22/04/16 DA - EXAM QUESTIONS ON TREES

2010 Q1

A Difference between training and test set

- The training set is used to grow the tree and to decide on splits
- The test set is used to evaluate the tree and to obtain an estimate of the misclassification rate
- The validation set is used to select which tree to use
- Split data into 50/25/25
- If the training set is very like the data, the probabilities predicted by the model will be the relative ~~of~~ of the real ones

- training set to decide on splits
- validation to grow trees
- test to evaluate trees

B Type of Data - Importance when evaluating splits

- The types of data determines which splits are possible, and thus, the number of possible splits
- For categorical variables, all splits are possible and, where there is a number of classes For example, if there are 3 classes, A, B, C The possible splits available are

  A v BC B v AC or C v BA

- For ordinal / interval variables, only distinct splits are possible. For example if there are 3 classes (1, 2, 3) the possible splits are ~~1 v 2 3, 1, 2 v 3 or 1, 3 v 2~~
  i.e. order is ~~not important~~ important only have 1, 2 v 3 or 1 v 23
- For continuous data the splitting value can take any number, but necessary within the range of values

D What does CP measure?

- CP measures cost complexity which is a penalty placed on each node for its complexity
- It is used in the bottom up approach (CART) to building classification trees, to prune branches or not
- Nodes with smallest CP are ~~removed first~~ removed first.
- Bigger value of cp implies a better branch
- Calculated at every branch in the tree

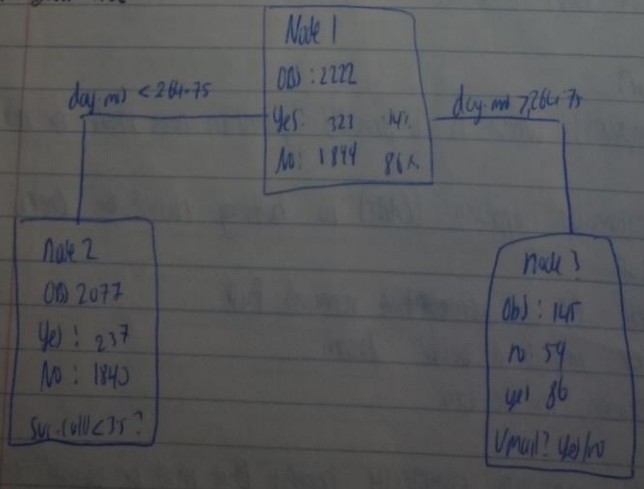
- The variable governs the minimum complexity benefit that must be gained in order to make a split worthwhile Default 0.1

- The CP is used to control the size of the decision tree and to select optimal tree size
- If the cost of adding another variable to the decision tree from the current node is above the value of CP, then tree building does not occur
- setting CP to zero will build a tree of maximum depth

- Grow maximal tree - largest possible tree
- R_α = cost + complexity measure of tree T
- Define cost as the misclassification rate = R(T)
- Complexity measure: number of # of terminal nodes
- $R_\alpha = R(T) + \alpha * |T|$ where α is penalty placed on complexity $|T|$ is # terminal nodes

E What is Xerror used for?
- The cross validation error is used to suggest an optimal sized tree
- It is used in the top down approach to building a classification tree as a sign of when to stop growing the tree
- Also used in the bottom up approach when pruning → Acts as an indicator to stop pruning
- In both cases, when the value of Xerror starts to increase, it is suggesting the optimal solution has been reached
- Acted with CP to determine CP value to choose

F. Draw Tree



DA - TREE EXAM QUESTIONS

- What is a surrogate split?
- A surrogate split is a predictor variable that's associated with the primary predictor variable(s).
- It splits in a fashion similar to the primary if too expensive or difficult to gather.

- A variable with possibly equivalent information
- A splitter that splits in a fashion similar to the primary splitter
- Reveals structure of the info in the variable
- If the primary (s) is too expensive or difficult to gather - use surrogate S* instead
- Use surrogate split if data is missing
- Surrogates are node dependent calculated at a local level
- Allows CART to deal with missing values in the future
  agree - measurement of strength of surrogate

H Interpret: tmul.mlys <65 to the right, agree = 1.00 adj = 1.00 (0 split)
- This result evaluates the surrogate splitter.
- The "agree" variable is a measure of similarity between the surrogate splitter and original splitter.
- Agree = 1 in this case suggests that the surrogate variable splits the data in exactly the same way as the original variable
- The "adj" variable refers to the decrease in impurity that could be achieved by using the surrogate as a splitter.
- A high value indicates a large decrease in impurity, and this indicates a useful split

2008 Q1C - GINI
- One way to evaluate splits in a classification tree is to use the concept of impurity
- A node which contains only one class is perfectly pure, while a node which contains an equal proportion of each class is least pure
- The goodness of a split is defined to be the decrease in impurity
- GINI is a common impurity function. It looks for the largest class in a dataset and strives to isolate it from all other classes.
- The GINI coefficient is the impurity of the node.

GINI index: $1 - \sum_i p_i^2$ Classification error: $1 - \max(p_i)$

2008 Q 1K

→ What is a profit matrix, how is it used in trees?

- A PM shows the profit that is gained or lost due to outcomes in the model
- It can be used for evaluating a classification tree or for classifying data
- For example, you can calculate the expected profit of "Event" and the expected profit of "non event"
- If for cate 1, then cate 1 is assigned as an event
- A PM is also used in the GINI splitting criterion to choose the best split which will minimise costs, at node level to calculate the misclassification rate, and at the level in the pruning process.

- Also used in splitting criterion GINI to choose best split → minimise cost.
- Allow pruning of the tree

When classifying cases, each one is assigned to the class that results in the highest expected profit

2011 Q 1F

→ How are xerror and xstd calculated?

- Xerror is the cross validation error
- Ideally we want to pick the tree with the lowest xerror ± 1 SD
- These are relative errors - multiply by root node to get misclassification rate
- Std is the standard deviation of the xerror

Xerror = relative error x root node error.

2012 Q 1A

→ Purpose of a tree in data mining context?

- Decision tree as a predictive model which maps observations about an item to conclusions about the item's target value
- Given a data set, aim is to predict the outcome of a new case, given the existing data
- Consists of a root node and various sub-level nodes. - Each node splits the data in two to determine the outcome of a class.
- Form of multivariate analysis
- Decision trees are produced by algorithms that identify various ways of splitting data into branch like segments

22/04/16 DA - TREE EXAM QUESTIONS

2017 Q1B (i)

i. Input Data.

- Can be continuous data - ie. numbers such as age, time or distance
- Ordinal data like School class or days of week - this is categorical data that can be ordered
- Binary variables - can be yes/no o/1 or male/female - special case of categorical data with only two outcomes.
- Categorical data - two or more levels/outcomes such as colour of prawns in ireland
- Negative values acceptable
- Missing data can be dealt with but better to have complete data.

2017 Q1B (ii)

ii. Selection of Split

- One way to evaluate split in a classification tree is to use the concept of impurity
- A node which contains only one class is perfectly pure while a node which has a 50/50 class split is "impure".
- The goodness of a split is defined to be the decrease in impurity.
- GINI or entropy are common impurity function.
- It tend at largest class in a dataset and serves to isolate it from other classes.
- The GINI coefficient is the impurity of the node.

- Assign an object at random from the node to class i with probability $p(i|t)$.
- The estimated probability that the belong to class j is $p(j|t)$
- Create estimated probability of misclassification under this rule

- Each split at any node gives us two children.
- We measure the impurity of the two children
- Choose split with the largest gain → R calls this improvement

2012 EXAM Q 1B (iii)

(iii) Selection of the tree: Size

- Two approaches
  - Bottom up: grow a big tree (maximal tree) and prune branches
  - Top down: Stop growing when there are no more useful splits

Cost complexity pruning: uses concept of misclassification

Grow maximal tree - largest possible

R_α = cost + complexity measure of the Tree T.

Define cost as the misclassification rate = $R(T)$

Complexity measure: function of # of terminal nodes

$R_\alpha = R(T) + \alpha |T|$ $|T|$ = # terminal nodes α penalty placed on complexity

For a single tree 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, $R(t)_\alpha$ 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 - adding to tree

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

Bigger values for α implies it is a better branch

Smaller values correspond to worse branches

- Calculate α for each node - prune the tree with the lowest value of α - weakest link
- Recalculate again, can continue pruning weakest node to get sequence of subtrees
- Set stopping size
- Use cross validation results
- Chose the tree with minimum misclassification

2012 Q 1B (iv)

iv Treatment of missing data.

- Trees can handle missing data via the use of surrogates
- A surrogate is a variable with possibly equivalent information as primary splitter
- A splitter that splits in a fashion similar to the primary splitter.

22/04/16 DA - TREE EXAM QUESTIONS.

- Recall structure of the MD of the course
- If the primary splitter is too expensive or difficult to gather → use surrogate instead
- Use surrogate split if data is missing
- Surrogates are node dependent calculated at the node level

2012 Q1D

→ Compare Trees with Logistic Regression

CART

- CART very bad at detecting linear structure, recognises it but cannot represent it effectively
- Can produce a very large tree in an attempt to represent a simple relationship
- Logistic regression good for linear relationship
- Many non linear structures can still be reasonably approximated with other structures
- Even incorrectly specified logistic regression can perform well
- LR provides a smooth continuous predicted probability of class membership.
- LR flexibility allowed with transformation of variables
- CART - automatic separation of relevant from irrelevant predictors
- Impervious to outliers
- Can handle outliers - CART
- Can handle missing data - CART
- CART requires only moderate supervision by
- CART can specify complexity of the model
- CART allows analysis on highly skewed messy data sets.
- CART can handle both categorical and quantitative data whereas LR can only handle binary categorical data
- CART automatically detects interactions in data sets. LR needs interactions to be manually defined
- CART more likely to overfit data, LR not.
- CART tends to work better with smaller data sets, LR tends to work better in linear domains (parametric v non parametric)
- CART relatively easier to explain

2013 Q1A

→ Missing Data

- May reduce number of cases or variables included in to data
- Could remove the variable or case which has missing info or else try and enter a new one

- look at % of each var that is missing and % of each case that is missing
- If one is missing, are you likely to miss another? look for patterns
- Visualise data and practice summary statistics.
- We can do nothing
- list wise deletion - only include if it has data for all cases
- Pair wise deletion - Data included if the case contains certain variables like AL, or DEG
- Omit variables with high % (10% +) of missing data
- Weighting for non response
- Imputation - Substitution of data
- Create new variable
- Model or unique in dealing with missing data

Imputation

- Method for substitution of missing data: regression mean
  - Copy from last
  - bootstrap
  - median/mode/mean, S.d. Shape of distribution
  - Distribution based replacement, calculated on percentile of variables distribution
  - Hot deck: divide sample into group and select value at random within group
  - Most frequently occurring variable

Disadvantage: Alters relationship between variables

- May increase biases in survey estimate
- Researcher may falsely treat data as a complex dataset
- Imputed values should be flagged

Multiple imputation - Create M datasets with imputed values (separate dataset)

- Combine results which should reflect uncertainty of missing data

24/04/16 DA - TREE EXAM QUESTIONS

Cross Validation Calculation
Divide the dataset into 5 groups $G_1, G_2, \ldots, G_5$ each of size s/n one for each group separately:
- Fit a full model on the dataset 'everyone except G_i' and determine $T_0, T_1, \ldots T_{am}$ for this reduced dataset
- Compute the predicted class for each observation in G_i, under each of the models T_0 for $1 \leq j \leq m$
- From this compute the risk for each subset

Sum over the G_i to get an estimate of risk for each β_j. For that β (complexity parameter) with smallest risk, compute T_0 for the full dataset, this is chosen as the best trimmed tree

Any value within one standard error of the achieved minimum cp is viewed as being the equivalent to the minimum (i.e. considered to be part of the flat plateau)

Importance of examining data prior to model building
- DA used to discover meaningful patterns and rules within a dataset.
- In order to identify any problems of features of the data, the first step should always be to explore the raw data graphically and produce a range of descriptive statistics
- By doing this prior to building any model, outliers and missing data can be identified and handled appropriately, thus increasing the accuracy of the data model
- Derived variables can also be created, bringing features of the data to light

- Outliers can have many problematic effects on a data model, including bias, distortion and faulty conclusion.
- Although there are a number of ways to handle outliers, it must first be decided what constitutes an outlier.
- This is largely a subjective task and method include visually identifying outliers; assuming a standard deviation to draw a definite line; or using method such as Grubbs test (using z-values).

Once an outlier is identified, the next step is to decide what to do with it. If it is a data entry error, it should be fixed; if it is due to experimental problem it should be removed or altered (by replacement or transformation); if it is as a result of bad luck, it should be left alone. If it is not clear what caused by a decision must be made as to whether it was by chance or by mistake. Outliers occurring by chance should be left untouched; mistake should be removed or altered

Missing data can also introduce bias to a model and lead to inaccurate conclusions There are numerous ways to handle. In order to choose the right method it is important to examine the data first. Listwise deletion removes entire cases with any missing data, Imputation methods substitute new values for any missing values; variables or data with a high % of missing data can be removed or weighting can be applied to non-response.

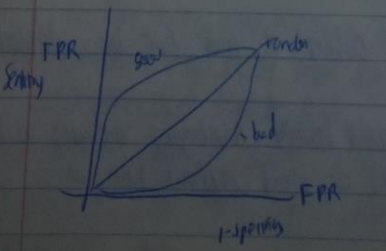
Raw data almost always needs to be cleaned before it can be used to build any sort of data model. Outliers and missing data can have a strong effect on a model and so, they must be dealt with appropriately.

ROC

A Receiver Operating Characteristic Curve is a plot of the true positive rate versus the false positive rate for different cut offs. It is used for model evaluation

actual \ predicted	True	False
True	TP	FN
False	FP	TN

An ROC curve plots the true positive rate (Sensitivity = $\frac{TP}{TP+FN}$) against the false positive rate (1-specificity (1- $\frac{TN}{TN+FP}$))



24/04/16 DA 11

ROC heaven is a line with points (0,0), (0,1), and (1,1) A random model will produce a straight line - this is because of random guessing
Apart from visually determining which is the better of the two models, we can calculate the area under the curve of each model (can be calculated). The better model will give a larger area

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

How to construct
- Need to calculate TPR and FPR (Sensitivity & 1-Specificity)
- Draw up a confusion table using a cut off of x
- Using the confusion table, derive Sensitivity = TP / (TP+FN) and Specificity = TN / (TN+FP)
- TPR = Sensitivity FPR = 1 - Specificity
- Repeat for multiple values of x and draw curve.

Shows
- Trade off Sensitivity and Specificity (Sensitivity increases, Specificity decreases).
- Closer the curve follows the left and upper border, means a more accurate model
- Area under curve is measure of test accuracy

- In certain scenarios we may wish to add an additional cut or weight to a model/diagnosis which could affect our decision
- For example, falsely misclassifying a component as damaged could cost the company more to replace it than it would to falsely misclassify it as undamaged
In ROC cut: $P(+) * [1-PP] * C(+|-) + P(-) * FP * C(-|+)$

CART vs LOGISTIC REGRESSION
- CART analysis allows you to perform analysis on a highly skewed noisy dataset. LR can be affected by noisy data and outliers
- CART can handle both categorical and quantitative data whereas LR can only handle binary categorical data

- Much easier to build a good model using CART, as a tree like structure automatically disregards insignificant variables. LR needs to be manually adjusted with different combinations of variables
- CART will automatically detect interactions in the dataset. LR needs interaction added in manually
- CART more likely to overfit the data and must be pruned manually to prevent this. LR less likely to overfit data
- CART tends to work better with real data sets, whereas LR tends to work better globally (parametric vs non-parametric)
- Easier to explain results of CART to non statisticians
- LR better at detecting linear relationships
- Both models can provide probabilities of class assignment
- Both require a target variable to be specified
- Both are supervised learning techniques.

Combined?
- Running a shallow tree
- Assigned each case to a terminal node.
- Treat terminal node as categorical variable
- Feed back into logistic regression model.

Choosing Best Tree
- Look at the CP plot for complexity trade off.
- Ensure tree is complex and deep enough to explain space of the data
- A tree too deep will overfit the data
- Look at misclassification rates and cross validation
- Look at the ROC curve

Surrogates
- A variable with possibly equivalent info. A surrogate is a splitter that splits in a fashion similar to a primary splitter.
- Used if data is missing and can act as the primary splitter.
- Surrogates also help reveal common patterns among predictors and clusters of.
- For example, a primary splitter may be present in the training data but often

making predictions on future data we have no way of knowing if that splitter will be available
- if a primary splitter is missing, the surrogate will be available to act as a primary splitter.

Creating Splits for DT be
- look at impurity at parent node
- Measure impurity of the two children
- Evaluate each case from the parent node and calculate impurity with an impurity function like 'gini'
- Choose the split which minimises impurity i.e. provides the most improvement.
- Split can be created with multiple impurity functions like Gini and Entropy

Can look at:
- Misclassification rate
- Does it make common sense?
- Look at t-test and chi-square test to check for differences.
- Look at impurity level

Importance of cleaning your data
- Identify inconsistencies in the data which may affect results
- Identify outliers early on and determine how to handle
- Ensure data is coded correctly
- To remove any unnecessary variables from the dataset and determine if they may impact the model.
- To check the missing data

Use of background knowledge
- Better understand the data and results from the model
- Use common sense when looking at data.
- Can check for irregularities that others may not notice
- Understand what the variables represent and tell if they have a relationship in reality
- Manually enter intervals to layer regions
- Determine acceptable cut off point and multicollinearity

Pruning A large Tree
- Normally always need to prune a tree
- Reduce size of decision tree by removing sections that provide little power to classify instances
- Goal: Reduce Complexity, improve predictive accuracy
- When growing a tree to its maximum depth, it tends to overfit data
- If a tree overfits the data will not perform as well with new data
- Need to reach a trade off between model complexity and accuracy
- looking at plot of CP should this
- A more complex model will take longer to compute

Determination of Splits in a tree
- Splits in tree are determined by the level of impurity in the nodes
- look at impurity of parent node
- Measure impurity of child nodes with impurity function (gini)
- Evaluate each rule from the parent node
- Choose the split which minimises the impurity level in R (this is known as the "most improvement"

Selection of the tree - In particular node size
2 approaches: Bottom up (Classical CART - Grows maximal tree then prunes branches)
Top Down - Stop growing when there are no more useful splits
- Classical CART - need to look at the misclassification rate of entire tree and of nodes
Cost complexity pruning:
- Calculate a value for $R(T) = R(T_t) / |\tilde{T}_t| - 1$
- Prune the node or branch with the lowest value of (weighted loss)
- Recalculate again
- Continue pruning weaker nodes to get a sequence of subtrees
- Choose tree with minimum misclassification or close to the minimum.
- Plot misclassification rate v # of nodes

24/04/16 DA

CART v LOGISTIC REGRESSION

1. Detection
- CART excels in the detection of local structure. Each half of the tree is analysed separately
- Effective capture of global features of data
- CART very bad at detecting linear structure and cannot represent it effectively
- LR - good for linear relationships. Many non-linear structures can still be reasonably approximated with a linear structure

2. Assumptions
- Decision tree assumes the split are axis parallel and will become more complex with the increase in number of features and multiple decision boundaries are possible
- On the other hand, LR assumes there is only one decision boundary that is smooth and non-linear

3. Overfitting
- Complex decision trees may overfit the data and tree will become unstable. Can prune tree to solve this.
- LR is intrinsically simple, it has low variance and so is less prone to overfitting

4. Speed
- While both algorithms are fast, LR has a faster run time on large datasets than CART

5. Interpretability
- Nice probabilistic interpretation
- Easy to visualise and interpret

6. Transformation
- CART does not require any transformations such as log or square root

26/04/16 DA EXAM PAPERS

Selection of Split

- One way to evaluate splits in a classification tree is the concept of impurity
- A node which contains only one class is perfectly pure while a node which has a 50/50 class split is "impure"
- The goodness of fit is defined to be the decrease in impurity
- GINI for $c=2$ $= 2\,p(1|t)\,p(2|t)$ or entropy generally used
- It looks at largest class in a dataset and serves to isolate it from other classes

- Look at impurity of parent node
- Each split gives 2 children
- Measure impurity of the two children
- Choose the split with the largest decrease in impurity - R calls this improvement

Cost Complexity Pruning

- Grow maximal tree - largest tree possible
- R_α = cost + complexity measure of tree
- Define cost of misclassification rate = $R(T)$
- Complexity measure: function of # of terminal nodes
- $R_\alpha = R(T) + \alpha * |T|$ 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, $R(t)_\alpha$ and $R(T_t)_\alpha$ increase but $R(T_t)_\alpha$ increases quicker
  The value of α when $R(t)_\alpha = R(T_t)_\alpha$ = price paid for complexity - called α_{eff}

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

Bigger value for α implies it is a better branch

ROC and Priors

TPR and TNR not dependent on prior probabilities but accuracy is

$$Acc = P(+) * TPR + P(-) * (1 - FPR)$$

$P(+)$ = proportion of + in population written as π in R

$p(-)$ = proportion of $(-)$ in population
$p(+) + p(-) = 1$
Can draw lines of iso accuracy on the ROC curve

$$TPR = \frac{Acc - p(-)}{p(+)} + \frac{p(-)}{p(+)} * FPR$$

Plot error vs $p(+)$
$Error = 1 - Acc = 1 - (p(+) * TPR + p(-) * (1 - FPR))$
$Err = p(+) FPR + p(-) FPR$
$= (FN - FP) * p(+) + FP$
New space y-axis: error x-axis: $p(+)$

- Points in ROC space are going to be mapped onto lines in "error/$p(+)$" space
- Allows us to see clearly how error rate varies according to $p(+)$