Statistical Databases and Registers with some datamining

a course in Survey Methodology and Official Statistics

Departement of Statistics
Stockholm University

October 2010

The course and its staff

The course consists of three parts: Databases, Registers and Data mining (10 lectures each) and the syllabus is

1. A real world example of how to build a database at KI: PhD Jan Hagberg, KI
2. Database theory: Professor Bo Sundgren, SU
3. Data editing: Senior Statistician Anders Norberg, SCB
4. Commersial registers: Per Weidenman, PAR
5. Registers are databases that contains processed data: Professor Bo Sundgren, SU
6. National and international registers. How to make a good presentation: Sr Business Intelligence Analyst Alf Fyhrlund, SCB and PhD student Zhang Zhongxing, Gapminder.
7. Statistical learning: FL Mikael Möller
8. Real world examples of data mining: Mathias Lanner, SAS institute

Literature

1. In database theory notes by Bo Sundgren.
2. In register theory the book by Wallgren, A. och Wallgren, B. *Register-based Statstics – Administrative Data for Statistical Purposes*. Chichester, Wiley.
3. In statistical learning the book by Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The elements of statistical learning*. This book may also be found as a free pdf.
4. In data editing the following articles by Anders Norberg: Swedish Editing Methods and a short article.
5. Lecture notes to be found at the course home page.

Examination

The exam consists of a total of 10 tasks and each task give at most 10 credits.

Eight of the tasks shall be solved at the exam date which is January the 14th, 2011 (answers in either svenska or English) or at reexam which is February 14th, 2011.

There is one assignment that consist of two tasks. This assignment should be finished and published in 2010.

Since each task give at most 10 credits a maximum of 100 credits is possible.

Course must be finished February 2011. Next available examination time is next time the course is given.

The written exam tasks

Database theory: The exam will consist of 3 questions.

Register theory: The exam will consist of 3 questions.

Statistical learning: The exam will consist of 2 questions.

Assignment: Consists of 2 tasks.

Final grade

Final grading is according to the following table

Betyg		**Poäng**
A	Excellent	$90-100$
B	Very good	$80-89$
C	Good	$70-79$
D	Satisfactory	$60-69$
E	Adequate	$50-59$
Fx	Insufficient	$30-49$
F	Fail	≤ 29

Examiner and coordinator: Mikael Möller

Database and register

The subject register is under development and hence there are confusing notations. The following is my personal view.

- A **database** is a collection of tables that obeys certain parsimonious criteria. Once a data is entered it is (almost) never changed.
- A **register** is a collection of tables where updating is common. A register is usually not parsimonious.
- Registers are split into two different types
  1. **Administrative** registers: primarily used in administrative information systems
  2. **Statistical** registers: primarily used for statistical information as sums, means, deviations and so on. They are usually based on data from the administrative registers

Specialized registers

There is also another division of the registers into **base** and **specialized** registers

- **Administrative base** registers are kept as a basic resource for public administration
- **Statistical base** registers are registers for statistics and they are based on administrative base registers
- **Administrative specialized** registers such as the vehicle register
- **Statistical specialized** registers are statistical registers based on several administrative registers

Example of registers

In Sweden we have four official administrative **base** registers

- Register on persons
- Register on property
- Business register
- Activity register

But each company have registers of their own.

- Register of customers
- Register of transactions
- Register on products
- and so on

Statistical learning

When we start to study the subject datamining we will realize that it is a question of supervised and unsupervised learning.

And with learning we mean estimation of parameters.

Hence statistical learning would be a better name for the statistical procedures in datamining.

Our intention is to study

1. Supervised learning
   1. Linear Regression and Nearest Neighbour
   2. Neural Networks
2. Unsupervised learning
   1. Cluster Analysis
   2. The Google PageRank Algorithm

Let's get started