Statistical modelling with huge volumes of real data: a case study

The purpose of this talk is to discuss a number of important issues arising when for a project we were asked to model a data set with more than 8 millions of observations and more than 1500 candidate variables. In fact, the dataset was not available but we could derive it from different databases. I would like to share a series of important issues that arise towards modelling such data and start from problems with the data themselves (volume, granularity, accuracy etc.) as well as how we can run simple methods like logistic regression for such dataset.

Περίληψη

Σκοπός της ομιλίας είναι να συζητήσουμε και να αναδείξουμε μερικά σημαντικά και ενδιαφέροντα προβλήματα που προκύπτουν όταν προσπαθήσει κανείς να εφαρμόσει γνωστές μεθοδολογίες σε τεράστιους όγκους δεδομένων οι οποίοι μάλιστα δεν είναι απαραίτητο να είναι αποθηκευμένοι και στην ίδια βάση δεδομένων. Στην περίπτωση που θα συζητήσουμε τα δεδομένα είναι πάνω από 8 εκατομμύρια παρατηρήσεις, και 1500 υποψήφιες μεταβλητές, που είναι αποθηκευμένα σε διάφορες βάσεις δεδομένων με αρκετά προβλήματα ομοιογένειας. Μια σειρά από σημαντικά προβλήματα, όχι μόνο μεγέθους, προκύπτουν και θα ήταν χρήσιμο να παρουσιαστούν καθώς και λύσεις όπου αυτό είναι δυνατόν. Τα δεδομένα είναι μεγάλα σε όγκο, με προβλήματα στον ορισμό και την ομοιογενείας τους αλλά και τη συχνότητα μέτρησης τους, καθιστώντας το πρόβλημα αρκετά δύσκολο και ενδιαφέρον. Σκοπός ήταν να τρέξουμε μοντέλα λογιστικής παλινδρόμησης για να προβλέψουμε το γεγονός που μας ενδιέφερε χρησιμοποιώντας τις πιο κατάλληλες μεταβλητές