

1 Introduction

This is an advanced course offered by the the University of East Anglia as part of the MSc in Knowledge Extraction

The course will be concerned with applications of statistical modelling in insurance and is primarily for graduate staff at Norwich Union. The lecture component will be taught and assessed from Thursday 3 May for 12 weeks. This taught component will be followed by an individual, supervised project. A syllabus and details of assessment are attached.

The teaching team will comprise

Dr. G.Janacek (CMP)	g.janacek@uea.ac.uk
Dr. Tony Bagnall (CMP)	ajb@cmp.uea.ac.uk

2 Aims

It is clear that in the limited time available for this course we cannot provide a complete technical background to statistical modelling. The aims of this course are consequently more modest. We attempt cover the main ideas and techniques of modelling and show how some critical ideas come together to enable us to produce useful models. We also show how things may go wrong.

We assume very little apart from some basic probability, statistics and calculus. Obviously a more comprehensive knowledge and appreciation of statistics will help enormously. A provisional syllabus is given below but there will undoubtedly be some variation to suite the class background and needs.

Since this course is part of a Data Mining MSc programme we will include some statistical ideas encountered in "Knowledge Discovery" software.

3 When and where

On this course one lecture per week will be held in house at Aviva Norwich. The times and places are as follows:

Please note that all classes are 09:30-11:30 starting on Thursday 3 May.

4 Assessment

The taught part of the course (CMPSMC2Y) carries 20 credits and will be assessed by two pieces of course work. A further project based on applications of the lectures is expected from student and this project(CMPSMC4Y) is worth an extra 20 credits. The projects will begin after Christmas and for the project each participant will have a project supervisor. Additional information and advice will of course be available before the projects start.

Please note for Aviva purposes student must take both CMPSMC2Y and CMPSMC4Y.

5 Syllabus

An approximate guide to content of the taught part of the course is as follows.

Background

1. *Week 1: Probability and distributions* Basic distributions and properties including loss distributions. Joint and conditional distributions. Bayes theorem. Conditional expectation.
2. *Week 2: Dealing with data.* Basic data exploration, graphical tools, EDA. Populations and sampling. Sampling distributions. The concept of likelihood, parameter estimation. Distributions of estimates. Confidence intervals. An overview of testing. An introduction to Bayes inference.
3. *Week 3: Linear modelling 1: Fitting the model.* Motivation for wanting to fit a linear model and the most commonly used method for doing so.
4. *Week 4: Linear modelling 2: Using the model.* Using the model to explain underlying relationships in the data and to predict new data.
5. *Week 5: Linear modelling 3: Assessing the model.* Assessing how good the model is at explaining variation in the data and at predicting new data.
6. *Week 6: Linear modelling 4: Searching for the best model.* Popular methods of searching for the best parsimonious model.
7. *Week 7: Generalized linear models.* Definitions and rationale. Glim for Gamma based models and other continuous non-normal cases. Data exploration and diagnostics.
8. *Week 8: Categorical data.* Binary data. modelling Binomials, over dispersion. Poisson data. Multinomials. Loglinear models for multiway tables.
9. *Week 9: More on model critique etc.* Case studies, Time series, Survival distributions
10. *Week 10: More useful methods*

11. *Week 11: Classification Trees.* An introduction to classification trees and a description of the implementation used in KnowledgeStudio
12. *Week 12: Bayesian Networks* An introduction to Bayesian networks.

The content and structure of the course is provisional and may vary depending on the composition of the class. The lecture scheme will be confirmed at the second session.

6 Books

There is not a book of the course that we can happily recommend, however the following may prove useful:

- *Problem Solving - A Statisticians's Guide* by C.Chatfield, Published by Chapman and Hall is excellent reading.
- The GLIM bible is of course *Generalized Linear Models* by McCullagh and Nelder (Chapman and Hall).

For the early part of the programme any reasonable statistics text should be sufficient. A random list of texts is:

6.1 Basic Theory

There are many good (and some unreadable) books which cover the basics of statistical inference.

- *Theoretical Statistics*, Chapman and Hall by Cox and Hinkley
- *Kendall's Advanced theory of statistics*, published by Arnold is now in many volumes
- *Statistical Inference*, Prentice Hall by Garthwaite, P Jolliffe,I and Jones,B
- *Parametric statistical inference*, Oxford, Lindsey, James K

7 Data Analysis etc

We will be talking about exploring data,

- *Understanding Data Analysis*, Open University Press, Erickson, B and Nosanchuck, T.
- *Exploratory Data Analysis*, Addison-Wesley, Tukey, J.

- *Permutation tests*, Springer, Good, P
- *An Introduction to the bootstrap*, Chapman and Hall, Efron, B and Tibshirani

8 Regression etc.

There are many regression books

- *Residuals and influence in regression*, Wiley by R. Dennis Cook and Sanford Weisberg is a bit different
- *Applied regression analysis*, Addison Wesley by Draper, N. R., Norman Richard is a classic
- *Applied nonparametric regression*, CUP, Hardle, W.
- *Methods and applications of linear models*, Wiley, Hocking, R. R.

9 GLIM type topics

- *Generalized linear Models*, Chapman and Hall/CRC, P. McCullagh, et al.
- *Generalized Linear Models: Applications in Engineering and the Sciences*, John Wiley, Raymond H. Myers, et al.
- *Statistical modelling in GLIM*, Oxford, Aitkin, Murray et al.
- *GLIM for ecologists*, Oxford, Crawley, Michael J.
- *An introduction to generalized linear models*, Chapman and Hall, Dobson, Annette J,
- *Generalized Linear Models: A Unified Approach*, Sage Publications Ltd (Quantitative Applications in the Social Sciences Series.), Jeff Gill.
- *Interpreting Probability Models: Logit, Probit and Other Generalized Linear Models*, Sage Publications Ltd (Quantitative Applications in the Social Sciences Series.), Tim Futing Liao.
- *Nonparametric regression and generalized linear models*, Chapman and Hall, Green, P. J.
- *Glim : an introduction*, Oxford, Healy, M. J. R.
- *Graphical Models in Applied Mathematical Statistics*, Wiley, J. Whitaker.

- *Categorical Data Analysis*, Wiley, A.Agresti.
- *Modelling Binary Data*, CRC Press. D.Collett.

10 Electronic documents

Copies of all the documents associated with this course will normally be available on the web, either

1. On Blackboard at <http://www.blackboard.uea.ac.uk>
2. Or on <http://www.uea.ac.uk/~gj/nunion/>

While we will try and convert documents to `http` if possible most complex documents will be in `pdf` and you will need a pdf reader.

G.Janacek
School of Computing Sciences
UEA, Norwich NR4 7TJ,UK

e-mail G.Janacek@uea.ac.uk
tel 44-(0)1603-591206
fax: 44-(0)1603-593345