10 marks
DAL 10 March 2017

Master of Technology in Enterprise Business Analytics

Neural Networks for Business Analytics

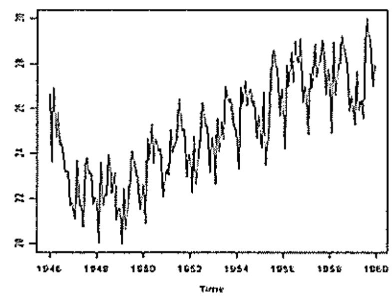
NN Forecasting Workshop

Dr. Barry Shepherd
Institute of Systems Science
National University of Singapore
E-mail: barryshepherd@nus.edu.sg

4/2/2017

Why use NN's for Forecasting?

- Time-Series approaches to Forecasting use only previous the values of the target variable
  - Naïve Approach, *e.g. next period sales = last period sales*
  - Moving Average, *e.g. next period sales = avg. of last N period sales*
  - *Weighted Moving Average,* e.g. $\text{sales}_{t+1} = (0.6^*\text{sales}_t + 0.3^*\text{sales}_{t-1} + 0.1^*\text{sales}_{t-2})$
  - Exponential Smoothing, *older data is given progressively less weight*
  - Box-Jenkins Methods (ARMA, ARIMA),...



A time series is simply a sequence of numbers collected at regular intervals over a period of time

- NN forecasting models can also incorporate causal factors as inputs
  - Assumption: the variables used for prediction (the independent variables) have some cause-and-effect relationship with the predicted variable (the dependent variable)
  - Correlations will suffice – but there should be some plausible link with the predicted event

Stock Index Prediction Workshop

- A data file contains daily values of the Straits Times Index (STI) from 1 Jan'04 to 31 Dec'13 . The fields are:
  - Date
  - Opening Price
  - Closing Price
  - Adjusted Close (similar to above but after minor accounting changes)
  - Highest value during the day
  - Lowest value during the day
  - Volume (total amount of traded stocks during the day)
- The goal is to predict the short term STI trend with sufficient accuracy for profitable use in a (simplified) trading scenario
- Tools, Your can choose either of:
  - Using R (full instructions in the workshop hand-out if you are unsure how to proceed)
  - Using SPSS Modeler (an example stream is given)

Task1: Predict Tomorrow's Closing Value

- There isn't a target field in the raw data – you need to create one.
  Create a target variable (tomorrow's close) by duplicating the column "close" but with a row offset of one, e.g.

Day1	Open	High	Low	Close	Tomorrow_Close
Day2	Open	High	Low	Close	Tomorrow_Close
Day3	Open	High	Low	Close	

  - *Use R or edit your copy of the data file in Excel or use the SPSS* ***History*** *node*
- Use all records with dates < Jan 1st 2012 for **training** and dates after this as **test** data
  - *In SPSS use a* ***Select*** *node with the expression: Date < datetime_date(2012,01,01)*
- Build the best model
  - Compare models using the mean absolute error (MAE) on the test data set
  - *(in SPSS modeler we can use an* ***Analysis*** *node to obtain this)*

Task2: Predict the STI Trend

- In practice, trading decisions are often made based on an assessment of market direction: bull or bear (increasing or decreasing) – hence a prediction of market direction (trend) and strength may be more useful
- We can predict trend either by:
  1. Subtract todays STI value from the predicted future STI value - if the value is positive then the trend is increasing, if negative then decreasing, else no-change
  2. Create a new target variable called **"trend"** with the values: increasing, decreasing, no-change and build a new prediction model using this target

> trend = increasing if (tomorrows close – todays close) > N
> trend = decreasing if (tomorrows close – todays close) < -N
> else trend = no-change
>
> *(Use N = 0 to begin with, then try increasing N to improve performance)*

- We will try both methods in this workshop and compare results

Task2: Predict the STI Trend

- The best test of the usefulness of a predictive model is to evaluate its performance in the context in which it will be used
- For this workshop we will evaluate the model in a simple simulated trading scenario

> <u>You bet $1 on the STI trend at the start of each day.</u>
>
> If you bet the STI would increase and it did increase then you gain $1.1 but if the STI decreased then you get back <u>$0.9*</u>
>
> If you bet the STI would decrease and it did decrease then you gain $1.1 but if the STI increased then you get back <u>$0.9</u>
>
> If there was no-change in the STI then you get $0.95 back

**For simplicity we assume 10% gain or loss after fees*

Would you make money with your model using these trading rules and trading over the period of the test data?

Deriving Trend Variables

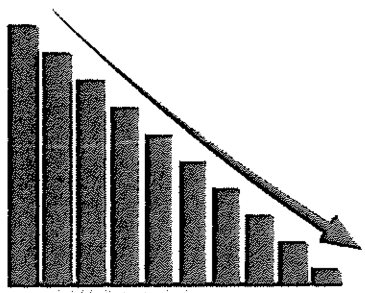
- Predictive model performance may be enhanced if we derive additional input variables that capture any trends in the raw model inputs
- **Example: Forecasting Telco Churn**
- If we can predict who might churn (close their account) then we can take counter measures – e.g. offer them a special deal

Personal data	age, gender, postcode etc
Account data	type of plan number of handsets, activation date etc
Billing data	number of calls, total $ spent, roaming activity etc
Call records (CDR's) huge quantity!	caller & receiver phone numbers date of call, time of call duration

Typical Telco Data available for churn modeling.

Trend Variables for Telco Churn

- We need to detect people whose behavior is changing in a way that suggests they might churn in the future
  - Are they using their phone less?
  - Is their billing amount decreasing over time?
- Example trend variables
  - Total calls over past N months
  - % change in call volume over time
  - Ratio of each months value to the total
  - Ratio between successive months & between the first & last month



FOR THIS WORKSHOP…
Derive extra input variables to help capture the trend.
E.g. todays close – yesterdays close
(for SPSS either edit the input data file or add a ***Derive*** *node to your stream)*

Telco Churn Sample Data

- One record per customer, each contains customer account details, usage last month and over whole tenure

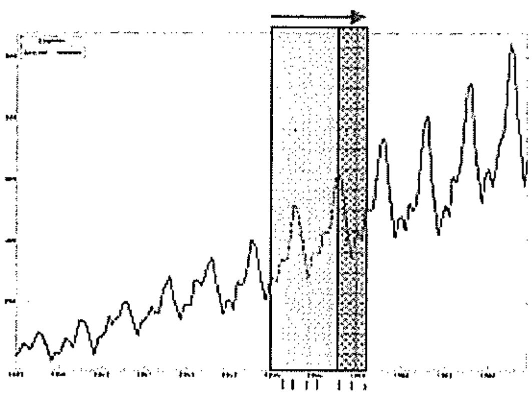
	Variable	Description	Measure
1	custid	Customer ID	nominal
2	region	Geographic indicator	scale
3	tenure	Months with service	scale
4	age	Age in years	nominal
5	marital	Marital status	scale
6	address	Years at current address	scale
7	income	Household income in thousands	ordinal
8	ed	Level of education	scale
9	employ	Years with current employer	nominal
10	retire	Retired	nominal
11	gender	Gender	nominal
12	reside	Number of people in household	scale
13	tollfree	Toll free service	nominal
14	equip	Equipment rental	nominal
15	callcard	Calling card service	nominal
16	wireless	Wireless service	nominal
17	longmon	Long distance last month	scale
18	tollmon	Toll free last month	scale
19	equipmon	Equipment last month	scale
20	cardmon	Calling card last month	scale
21	wiremon	Wireless last month	scale
22	longten	Long distance over tenure	scale
23	tollten	Toll free over tenure	scale
24	equipten	Equipment over tenure	scale
25	cardten	Calling card over tenure	scale
26	wireten	Wireless over tenure	scale
27	multline	Multiple lines	nominal
28	voice	Voice mail	nominal
29	internet	Internet	nominal
30	callid	Caller ID	nominal
31	callwait	Call waiting	nominal
32	forward	Call forwarding	nominal
33	confer	3-way calling	nominal
34	ebill	Electronic billing	nominal
35	loglong	Log-long distance	scale
36	logtoll	Log-toll free	scale
37	logequi	Log-equipment	scale
38	logcard	Log-calling card	scale
39	logwire	Log-wireless	scale
40	lninc	Log-incom	scale
41	custcat	Customer category	nominal
42	churn	Churn within last month	nominal

Target

CDR's were aggregated to generate the monthly and tenure usage

Training NN's for Forecasting

- Sliding Window Method
  - Strict division of data into training and test sets based on time
  - Refresh the model on a regular basis



Look-back period

Predict ahead period

Issues

- How far to look back?
- How frequently to update the model?
- How far to predict ahead? further ahead is generally less accurate

Example: Telco Churn

- Predict "who will close their account NEXT MONTH"
- Use monthly account activity summaries (billing cycle)
  - Assuming 6 months of data....
  - Compute trend variables from first 5 months
  - Compute target field from the 6th month (T/F ~ account closed in this month)
  - Use model to predict churners in the 7th month

	1	2	3	4	5	6	7	8
	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Build the model	Compute trends					Target known		
Use the model		Scoring Set					Target unknown	

Example: Telco Churn

- In practice, August billing data is not available until mid September,
  - Hence we can't run the model to predict Sep churners until mid-Sep
  - Gives the marketing department no time to take preventative action
- We need a gap (latency) => must predict 1 or 2 months ahead
- E.g.

	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Build the model	Compute trends					Target known			
Use the model			Scoring Set					?	

Example: Telco Churn

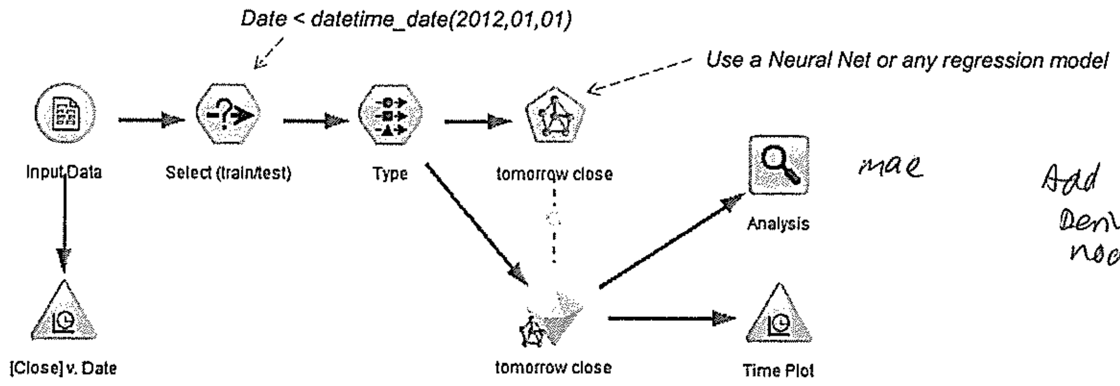
- Experiment with different training schemes to see which generates the best performing model
- For example, if the derived trend variables are based only on the previous 2 months then more training data can be derived from the available 6months data

Training Set = union of all yellow areas

Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
trends			target				
	trends			target			
		trends			target		
				Scoring			?

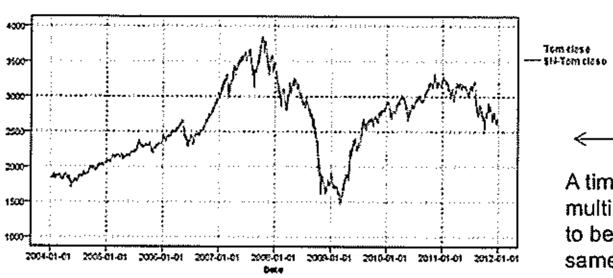
The training set now also includes May & Jun churners

Stock Index Prediction: Sample SPSS Stream

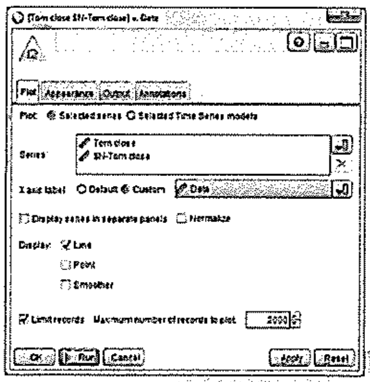


mae

Add Deriv node.



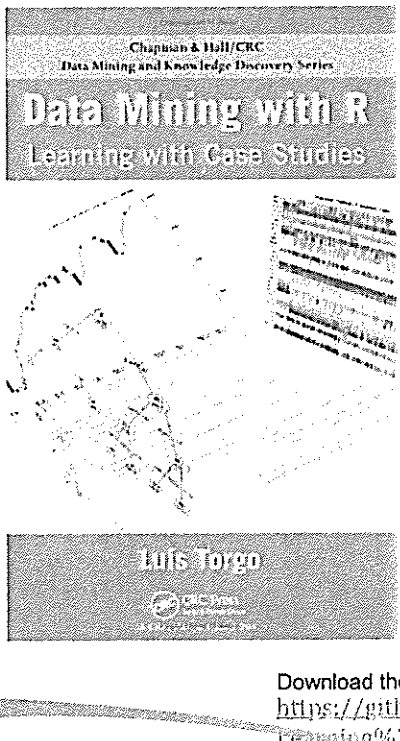
A time plot allows multiple time series to be plotted on the same axis



Workshop Report Guidelines

- This workshop counts for 10 marks – work in teams
- Hand in your best SPSS modeler stream file and/or your R code
  PLUS a short report by the last day of the unit (upload to IVLE or email to me).
  Include your modified input data file too – I may want to re-run your model.
- Give your report & stream a file name that is unique to your team. Insert all of your team member names into the report.
- Report Guidelines
  - List the variables used as model inputs and as model target
  - Describe any data transformations & new variables created
  - Describe the NN architecture you used: # nodes etc. Δ la
  - List the MAE for you best model for task1. Compute for both training and test set
  - Show the confusion matrix for the model for task2. – trading
  - State how much money would you win or loose using the model
  - (Optional) Add the Nikkei and S&P data as additional model inputs & compare performances

Similar Workshop and R Code at...



Chapter 3

Predicting Stock Market Returns

This second case study tries to move a bit further in terms of the use of data mining techniques. We will address some of the difficulties of incorporating data mining tools and techniques into a concrete business problem. The specific domain used to illustrate these problems is that of automatic stock trading systems. We will address the task of building a stock trading system based on prediction models obtained with daily stock quotes data. Several models will be tried with the goal of predicting the future returns of the S&P 500 market index. These predictions will be used together with a trading strategy to reach a decision regarding the market orders to generate. This chapter addresses several new data mining issues, among which are (1) how to use R to analyze data stored in a database, (2) how to handle prediction problems with a time ordering among data observations (also known as time series), and (3) an example of the difficulties of translating model predictions into decisions and actions in real-world applications.

Download the book at:
https://github.com/hudoop/Rstudy/blob/master/Data%20Mining%20with%20R-Learning%20with%20Case%20Studies(Luis%20Torgo%202011).pdf

