The Gold Mine of the 21st Century
Statistical Learning, Data Mining and Visualization

February 24, 2014

Krzysztof Podgorski
School of Economics and Management
Lund University

LUND UNIVERSITY

Motto

Nothing is more practical than a good theory.

Vladimir Vapnik*

*in *Statistical Learning Theory*. John Wiley, New York (1998)

How can business benefit from data mining?

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

- identify the targets most likely to maximize return on investment in future mailings

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

- identify the targets most likely to maximize return on investment in future mailings
- forecasting bankruptcy and other forms of default

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

- identify the targets most likely to maximize return on investment in future mailings
- forecasting bankruptcy and other forms of default
- identifying segments of a population likely to respond similarly to given events

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

- identify the targets most likely to maximize return on investment in future mailings
- forecasting bankruptcy and other forms of default
- identifying segments of a population likely to respond similarly to given events
- data mining tools sweep through databases to identify patterns in the buying activities to detect fraudulent transactions

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

- identify the targets most likely to maximize return on investment in future mailings
- forecasting bankruptcy and other forms of default
- identifying segments of a population likely to respond similarly to given events
- data mining tools sweep through databases to identify patterns in the buying activities to detect fraudulent transactions
- identifying anomalous data representing data entry error

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

- identify the targets most likely to maximize return on investment in future mailings
- forecasting bankruptcy and other forms of default
- identifying segments of a population likely to respond similarly to given events
- data mining tools sweep through databases to identify patterns in the buying activities to detect fraudulent transactions
- identifying anomalous data representing data entry error
- search for patterns in human genome to detect genetic conditioning of certain diseases

How can business benefit from data mining?

Automated prediction of trends that traditionally required extensive statistical analysis and specialized expertise.

- identify the targets most likely to maximize return on investment in future mailings
- forecasting bankruptcy and other forms of default
- identifying segments of a population likely to respond similarly to given events
- data mining tools sweep through databases to identify patterns in the buying activities to detect fraudulent transactions
- identifying anomalous data representing data entry error
- search for patterns in human genome to detect genetic conditioning of certain diseases

A number companies in retail, finance, health care, manufacturing, transportation, and aerospace are already using data mining to take advantage of historical data.

Outline

What is statistical learning?

OBSERVE → Observe a phenomenon and collect data ↔ **Data mining** – analysis of (often large) data to find relationship to summarize in novel ways that are useful for the data owner

↓

MODEL → Propose a model of that phenomenon ↔ **Inference** – identification of the model that well describes the relations found in the data

↓

PREDICT → Use the model to make predictions ↔ **Prediction** – making decisions with quantified uncertainty based on the model

How statistical data mining different from statistics?

Similarities

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data $\rightarrow$ Model $\rightarrow$ Prediction

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data $\rightarrow$ Model $\rightarrow$ Prediction
- Statistical data mining in its narrower meaning is a part of statistical learning that deals with searching for a possible model that maybe attached to the data

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data $\rightarrow$ Model $\rightarrow$ Prediction
- Statistical data mining in its narrower meaning is a part of statistical learning that deals with searching for a possible model that maybe attached to the data
- Statistical data mining is using statistical (uncertainty) modeling as its methodological foundation – this differs it from data mining as understood by a computer analyst

Differences

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data $\rightarrow$ Model $\rightarrow$ Prediction
- Statistical data mining in its narrower meaning is a part of statistical learning that deals with searching for a possible model that maybe attached to the data
- Statistical data mining is using statistical (uncertainty) modeling as its methodological foundation – this differs it from data mining as understood by a computer analyst

Differences

- Statistical data mining is typically dealing with much more complex data than the standard statistics

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data $\rightarrow$ Model $\rightarrow$ Prediction
- Statistical data mining in its narrower meaning is a part of statistical learning that deals with searching for a possible model that maybe attached to the data
- Statistical data mining is using statistical (uncertainty) modeling as its methodological foundation – this differs it from data mining as understood by a computer analyst

Differences

- Statistical data mining is typically dealing with much more complex data than the standard statistics
- Emphasize is on algorithmic and computational methods to discover a model (learning from the data) rather than on analytical results for developed models

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data → Model → Prediction
- Statistical data mining in its narrower meaning is a part of statistical learning that deals with searching for a possible model that maybe attached to the data
- Statistical data mining is using statistical (uncertainty) modeling as its methodological foundation – this differs it from data mining as understood by a computer analyst

Differences

- Statistical data mining is typically dealing with much more complex data than the standard statistics
- Emphasize is on algorithmic and computational methods to discover a model (learning from the data) rather than on analytical results for developed models
- By using computational tools and algorithm, the methodological aspect is pushed in the background:

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data $\rightarrow$ Model $\rightarrow$ Prediction
- Statistical data mining in its narrower meaning is a part of statistical learning that deals with searching for a possible model that maybe attached to the data
- Statistical data mining is using statistical (uncertainty) modeling as its methodological foundation – this differs it from data mining as understood by a computer analyst

Differences

- Statistical data mining is typically dealing with much more complex data than the standard statistics
- Emphasize is on algorithmic and computational methods to discover a model (learning from the data) rather than on analytical results for developed models
- By using computational tools and algorithm, the methodological aspect is pushed in the background:
  - **automated process of statistical learning performed by computers!**

How statistical data mining different from statistics?

Similarities

- Statistical data mining in its broader meaning is identified as statistical learning which is a part of statistics since it is based on the same fundamental scheme of inference: Data $\rightarrow$ Model $\rightarrow$ Prediction
- Statistical data mining in its narrower meaning is a part of statistical learning that deals with searching for a possible model that maybe attached to the data
- Statistical data mining is using statistical (uncertainty) modeling as its methodological foundation – this differs it from data mining as understood by a computer analyst

Differences

- Statistical data mining is typically dealing with much more complex data than the standard statistics
- Emphasize is on algorithmic and computational methods to discover a model (learning from the data) rather than on analytical results for developed models
- By using computational tools and algorithm, the methodological aspect is pushed in the background:
  - **automated process of statistical learning performed by computers!**
  - **no longer require statistical expertise to put hands on the data!**

Outline

Classification problem

Classification problem

- **Overall goal**: We observe certain features of an object and we want decide to which category (or class, or population) this object belongs.

Classification problem

- **Overall goal**: We observe certain features of an object and we want decide to which category (or class, or population) this object belongs.
- The classification of an object to a class is made through a classification rule.

Classification problem

- **Overall goal**: We observe certain features of an object and we want decide to which category (or class, or population) this object belongs.
- The classification of an object to a class is made through a classification rule.
- **Goal**: Find an effective classification rule.

Learning, validation, and testing

- **Data**: By collecting relevant data we we want to

Learning, validation, and testing

- **Data**: By collecting relevant data we we want to
  - **Learn** how to **discriminate** between classes, i.e. let an algorithm run through the data to identify relevant features for the classification problem and to develop several reasonable classification rules

Learning, validation, and testing

- **Data**: By collecting relevant data we we want to
  - **Learn** how to **discriminate** between classes, i.e. let an algorithm run through the data to identify relevant features for the classification problem and to develop several reasonable classification rules
  - **Verify** how these methods perform on actual data sets and decide for the optimal method
  - **Test** how the optimal method performs on a data set that was not used yet for the discrimination and method selection stages.

Data allocation

Data allocation

- Allocate data, for example 50% for the learning phase (discrimination), 25% for validation (model/method selection), and 25% for testing phase (model assessment)

Data allocation

- Allocate data, for example 50% for the learning phase (discrimination), 25% for validation (model/method selection), and 25% for testing phase (model assessment)
- **Model/method selection**: estimating the performance of different models or methods in order to choose the best one.

Data allocation

- Allocate data, for example 50% for the learning phase (discrimination), 25% for validation (model/method selection), and 25% for testing phase (model assessment)
- **Model/method selection**: estimating the performance of different models or methods in order to choose the best one.
- **Model assessment**: having chosen a final model, estimating its prediction error on new data.

Outline

Email spam – classification problem

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks
- Objective: automatic spam detector – predicting whether the email was junk email, or spam

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks
- Objective: automatic spam detector – predicting whether the email was junk email, or spam
- Classification problem: the outcomes are discrete (bi-) valued

Classifier: which features to use and how

Classifier: which features to use and how

- Average percentage of words or characters in an e-mail message:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

Classifier: which features to use and how

- Average percentage of words or characters in an e-mail message:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Learning method has to decide which features to use and how

Classifier: which features to use and how

- Average percentage of words or characters in an e-mail message:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Learning method has to decide which features to use and how
- We might use a rule such as

```
if (%george < 0.6) & (%you > 1.5) then spam
else email.
```

Classifier: which features to use and how

- Average percentage of words or characters in an e-mail message:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Learning method has to decide which features to use and how
- We might use a rule such as

```
if (%george < 0.6) & (%you > 1.5) then spam
else email.
```

- Another form of a rule might be:

```
if (0.2  %you  0.3  %george) > 0 then spam
else email.
```

Classifier: which features to use and how

- Average percentage of words or characters in an e-mail message:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Learning method has to decide which features to use and how
- We might use a rule such as

```
if (%george < 0.6) & (%you > 1.5) then spam
else email.
```

- Another form of a rule might be:

```
if (0.2  %you  0.3  %george) > 0 then spam
else email.
```

- The problem is not 'symmetric': we want to avoid filtering out good email, while letting spam get through is not desirable but less serious in its consequences

Email spam – classification problem

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks
- Objective: automatic spam detector – predicting whether the email was junk email

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks
- Objective: automatic spam detector – predicting whether the email was junk email
- Coded: `spam` as 1 and `email` as zero

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks
- Objective: automatic spam detector – predicting whether the email was junk email
- Coded: `spam` as 1 and `email` as zero
- Training set: 3065 observations (messages) – the method will be based on these observations

Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks
- Objective: automatic spam detector – predicting whether the email was junk email
- Coded: `spam` as 1 and `email` as zero
- Training set: 3065 observations (messages) – the method will be based on these observations
- Test set: 1536 messages randomly chosen – the method will be tested on these observation

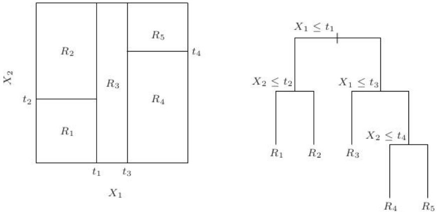
Email spam – classification problem

- Training data: 4601 email messages the true outcome (email type) email or spam is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks
- Objective: automatic spam detector – predicting whether the email was junk email
- Coded: `spam` as 1 and `email` as zero
- Training set: 3065 observations (messages) – the method will be based on these observations
- Test set: 1536 messages randomly chosen – the method will be tested on these observation
- Validation data set is not specified

Binary partition = binary tree

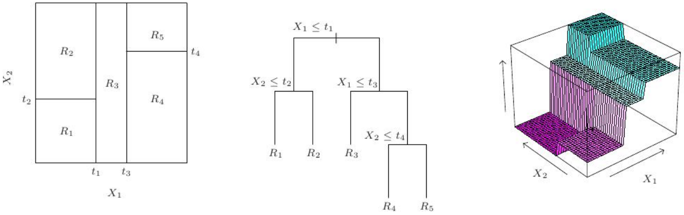
Binary partition = binary tree

- A binary partition can be presented by a sequence of decisions that can be represented as a decision tree T



Binary partition = binary tree

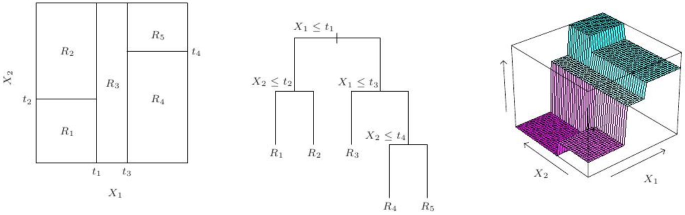
- A binary partition can be presented by a sequence of decisions that can be represented as a decision tree T



- A fit that is piecewise constant over the binary partition

Binary partition = binary tree

- A binary partition can be presented by a sequence of decisions that can be represented as a decision tree T



- A fit that is piecewise constant over the binary partition
- How to choose the values over each partition?

Data minining in action

Data minining in action

- Computer evaluates the optimal spliting points and **'grows'** a tree
- It does it in a 'gready' way to get optimal accuracy within the learning/training set.

Data minining in action

- Computer evaluates the optimal spliting points and **'grows'** a tree
- It does it in a 'gready' way to get optimal accuracy within the learning/training set.
- The obtained tree is typically over-fitting the data (too many nodes comparing to the number of the data points).

Data minining in action

- Computer evaluates the optimal spliting points and **'grows'** a tree
- It does it in a 'gready' way to get optimal accuracy within the learning/training set.
- The obtained tree is typically over-fitting the data (too many nodes comparing to the number of the data points).
- Reduction of the tree size by cutting some of the branches of an overgrown tree – **prunning**.

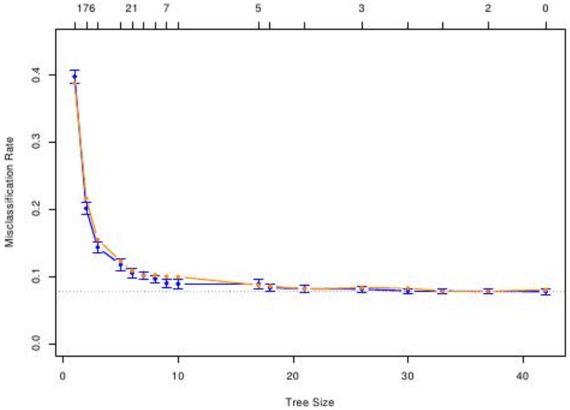
Data minining in action

- Computer evaluates the optimal spliting points and **'grows'** a tree
- It does it in a 'gready' way to get optimal accuracy within the learning/training set.
- The obtained tree is typically over-fitting the data (too many nodes comparing to the number of the data points).
- Reduction of the tree size by cutting some of the branches of an overgrown tree – **prunning**.
- After evaluation of the 'gready' tree, it is **pruned** to simplify the tree without losing the accuracy – the validation set
- Eventually the chosen tree is tested to report actual accuracy – the testing set.

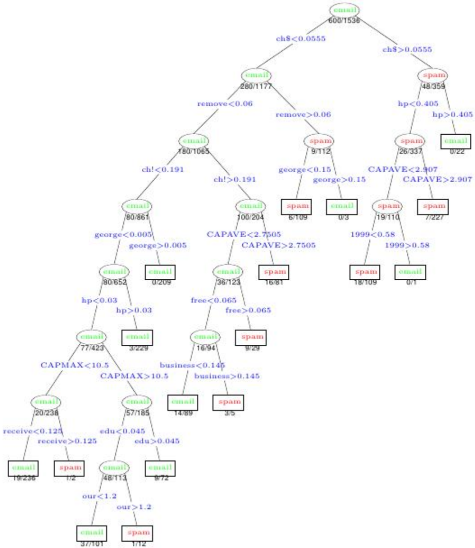
Spam example

Spam example

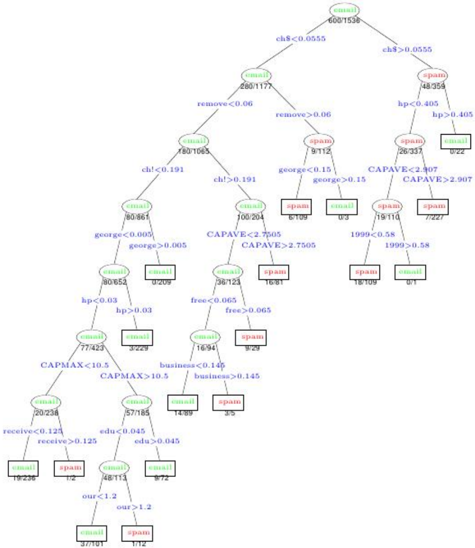
- 10-fold cross-validation error rate as a function of the size of the pruned tree, along with ± 2 standard errors of the mean, from the ten replications.



Pruned tree and conclusions

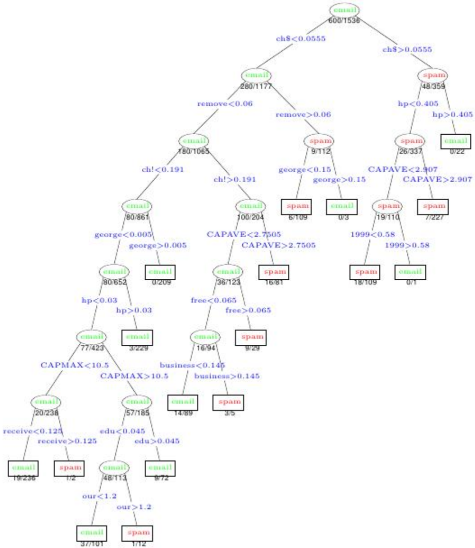


Pruned tree and conclusions



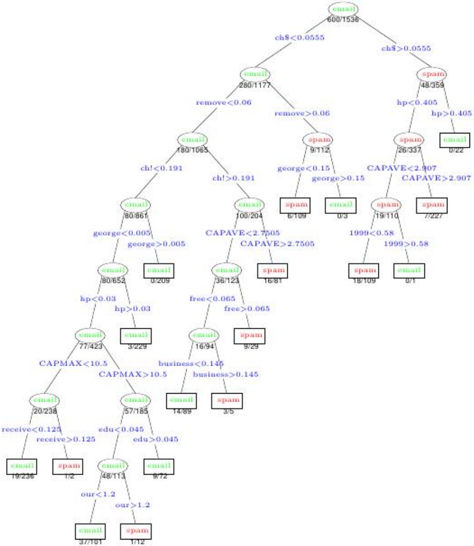
- The error flattens out at around 17 terminal nodes

Pruned tree and conclusions



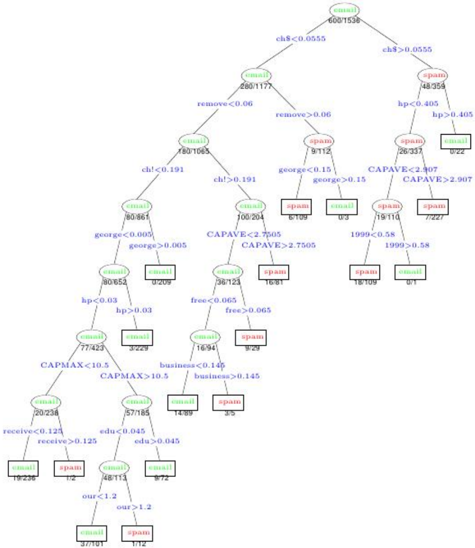
- The error flattens out at around 17 terminal nodes
- The pruned tree is shown.

Pruned tree and conclusions



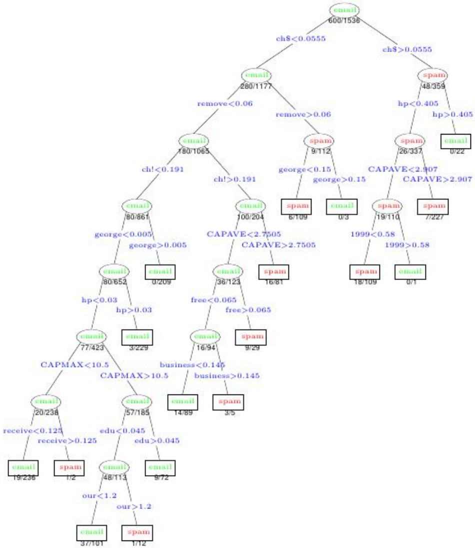
- The error flattens out at around 17 terminal nodes
- The pruned tree is shown.
- Of the 13 distinct features chosen by the tree, 11 overlap with the 16 significant features in the additive model.

Pruned tree and conclusions



- The error flattens out at around 17 terminal nodes
- The pruned tree is shown.
- Of the 13 distinct features chosen by the tree, 11 overlap with the 16 significant features in the additive model.
- The split variables are shown in blue on the branches, and the classification is shown in every node.

Pruned tree and conclusions



- The error flattens out at around 17 terminal nodes
- The pruned tree is shown.
- Of the 13 distinct features chosen by the tree, 11 overlap with the 16 significant features in the additive model.
- The split variables are shown in blue on the branches, and the classification is shown in every node.
- The numbers under the terminal nodes indicate misclassification rates on the test data.

Discussion and interpretation of the results

TABLE 9.3. *Spam data: confusion rates for the 17-node tree (chosen by cross–validation) on the test data. Overall error rate is 9.3%.*

	Predicted	
True	email	spam
email	57.3%	4.0%
spam	5.3%	33.4%

Discussion and interpretation of the results

TABLE 9.3. *Spam data: confusion rates for the 17-node tree (chosen by cross–validation) on the test data. Overall error rate is 9.3%.*

	Predicted	
True	`email`	`spam`
`email`	57.3%	4.0%
`spam`	5.3%	33.4%

- Interpretation in terms of sensitivity and specificity:

Discussion and interpretation of the results

TABLE 9.3. *Spam data: confusion rates for the 17-node tree (chosen by cross–validation) on the test data. Overall error rate is 9.3%.*

	Predicted	
True	email	spam
email	57.3%	4.0%
spam	5.3%	33.4%

- Interpretation in terms of sensitivity and specificity:
  - Sensitivity: probability of predicting spam given true state is spam.

Discussion and interpretation of the results

TABLE 9.3. *Spam data: confusion rates for the 17-node tree (chosen by cross–validation) on the test data. Overall error rate is 9.3%.*

	Predicted	
True	`email`	`spam`
`email`	57.3%	4.0%
`spam`	5.3%	33.4%

- Interpretation in terms of sensitivity and specificity:
  - Sensitivity: probability of predicting spam given true state is spam.
  - Specificity: probability of predicting e-mail given true state is e-mail.

$$\textit{Sensitivity} = \frac{33.4}{33.4 + 5.3} = 86.3\%$$

$$\textit{Specificity} = \frac{57.3}{57.3 + 4.0} = 93.4\%$$

Marketing study
CLASSIFICATION

Final Project for Data Mining Course

Lecturer:
Krzysztof Podgorski

Prepared by:
Patrik Takeuchi
&
Nima Shariati

Objective

In this project, the task is to predict which customers are most likely to respond to a direct mail marketing promotion using the clothing-store data set collected on 50 input variables and one response for 21,740 customers.

Input Variables

- Customer ID: unique, encrypted customer identification
- Zip code
- Number of purchase visits
- Total net sales (i.e. amount spent on all purchases)
- Average amount spent per visit (it should be the ratio of the previous two)
- Amount spent at each of four different franchises (four variables)
- Amount spent in the past month, the past three months, and the past six months
- Amount spent the same period last year
- Gross margin percentage
- Number of marketing promotions on file
- Number of days the customer has been on file
- Number of days between purchases
- Markdown percentage on customer purchases
- Number of different product classes purchased
- Number of coupons used by the customer
- Total number of individual items purchased by the customer
- Number of stores the customer shopped at
- Number of promotions mailed in the past year

Input Variables (continue)

- Number of promotions responded to in the past year
- Promotion response rate for the past year
- Product uniformity (low score = diverse spending patterns)
- Lifetime average time between visits
- Microvision lifestyle cluster type
- Percent of returns
- Flag: credit card user
- Flag: valid phone number on file
- Flag: Web shopper
- 15 variables providing the percentages spent by the customer on specific classes of clothing, including sweaters, knit tops, knit dresses, blouses, jackets, career pants, casual pants, shirts, dresses, suits, outerwear, jewelry, fashion, legwear, and the collectibles line; also a variable showing the brand of choice (encrypted)

and the response (target) variable is the response to promotion.

Response to marketing campaign

	Count	Percentage
Non-Responsive	18,129	83.39%
Responsive	3,611	16.61%
Total	21,740	100.00%

Transformation to achieve symmetry, Binary variables and Standardization of variables (continue)

- Standardization
  - Standardization of the values are done as to avoid the difference of variability of the variables. To achieve this we will subtract the mean and divide by the standard deviation thus giving us a mean of zero and a standard deviation of one

Relationship between Features (Predictors) and Outcomes (Response)

		Mean	Median	Standard Deviation	Minimum	Maximum	Correlation
1	LTFREDAY	0.0001	0.0290	1.0000	(6.2184)	1.9407	(0.4339)
2	FRE	(0.0000)	(0.0468)	1.0000	(1.2218)	3.8531	0.4000
3	STYLES	(0.0001)	(0.0721)	1.0000	(2.1644)	4.1306	0.3687
4	RESPONDED	(0.0000)	(0.8328)	1.0000	(0.8328)	3.1178	0.3370
5	MON	(0.0000)	(0.0868)	1.0000	(5.8436)	4.4985	0.3335
6	SMONSPEND	0.0000	(0.0663)	1.0000	(1.1043)	10.3816	0.3315
7	CLASSES	(0.0001)	0.1388	1.0000	(2.1352)	2.4476	0.3284
8	FREDAYS	0.0001	0.0321	1.0000	(5.5009)	2.0399	(0.3231)
9	RESPONSERATE	(0.0000)	(0.8505)	1.0000	(0.8505)	2.3078	0.3226
10	REC	0.0001	0.2206	1.0000	(3.4705)	1.2792	(0.2959)
11	HI	0.0001	(0.0021)	1.0000	(8.8038)	2.6371	(0.2909)
12	STORES	(0.0001)	0.0440	1.0000	(1.1389)	3.8858	0.2856
13	COUPONS	(0.0000)	(0.6921)	1.0000	(0.6921)	1.4449	0.2705
14	CC_CARD	(0.0000)	(0.7891)	1.0000	(0.7891)	1.2672	0.2411
15	OMONSPEND	0.0000	(0.5160)	1.0000	(0.5160)	1.9379	0.2378
16	TMONSPEND	0.0000	(0.9199)	1.0000	(0.9199)	1.0870	0.2330
17	PROMOS	(0.0001)	0.2393	1.0000	(2.5379)	2.4041	0.2040
18	PKNIT_TOPS	(0.0000)	(0.7709)	1.0000	(0.7709)	1.2972	0.1992
19	PFASHION	0.0000	(0.7302)	1.0000	(0.7302)	1.3694	0.1909
20	MAILED	(0.0001)	0.1127	1.0000	(1.7611)	1.3462	0.1878
21	PERCRET	(0.0000)	(0.5568)	1.0000	(0.5568)	19.7041	0.1872
22	PCAS_PNTS	0.0000	(0.8632)	1.0000	(0.8632)	1.1584	0.1791
23	PBLOUSES	0.0000	0.8460	1.0000	(1.1820)	0.8460	0.1712
24	PKNIT_DRES	(0.0000)	(0.6654)	1.0000	(0.6654)	1.5028	0.1708
25	PSHIRTS	(0.0000)	(0.8862)	1.0000	(0.8862)	1.1284	0.1694
26	PDRESSES	(0.0000)	(0.7118)	1.0000	(0.7118)	1.4049	0.1677
27	PREVPD	(0.0000)	(0.5621)	1.0000	(0.5621)	1.7791	0.1671
28	CCSPEND	(0.0001)	(0.0120)	1.0000	(8.0943)	4.2208	0.1669

Allocation of data

- 50% of the data was used for the learning phase
- 25% of the data was allocated for Validation of model/method selection
- 25% of the data was allocated for testing phase model assessment

Misclassification Cost

Average amount spent per visit

	AVRG
Mean	113.89
Median	92.07
Standard Deviation	87.25
Minimum	0.49
Maximum	1,919.88
Number of Customers	21,740

Let's assume that the profit ranges from 30% to 20% which would be normal for retail clothing. For our calculation we will assume that the profit is 25% thus making the average profit per visit to (113.89*0.25=28.47) 28.47 USD.

Cost for direct mail marketing promotions

First class letters (1 oz.)	0.49
Cost for letter	1.00

Misclassification Cost (continue)

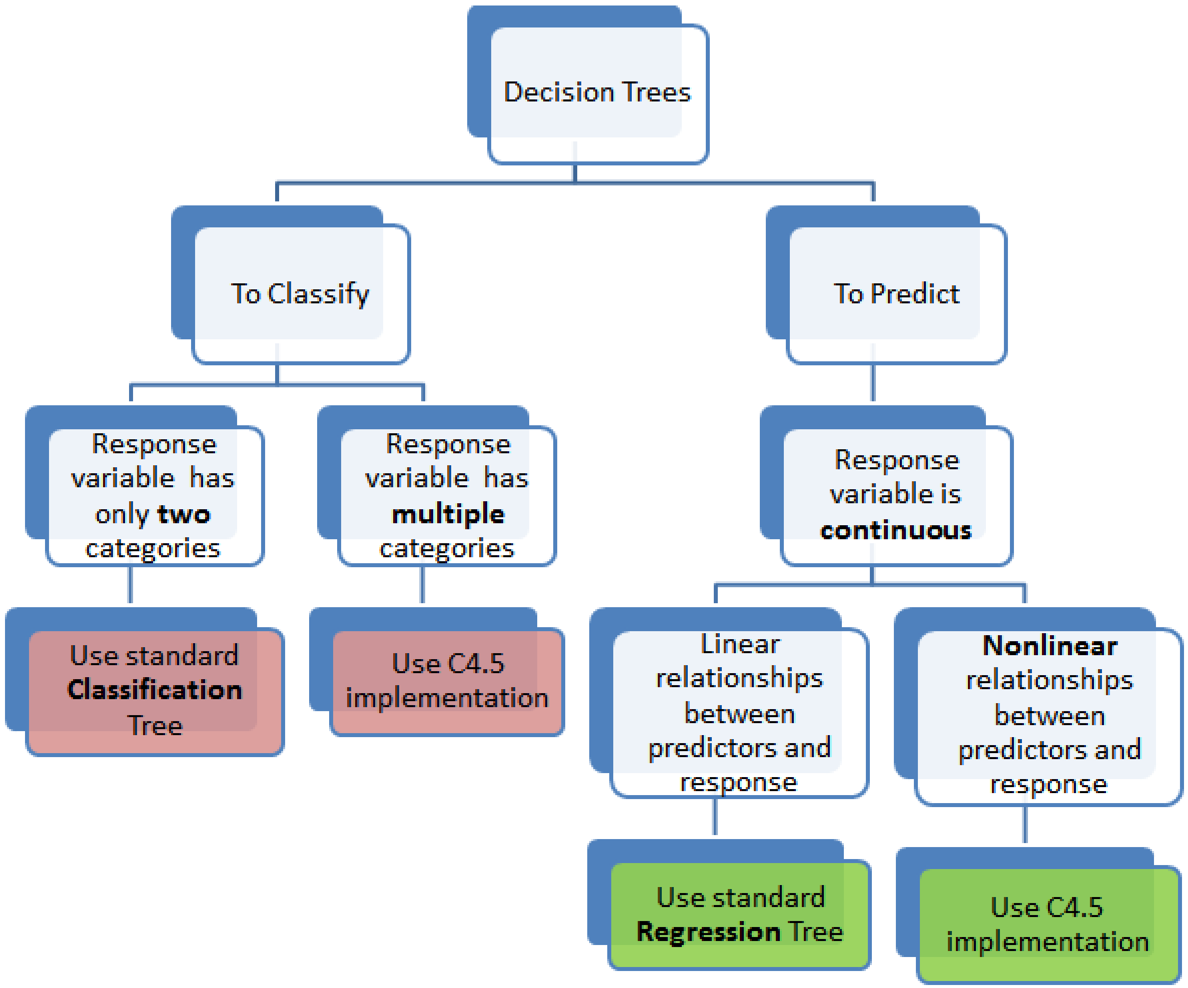
Misclassification costs

Actual Group	Predicted Group	
	Non-Responsive to promotion	Responsive to promotion
Non-Responsive to promotion	TRUE No Contact USD 0.00	FALSE Promotion sent USD 1.49
Responsive to promotion	FALSE No Contact USD 28.47	TRUE Promotion sent -USD 26.98

Cost Matrix

0	1
19.11	0

Classification models and Evaluation



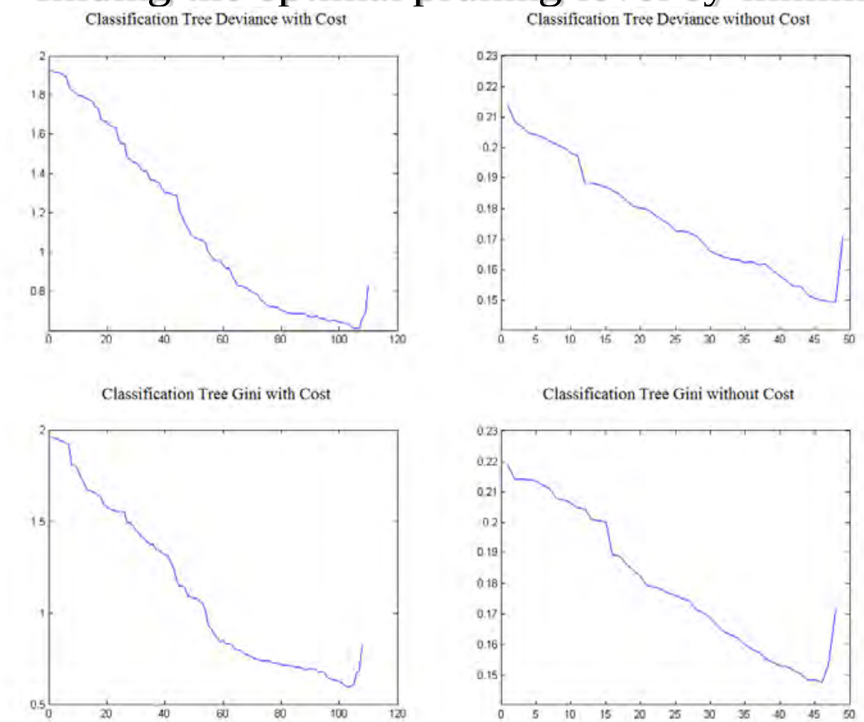
Classification models and Evaluation (continue)

- Classification Tree
  - Classification Tree using Deviance as splitting method with cost
  - Classification Tree using Deviance as splitting method without cost
  - Classification Tree using Gini Index as splitting method with cost
  - Classification Tree using Gini Index as splitting method without cost

Classification models and Evaluation (continue)

- Pruning Classification Tree

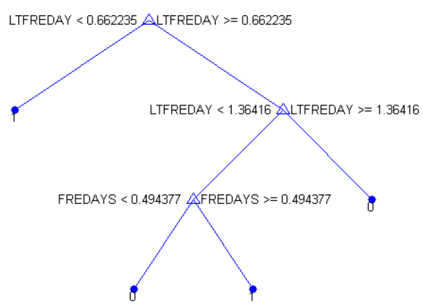
  Cross validation was used to determine the size of the trees by finding the optimal pruning level by minimizing cross-validated loss



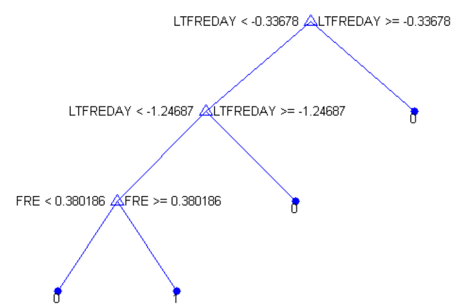
Classification models and Evaluation (continue)

- ➢ Pruned Classification Trees

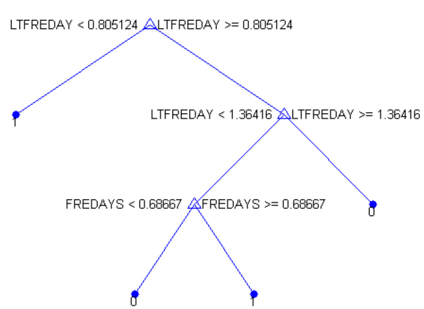
Classification Tree
Splitting Metod Deviance
With Cost



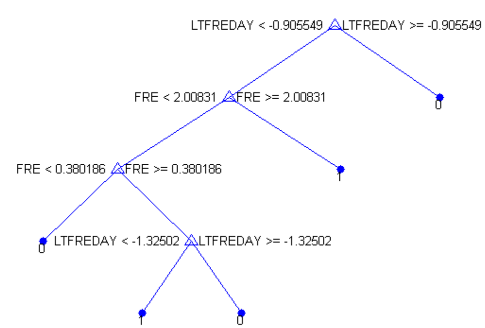
Classification Tree
Splitting Metod Deviance
Without Cost



Classification Tree
Splitting Metod Gini
With Cost



Classification Tree
Splitting Metod Gini
Without Cost



Validation of model/method

- Validation data are classified using 4 different trees to see which classification tree performs best

Confusion Matrix in Counts

Deviance with Cost

Actual	Predicted Non-Res	Predicted Res	Missclass ratio
Non-Res	832	3730	0.81762
Res	0	873	0.00000
		APER=	0.68629

Deviance without Cost

Actual	Predicted Non-Res	Predicted Res	Missclass ratio
Non-Res	4358	204	0.04472
Res	582	291	0.66667
		APER=	0.14462

Gini with Cost

Actual	Predicted Non-Res	Predicted Res	Missclass ratio
Non-Res	797	3765	0.82530
Res	0	873	0.00000
		APER=	0.69273

Gini without Cost

Actual	Predicted Non-Res	Predicted Res	Missclass ratio
Non-Res	4380	182	0.03989
Res	611	262	0.69989
		APER=	0.14591

Confusion Matrix in USD

Deviance with Cost

Actual	Predicted Non-Res	Predicted Res	Total Cost
Non-Res	USD 0.00	USD 5,557.70	USD 5,557.70
Res	USD 0.00	(USD 23,553.54)	(USD 23,553.54)
		Total Cost	(USD 17,995.84)

Deviance without Cost

Actual	Predicted Non-Res	Predicted Res	Total Cost
Non-Res	USD 0.00	USD 303.96	USD 303.96
Res	USD 16,569.54	(USD 7,851.18)	USD 8,718.36
		Total Cost	USD 9,022.32

Gini with Cost

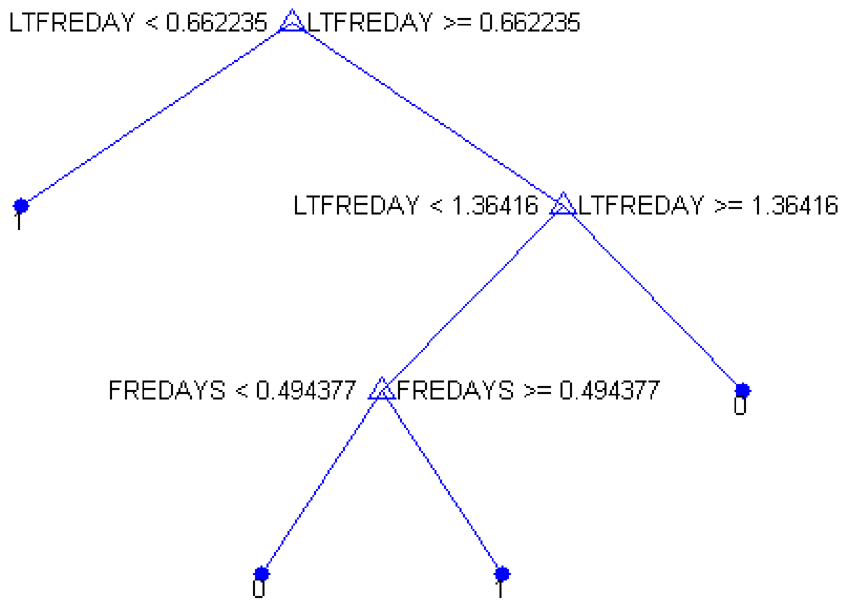
Actual	Predicted Non-Res	Predicted Res	Total Cost
Non-Res	USD 0.00	USD 5,609.85	USD 5,609.85
Res	USD 0.00	(USD 23,553.54)	(USD 23,553.54)
		Total Cost	(USD 17,943.69)

Gini without Cost

Actual	Predicted Non-Res	Predicted Res	Total Cost
Non-Res	USD 0.00	USD 271.18	USD 271.18
Res	USD 17,395.17	(USD 7,068.76)	USD 10,326.41
		Total Cost	USD 10,597.59

Assessment of model

- Model chosen was the classification tree with splitting criteria deviance including misclassification cost



Assessment of model (continue)

Results

Confusion Matrix chosen classification tree

Deviance with Cost in Count

Actual	Predicted Non-Res	Predicted Res	Missclass ratio
Non-Res	896	3638	0.80238
Res	2	899	0.00222
		APER=	0.669733

Deviance with Cost in USD

Actual	Predicted Non-Res	Predicted Res	Total Cost
Non-Res	USD 0.00	USD 5,420.62	USD 5,420.62
Res	USD 56.94	(USD 24,255.02)	(USD 24,198.08)
		Total Cost	(USD 18,777.46)

Thank you