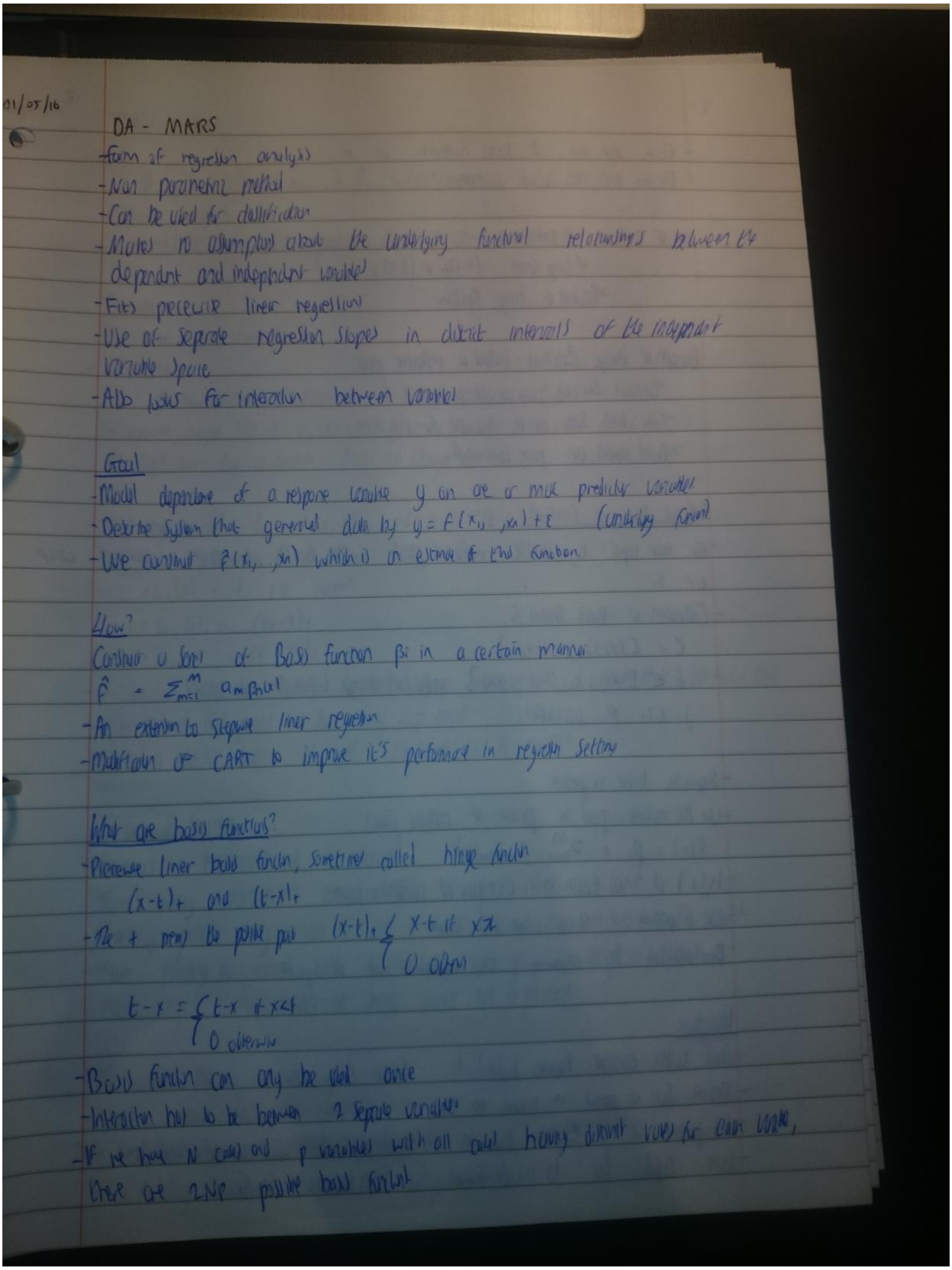


01/05/16

DA - MARS

- Form of regression analysis
- Non parametric method
- Can be used for classification
- Makes no assumptions about the underlying functional relationships between the dependent and independent variables
- Fits piecewise linear regressions
- Use of separate regression slopes in distinct intervals of the independent variable space
- Also allows for interaction between variables

Goal

- Model dependence of a response variable y on one or more predictor variables
- Describe system that generates data by $y = f(x_1, \ldots, x_n) + \varepsilon$ (underlying function)
- We construct $\hat{f}(x_1, \ldots, x_n)$ which is an estimate of this function.

How?

Construct a set of basis function β_i in a certain manner

$\hat{f} = \Sigma_{m=1}^{M} a_m \beta_m(x)$

- An extension to stepwise linear regression
- Modification of CART to improve it's performance in regression setting

What are basis functions?

- Piecewise linear basis function, sometimes called hinge function

$(x-t)_+$ and $(t-x)_+$

- The + means the positive part $(x-t)_+ \begin{cases} x-t & \text{if } x \geq t \\ 0 & \text{otherwise} \end{cases}$

$t-x = \begin{cases} t-x & \text{if } x<t \\ 0 & \text{otherwise} \end{cases}$

- Basis function can only be used once
- Interaction has to be between 2 separate variables
- If we have N cases and p variables with all cases having distinct values for each variable, there are $2Np$ possible basis functions

- Each knot has 2 basis functions < >
- Doesn't take into account interaction

One of 3 forms • A constant
• A hinge function $(x-t)_+$ or $(t-x)_+$
• Product of hinge function

Graph of hinge functions - called a reflective pair
• Continues them as pairs at the beginning
• Takes both sides of a variable for the slope
• Build model then prune backwards

Reflected Pair of Functions

- For each input X_i, calculate reflected pairs with knots at each individual value of X_i
- Collection of Basis functions C
  $C = [(x_j - t)_+, (t - x_j)_+]$
  $t \in [x_{1j}, x_{2j}, \ldots, x_{nj}]$ individual unique values for each variable
  $j = 1, 2, \ldots p$ variables

- Stepwise linear regression
- Use the reflected pairs or product of reflected pairs
  $F(x) = \beta_0 + \Sigma_{m=1}^{M} \beta_m h_m(x)$
- $h_m(x)$ is a function in C - pair of reflected functions
- Or a product of 2 or more such functions
- βm estimated by minimising the residual sum of squares

Procedure

- Start with constant function $h_0(x) = 1$ (like fitting mean of y to data)
- Define set of terms in model as Φ
- All products of a function h_m in the model set Φ with one of reflected pairs in C
- We add the to the model Φ a term of the form:

01/05/16

$$\hat{\beta}_{M+1} h_l(x) \cdot (X_j - t)_+ + \hat{\beta}_{M+2} h_l(x) \cdot (t - X_j)_+ \qquad h_l \in \phi$$

- Calculate weights.
- We add the term that produces the largest decrease in training error - SSE

Example

- First term is a constant h_0
- Second term:
  - → now consider a function in the form of $\beta_1 (X_j - t)_+ + \beta_2 (t - X_j)_+$
- For example, suppose this is $\beta_1 (X_7 - t)_+ + \beta_2 (t - X_7)_+$ for some value of t
- Multiplication by a constant does not change things

- Next stage we add $h_m(x) \cdot (X_j - t)_+$ and $h_m(x) \cdot (t - X_j)_+$
- We have 3 choices for h_m:
  - $h_0(x) = 1$ i.e. constant.
  - $h_1(x) = (X_7 - t)_+$
  - $h_2(x) = (t - X_7)_+$
- h_1, h_2 are interaction terms, can only have interaction with a variable that is already in
- At the end of this process we have a very large model
- We set a number limit of # of terms in the model

Hierarchical Structure

- There is a hierarchical structure.
- Two way interactions are included if main effects are there
- A 4-way product only included if one of it's 3 way components is in the model.
- Have facility to restrict order of interaction and particular terms in interaction
- each input can appear at most, once in a product

Forward Pass

Adds terms in pairs until:

- Reach maximum number of terms
- Adding a term changes R^2 by less than 0.001

- Reached an R^2 of 0.999 or more
- GRSq < -10
- No new term increases R^2
- Default for n_k is $\min(200, \max(20, 2*ncol(x))) + 1$

- GRSq - EARTH's estimate of the generalisation performance of the model

Backward Pass

- Need to prune the number of basis functions
- Can specify maximum number of terms here - nprune
- Assume we have n_b basis functions
  For each subset 1. n_b find best subset in terms of lowest RSS
- Then look at each subset and calculate GCV to find lowest value
- Gives us a series of basis functions to use
- Calculate the coefficients, residuals and fitted values using lm (no standard errors)

Comments

- Typically overfits
- Go backwards and reduce and remove terms which cause smallest decrease in RSS
- Estimate model for each size λ - # terms in model
- Use cross validation to estimate λ

GCV formula: $$GCV(\lambda) = \frac{\sum_{i=1}^{N} (y_i - \hat{f}_\lambda(x_i))^2}{(1-M(\lambda)/N)^2}$$

- m(λ) is the effective # of parameters in model
- r linearly independent basis functions
- k knots
- $M(\lambda) = r + 3K$
- Substitute for cross validation

GRSq

$$GRSq = 1 - \frac{GCV}{GCV_{null}}$$

- Where GCV null is the ~~mean~~ GCV of an intercept only model

01/05/16

DA-MARS

- The GCV and GRSq are measures of the generalisation ability of the model, ie how well the model would predict using data not in the training set.
- like adjusted R^2

OUTPUT PLOTS

1- Model Selection

- Plot of R^2 and GRSq
- R^2 a normalised form of the RSS
- GRSq - Measure of generalisation ability of model
- Should not run together - there should be an increased penalty being applied to the GCV as # of model parameters increase
- GCV - Generalised cross validation : trade off goodness of fit against model complexity

2-Residuals vs Fitted

- Shows residuals of each value of predicted response : (remainder of observed - expected value).
- Should be as close to zero as possible
- Constant variance of residuals not as important as in linear model
- Highlights outliers

3.-Cumulative Distribution

- Cumulative sum of absolute value of residuals
- Ideally, start at 0 and shoot straight up to one
- Can calculate mean etc.

4 - Quantile Quantile

- Compares the distribution of residuals to the assumed normal distribution
- Want all points to fall on a straight line
- Look for outliers

Variable Importance

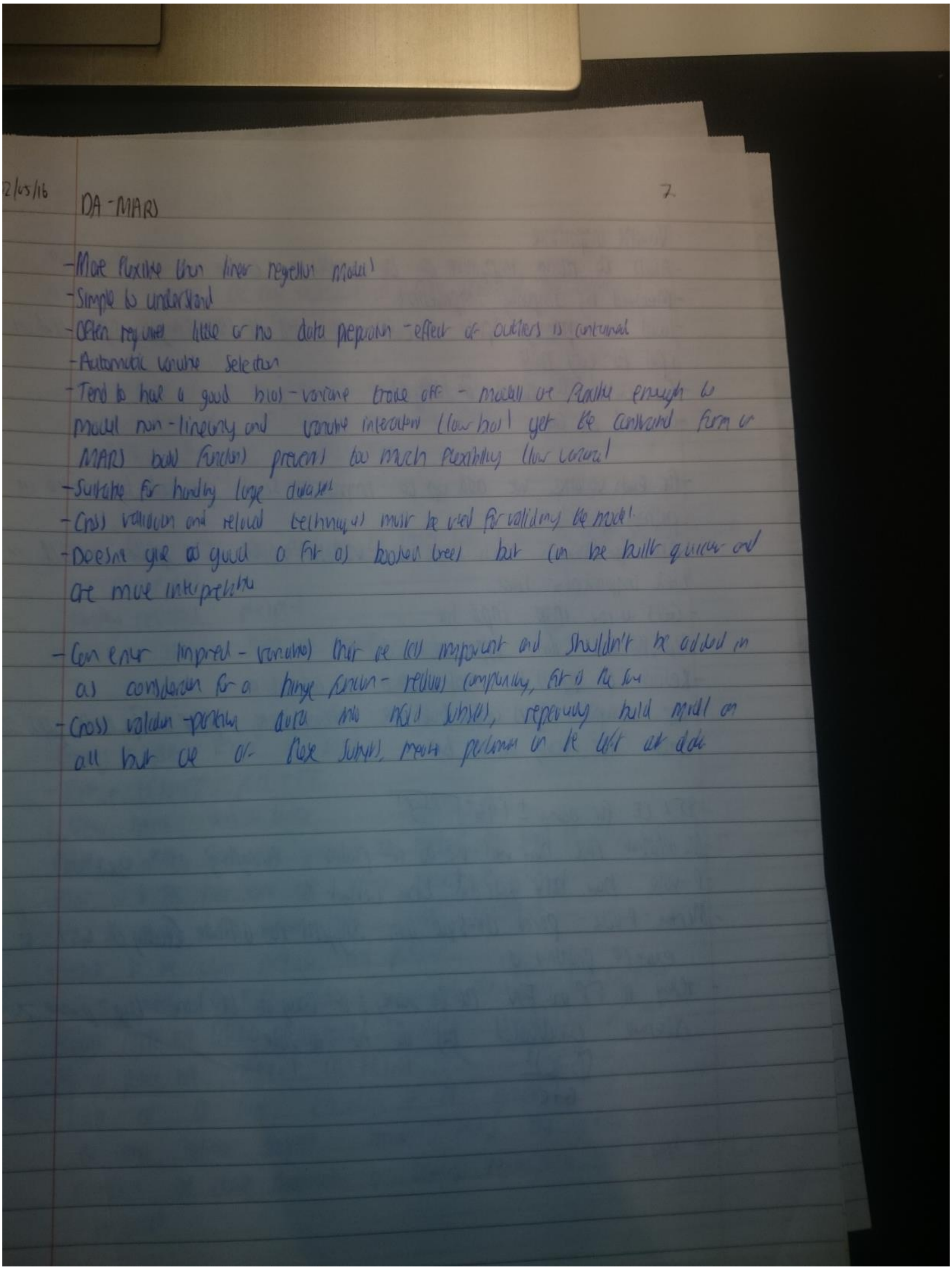
- Loose relationship between var importance in model computed
- Variance of variable importance is high
- Run different datasets ⇒ bootstrap ⇒ may get different order
- Highly correlated variables - one variable is chosen over the other
- In interactions, each variable gets credit for entire term
- Uses 3 different criteria:
  - # times variable appears in subsets ~~relative to previous subsets~~ of various sizes
  - Decrease in RSS for each subset relative to previous subset
  - Decrease in GCV for each " "
- Normalised so largest decrease is 100

ie variable with "highest importance" will have large number for nsubsets and numbers close to 100 for GCV and RSS

Comments

- Can calculate CI and PI
- Can develop a variance model
- Assumes errors are independent
- Ability to operate locally
- Regression surface built up parsimoniously using non-zero components only where they are needed
- Can use GLM method after the basis matrix has been ~~created~~ constructed to determine final weights

- Can be computationally slow to fit the model and far more complex than a tree
- Has advantage of being able to be used as binary or continuous outcome
- Non parametric regression procedure - makes no assumption about underlying functional relationship between dependent and independent variables
- Looks for and identifies interactions
- Typically accurate.
- Automatically models non-linearities and interactions



2/05/16

DA - MARS

- More flexible than linear regression models
- Simple to understand
- Often requires little or no data preparation - effect of outliers is contained
- Automatic variable selection
- Tend to have a good bias - variance trade off - models are flexible enough to model non-linearity and variable interactions (low bias) yet the constrained form of MARS basis functions prevents too much flexibility (low variance)
- Suitable for handling large datasets
- Cross validation and related techniques must be used for validating the model.
- Doesn't give as good a fit as boosted trees but can be built quicker and are more interpretable

- Can enter linear - variables that are less important and shouldn't be added in as consideration for a hinge function - reduces complexity, fit is the same
- Cross validation - partition data into n(k) subsets, repeatedly build model on all but one of these subsets, measure performance on the left out data

Variable importance

- Assess the relative importance of the variables in a tree
- Measured by impurity improvement
- looks at primary splitter for each node and all the surrogate splits listed in report for every node
- Can control the number of surrogates
- Calculated over tree

- For each variable we add up the improvement scores generated by variable in primary splitters.
- Also go through each node this variable used as a splitter and add in these improvement scores
- Gets a raw ~~imp~~ input score
- Variables that don't appear get a zero
- Rescale the results so best variable is 100 and rest [illegible]
- Top competitor gets credit but the second best splitter in a node gets zero credit for being second best

- 95% CI for accuracy $\pm 1.96\sqrt{\frac{p(1-p)}{N}}$
- No information rate: how well we do if predict everything as 'yes'
- P-value: how better you did than without it
- McNemar-P-value - paired chi-square test, suggests no difference (fluky or 0.5) in example probably of i.
- Testing if FP and FN are the same, if they are the same high p-value [illegible] otherwise concluded they are not the same

$$\frac{(B-C)^2}{B+C}$$

16/04/16 DA - Multivariate Adaptive Regression Splines MARS

What are MARS?

- Form of regression analysis
- Non parametric method
- Can be used for classification
- Makes no assumptions about the underlying functional relationships between the dependant and independent variables
- Fits peicewise linear regressions
- Use of seperate regression slopes in distinct intervals of the independent variable space
- Also looks for interaction between variables
- Divides space - "very stepwise regressiony"

Goal

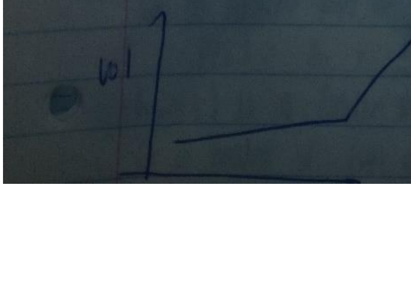
- Model the dependance of a response variable y on one or more predictor variables.
- Set of data $y_i, x_{1i}, \ldots, x_{ni}$
- Describe system that generated data by $y = f(x_1, \ldots, x_n) + \varepsilon$ (underlying function)
- We construct $\hat{f}(x_1, \ldots, x_n)$ which is an estimate of this function

How?

- Construct a series of basis functions β_i in a certain manner
- $\hat{f} = \Sigma_{m=1}^{M} a_m \beta_m(x)$
- An extension to stepwise linear regression
- Modification of CART to improve its performance in regression setting

Cherry Tree Example

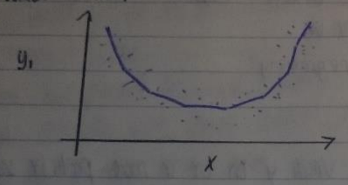
- Measurement of girth, height, volume of timber of 31 trees
- Want to predict volumn based on girth of tree
- Can fit a SLR of girth to volumn



- Given value for intercept, h(13.8 - Girth) value for when $G < 13.8$ and $h(G - 13.8)$ when $G > 13.8$
- GCV - like cross validation
- RSq - like R^2
- GRSq

- Result is value of h() function as the basis values.

Quadratic example



$y_1 \sim x$

Returns 8 parameters → intercept and 7 h() function

- Can also use interactions in model to predict.

Form of Models

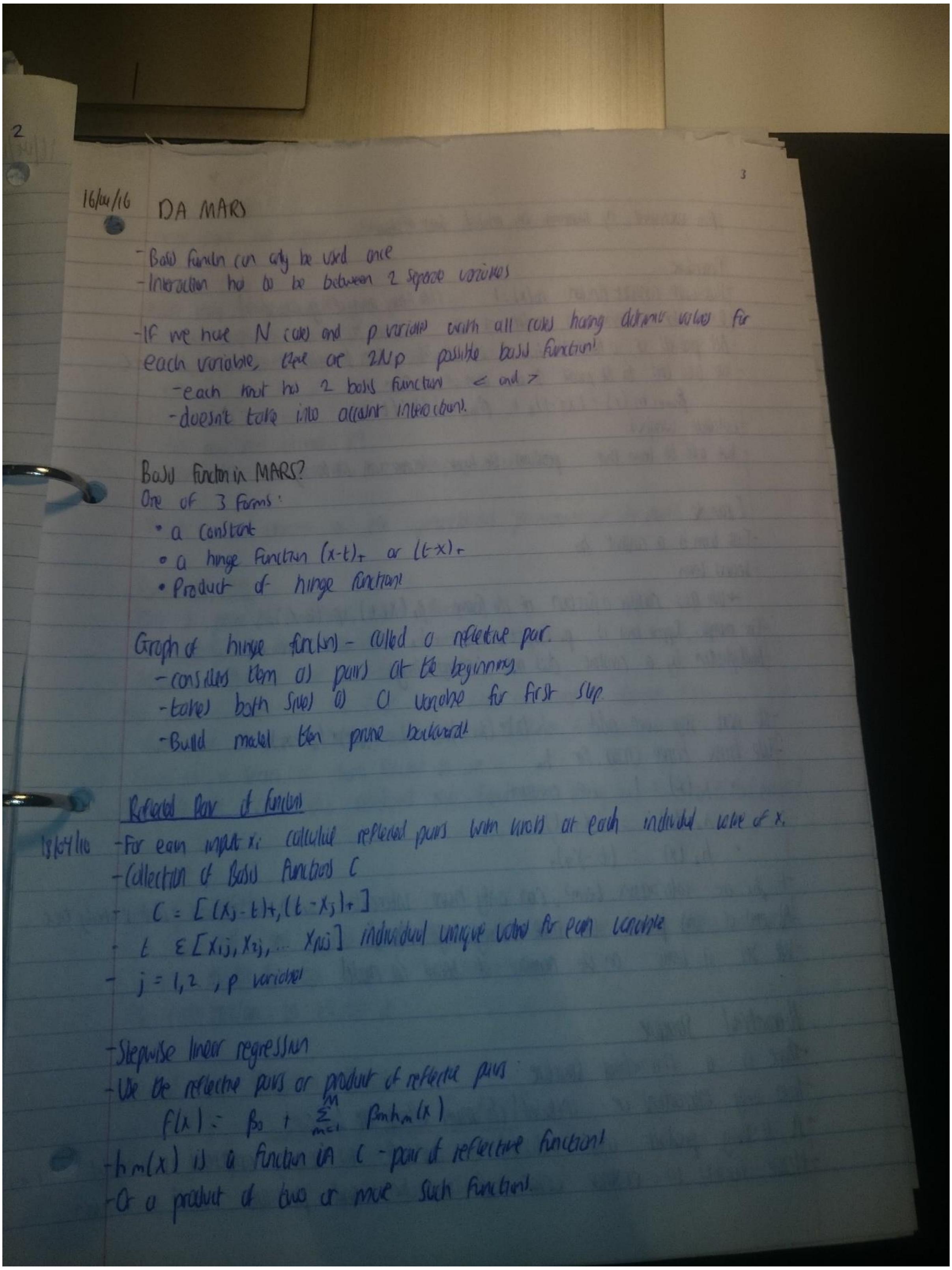
- MARS builds model in form of $\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$
- c_i are coefficients
- $B_i(x)$ are called basis function

What are basis function?

- Piecewise linear basis functions sometimes called hinge function
  $(x-t)_+$ and $(t-x)_+$
- The + means the positive part: $(x-t)_+ = \begin{cases} x-t & \text{if } x > t, \\ 0 & \text{otherwise} \end{cases}$

$(t-x) = \begin{cases} t-x & \text{if } x < t \\ 0 & \text{otherwise} \end{cases}$

- Sometimes called hinge function



16/04/16 DA MARS

- Basis function can only be used once
- Interaction has to be between 2 separate variables

- If we have N cases and p variables with all cases having distinct values for each variable, there are 2Np possible basis functions
  - each knot has 2 basis function < and >
  - doesn't take into account interactions.

Basis function in MARS?

One of 3 Forms:

- a constant
- a hinge function $(x-t)_+$ or $(t-x)_+$
- Product of hinge functions

Graph of hinge functions - called a reflective pair.

- considers them as pairs at the beginning
- takes both sides of a variable for first step
- Build model then prune backwards

Reflected Pair of functions

18/04/16 - For each input x_i calculate reflected pairs with knots at each individual value of x.

- Collection of Basis Functions C
- $C = [(x_j - t)+, (t - x_j)+]$
- $t \in [x_{1j}, x_{2j}, \dots x_{Nj}]$ individual unique value for each variable
- $j = 1, 2, p$ variables

- Stepwise linear regression
- Use the reflective pairs or product of reflective pairs

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x)$$

- $h_m(x)$ is a function in C - pair of reflective functions
- Or a product of two or more such functions.

β_m estimated by minimising the residual sum-of-squares

Procedure

- Start with constant function $h_0(x) = 1$ (like fitting mean of y to data)
- Define set of terms in model as ϕ
- All products of a function h_m in the model set ϕ with one of reflected pairs in C
- We add this to the model ϕ a term of the form

$$\beta_{m+1} h_1(x) \cdot (x_j - t)_+ + \beta_{m+1} h_1(x) \cdot (t - x_j)_+ \qquad h_1 \Sigma \phi$$

- Calculate weights
- We add the term that produces the largest decrease in the training error - SSE

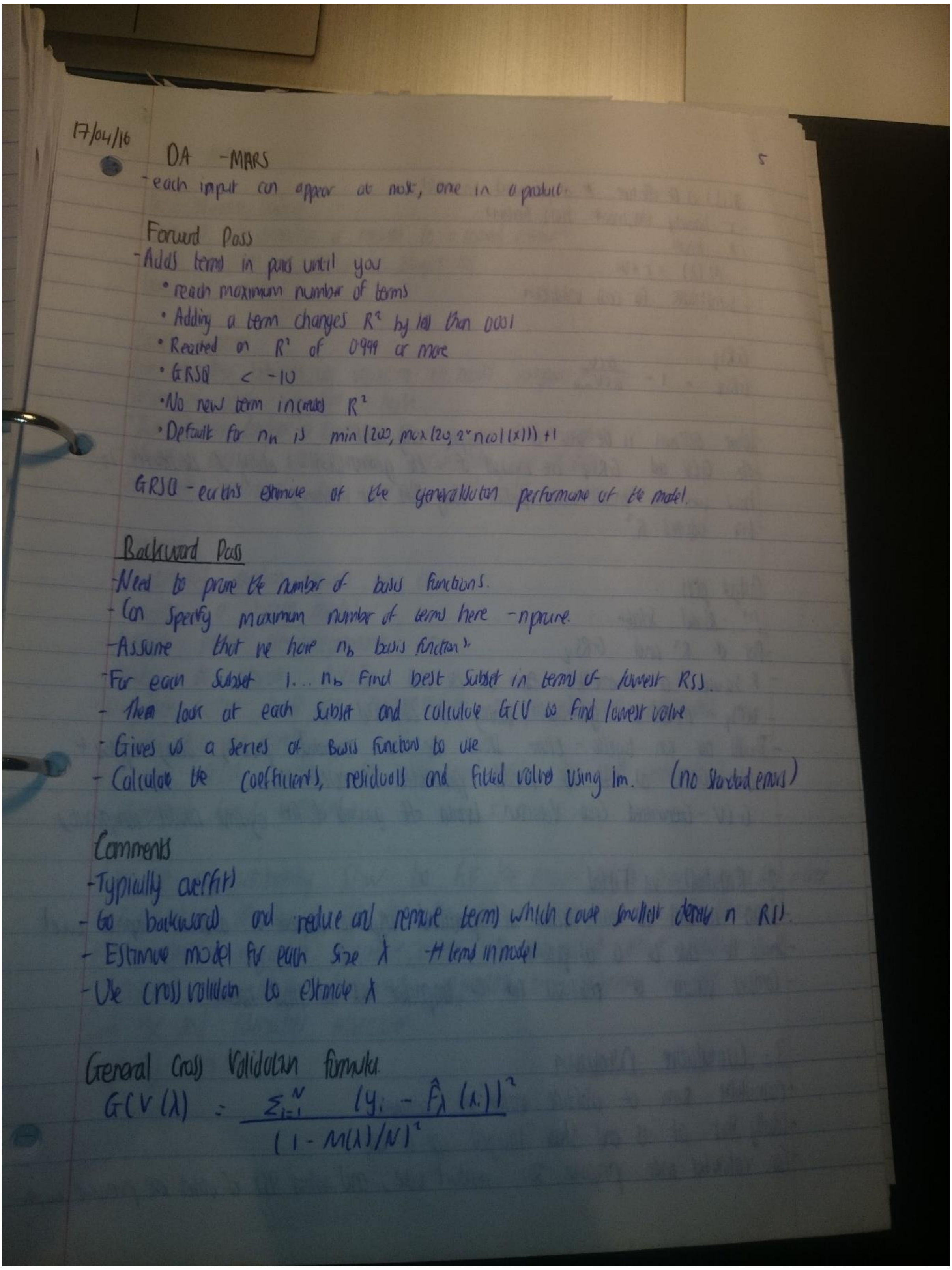
Example

- First term is a constant h_0
- Second term:
  - → We now consider a function of the form $\beta_1 (x_j - t)_+ + \beta_2 (t - x_j)_+$
- For example, suppose this is $\beta_1 (x_7 - t)_+ + \beta_2 (t - x_7)_+$ for some value of t.
- Multiplication by a constant does not change things

- The next stage we add $h_m(x) \cdot (x_j - t)_+$ and $h_m(x) \cdot (t - x_j)_+$
- We have three choices for h_m:
  - $h_0(x) = 1$ i.e. constant
  - $h_1(x) = (x_7 - t)_+$
  - $h_2(x) = (t - x_7)_+$
- h_1, h_2 are interaction terms, can only have interaction with a variable that is already there
- At end of this process we have a very large model
- We set a limit on the number of terms in model

Hierarchical Structure

- There is a hierarchical structure
- Two way interactions are included if main effects are there
- A 4-way product only included if one of its 3-way components is in model already
- Have facility to restrict order of interaction and particular terms in interaction



17/04/16

DA -MARS

5

- each input can appear at most, once in a product.

Forward Pass

- Adds terms in pairs until you
  - reach maximum number of terms
  - Adding a term changes R^2 by less than 0.001
  - Reached an R^2 of 0.999 or more
  - GRSQ < -10
  - No new term increases R^2
  - Default for n_k is min(200, max(20, 2*ncol(x))) +1

GRSQ - earth's estimate of the generalisation performance of the model.

Backward Pass

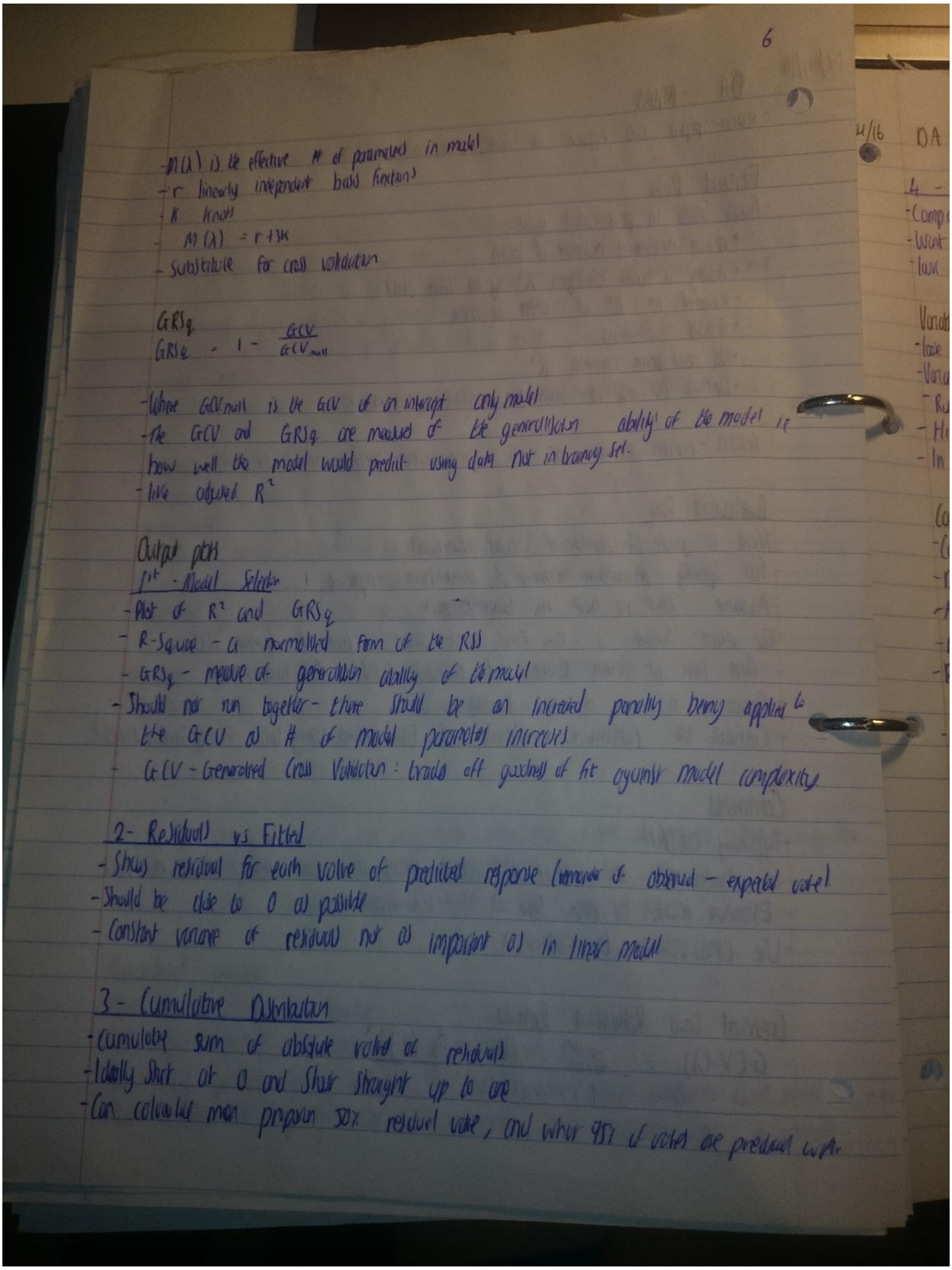
- Need to prune the number of basis functions.
- Can specify maximum number of terms here - nprune.
- Assume that we have n_b basis functions.
- For each subset 1... n_b find best subset in terms of lowest RSS.
- Then look at each subset and calculate GCV to find lowest value
- Gives us a series of basis functions to use
- Calculate the coefficients, residuals and fitted values using lm. (no standard errors)

Comments

- Typically overfits
- Go backwards and reduce and remove terms which cause smallest decay in RSS.
- Estimate model for each size λ - # terms in model
- Use cross validation to estimate λ

General Cross Validation formula

$$GCV(\lambda) = \frac{\sum_{i=1}^{N} (y_i - \hat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2}$$



- $M(\lambda)$ is the effective # of parameters in model
- r linearly independent basis functions
- K knots
- $M(\lambda) = r + 3K$
- Substitute for cross validation

GRSq

$$GRSq = 1 - \frac{GCV}{GCV_{null}}$$

- Where GCVnull is the GCV of an intercept only model
- The GCV and GRSq are measures of the generalisation ability of the model ie how well the model would predict using data not in training set.
- like adjusted R^2

Output plots

1st - Model Selection

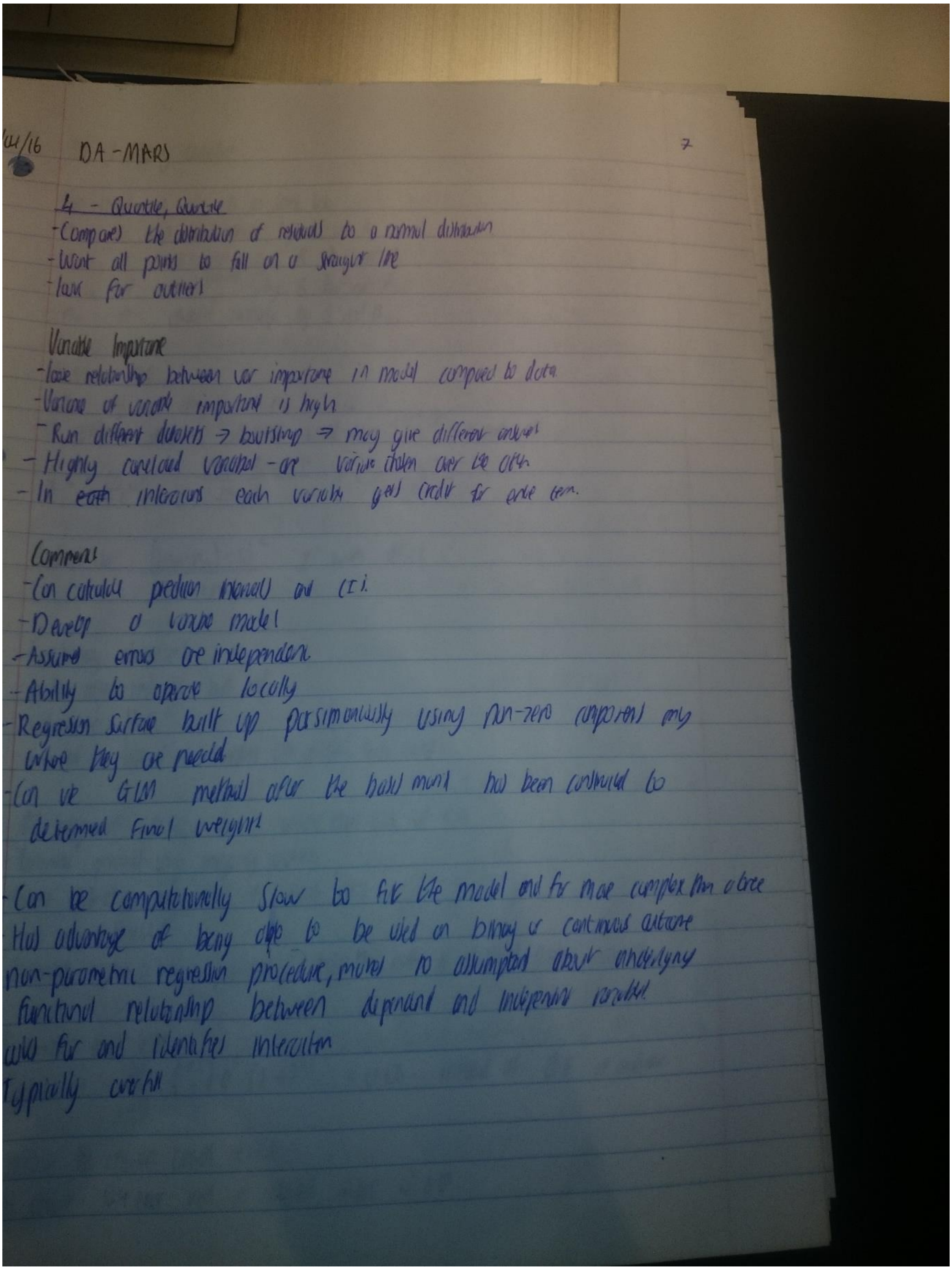
- Plot of R^2 and GRS_q
- R-Square - a normalised form of the RSS
- GRS_q - measure of generalisation ability of the model
- Should not run together - there should be an increased penalty being applied to the GCV as # of model parameters increases
- GCV - Generalised Cross Validation: trades off goodness of fit against model complexity

2- Residuals vs Fitted

- Shows residual for each value of predicted response (remainder of observed - expected value)
- Should be close to 0 as possible
- Constant variance of residuals not as important as in linear model

3- Cumulative Distribution

- Cumulative sum of absolute values of residuals
- Ideally start at 0 and shoot straight up to one
- Can calculate mean proportion 50% residual value, and where 95% of values are predicted with



/11/16 DA-MARS

4 - Quantile, Quantile

- Compares the distribution of residuals to a normal distribution
- Want all points to fall on a straight line
- look for outliers

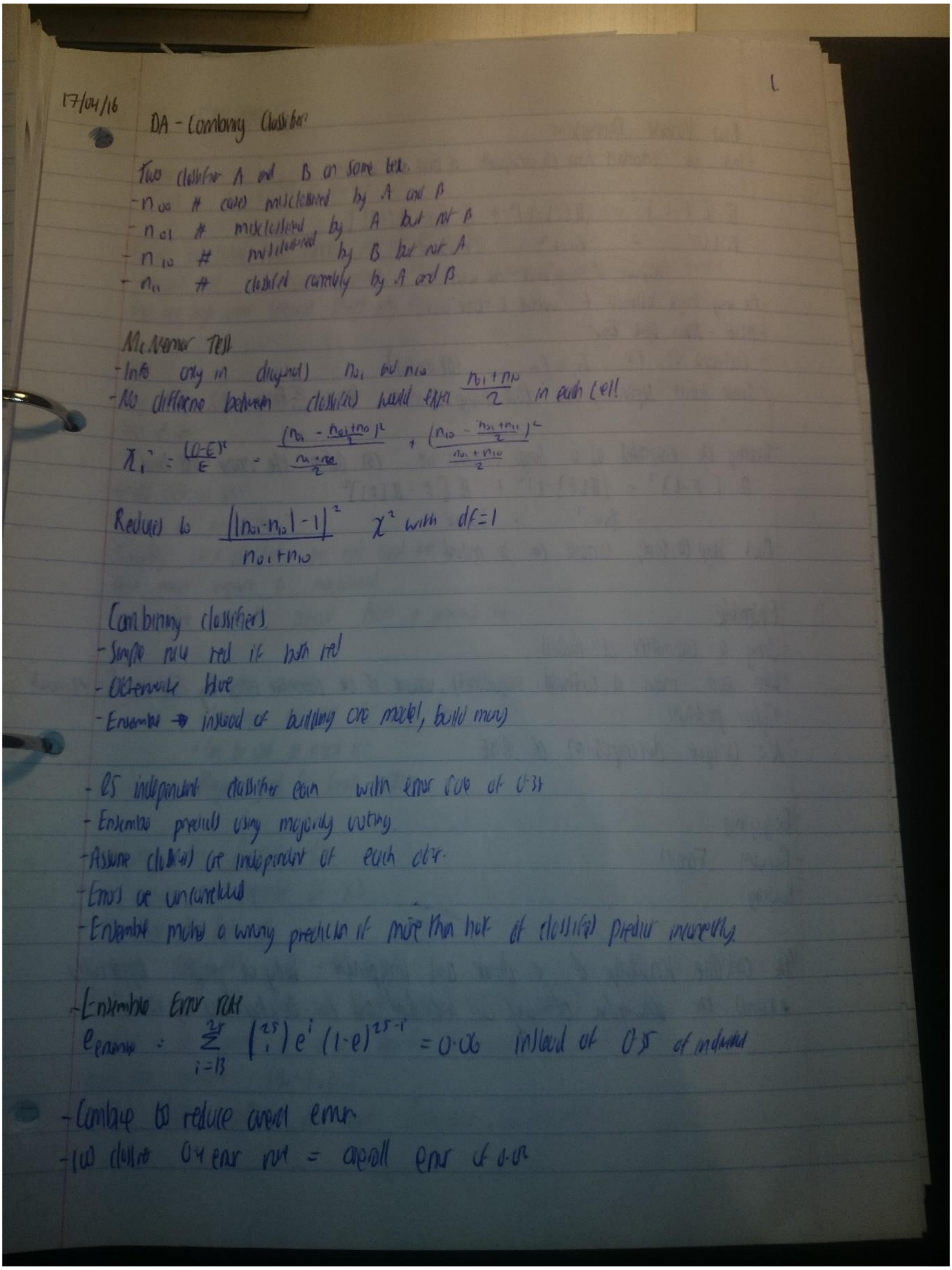
Variable Importance

- loose relationship between var importance in model compared to data
- Variance of variable importance is high
- Run different datasets ⇒ bootstrap ⇒ may give different answers
- Highly correlated variables - one variable chosen over the other
- In ~~each~~ interactions each variable gets credit for entire term.

Comments

- Can calculate prediction intervals and CI.
- Develop a viable model
- Assumes errors are independent
- Ability to operate locally
- Regression surface built up parsimoniously using non-zero components only where they are needed
- Can use GLM method after the basis model has been constructed to determined final weights

- Can be computationally slow to fit the model and far more complex than a tree
- Has advantage of being able to be used on binary or continuous outcome
- non-parametric regression procedure, makes no assumption about underlying functional relationship between dependent and independent variables
- allows for and identifies interaction
- Typically overfit



17/04/16

DA - Combining Classifiers

Two classifier A and B on same test.

- n_{00} # cases misclassified by A and B
- n_{01} # misclassified by A but not B
- n_{10} # misclassified by B but not A
- n_{11} # classified correctly by A and B

McNemar Test

- Info only in diagonals n_{01} and n_{10}
- No difference between classifiers would expect $\frac{n_{01}+n_{10}}{2}$ in each cell!

$$\chi_1^2 = \frac{(O-E)^2}{E} = \frac{\left(n_{01} - \frac{n_{01}+n_{10}}{2}\right)^2}{\frac{n_{01}+n_{10}}{2}} + \frac{\left(n_{10} - \frac{n_{01}+n_{10}}{2}\right)^2}{\frac{n_{01}+n_{10}}{2}}$$

Reduces to $\frac{\left(|n_{01}-n_{10}|-1\right)^2}{n_{01}+n_{10}}$ χ^2 with $df=1$

Combining classifiers

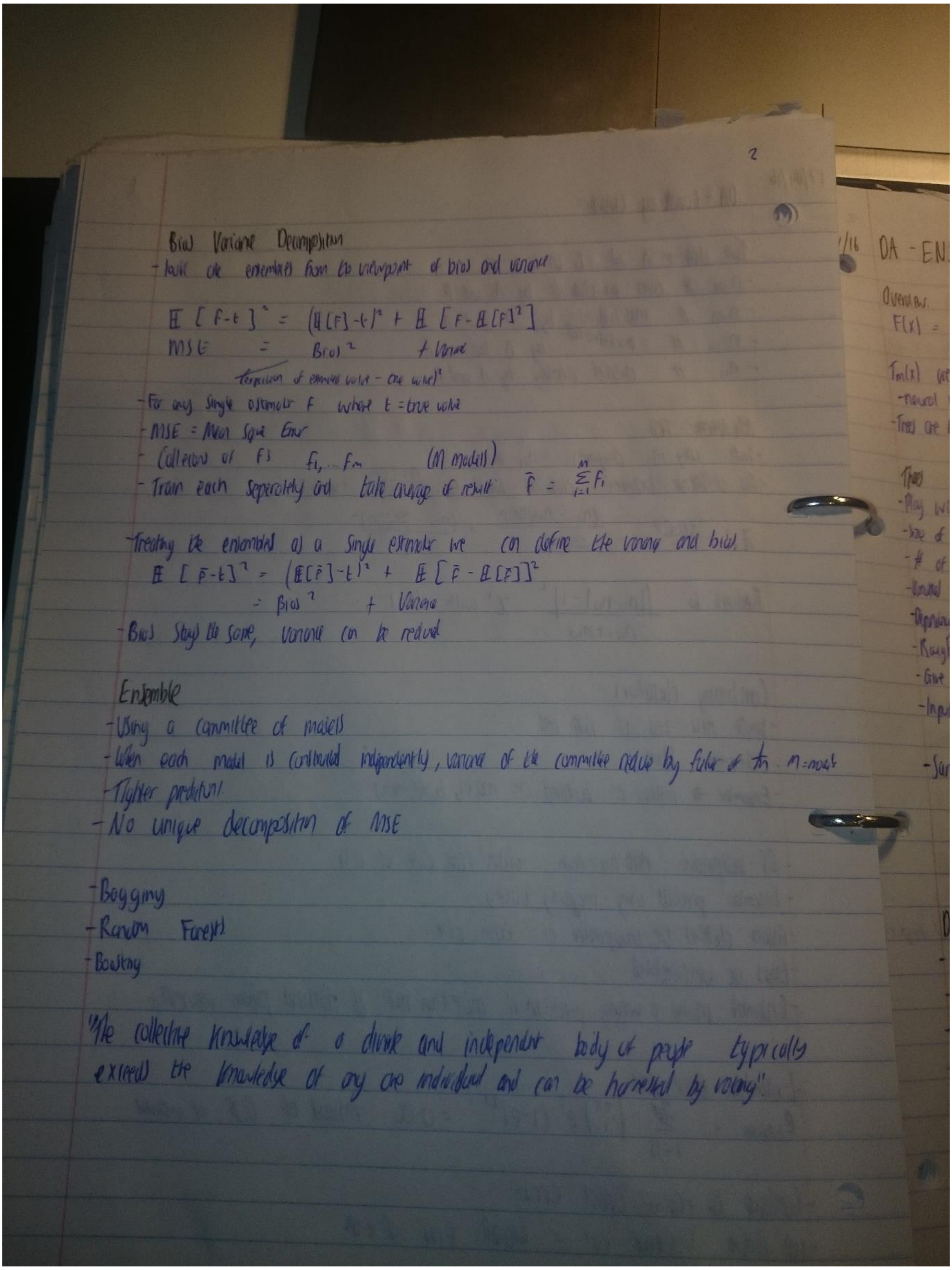
- Simple rule red if both red
- Otherwise blue
- Ensemble → instead of building one model, build many

- 25 independent classifier each with error rate of 0.35
- Ensemble predicts using majority voting
- Assume classifiers are independent of each other.
- Errors are uncorrelated
- Ensemble makes a wrong prediction if more than half of classifiers predict incorrectly

- Ensemble Error rate

$$e_{ensemble} = \sum_{i=13}^{25} \binom{25}{i} e^i (1-e)^{25-i} = 0.06 \text{ instead of } 0.35 \text{ of individual}$$

- Combine to reduce overall error
- 100 classifiers 0.4 error rate = overall error of 0.02



Bias Variance Decomposition

- look at ensembles from the viewpoint of bias and variance

$$\mathbb{E}[f-t]^2 = (\mathbb{E}[f]-t)^2 + \mathbb{E}[f-\mathbb{E}[f]^2]$$

$$MSE = Bias^2 + Variance$$

(expection of estimated value - true value)^2

- For any single estimator f where t = true value
- MSE = Mean Square Error
- Collection of fs $f_1, \ldots f_m$ (M models)
- Train each seperately and take average of result $\bar{f} = \sum_{i=1}^{m} f_i$

- Treating the ensembles as a single estimator we can define the variance and bias

$$\mathbb{E}[\bar{f}-t]^2 = (\mathbb{E}[\bar{f}]-t)^2 + \mathbb{E}[\bar{f}-\mathbb{E}[\bar{f}]]^2$$

$$= Bias^2 + Variance$$

- Bias stays the same, variance can be reduced

Ensemble

- Using a committee of models
- When each model is constructed independently, variance of the committee reduce by factor of $\frac{1}{m}$. m=models
- Tighter predictions
- No unique decomposition of MSE

- Bagging
- Random Forests
- Boosting

"The collective knowledge of a diverse and independent body of people typically exceeds the knowledge of any one individual and can be harnessed by voting".