the confidence we can place on our model and ultimately its predictive power
- NN cannot deal with missing data, and as such more extreme ad hoc such of variable or case deletion may be necessary to ensure the model can be accurately continued

21/05/16

DA: 2015 EXAM PAPER

CART vs LOGISTIC REGRESSION

Data

- CART analysis allows you to perform analysis on a highly skewed messy dataset
- LR can be affected by messy data and outliers
- CART can handle both categorical and quantitative data, whereas LR can only handle binary categorical data
- CART tends to work better with local datasets, whereas LR tends to work better globally (parametric vs non-parametric)

Detection

- CART excel at the detection of local structure. Even half of the tree is analysed separately CART automatically disregards insignificant variables
- CART automatically identifies interactive terms
- CART bad at detecting linear structure and cannot represent it effectively
- LR good for linear relationships. Many non-linear structures can be reasonably approximated with a linear structure
- LR needs to be adjusted manually with different combinations of variables - no detection of interactive terms

Assumptions

- The decision tree assumes the splits are axis parallel and will become more complex with the increase in number of factors and multiple decision boundaries are possible
- LR assumes there is only one decision boundary that is smooth and non-linear.

Overfitting

- Complex decision trees may overfit the data and they will become unstable (could prune tree)
- LR is intrinsically simple, it has low variance and so is less prone to overfitting

Speed
- While both algorithms are fast, LR has a faster run time on large data sets than decision trees

Interpretability
- CART and LR - nice probabilistic interpretation
- CART only to visualize and interpret for non-statisticians.

Transformation
- CART does not require transformations of variables with log or square root etc
- LR will light or point towards

How can they be combined
- Run a shallow tree
- Assigned each case to a terminal node
- Treat terminal nodes as categorical variables.
- Feed back into LR model