Machine Learning, Data Mining and Intelligent Data Analysis in Seismology

Matthias Ohrnberger, Institute of Earth and Environmental Sciences, University of Potsdam, Germany

When reading this keynote's title, one might immediately wonder about the wording. "Machine Learning", "Data Mining" and "Intelligent Data Analysis"? Isn't that the same thing or is it important to make those differences? I would like to answer either way. Yes it is the same thing when one chooses to view the idea of "letting the data speak" or "extracting patterns from data" as unifying basic concept for those terms. On the other hand, at least in my personal view, there are slight differences between these expressions that are worth considering. The term "Machine Learning" (ML) certainly emphasizes the learning aspect in an algorithmic and/or conceptual sense. "Data mining" (DM) on the other hand puts its focus on the potential increase of knowledge given the data. As in real "mining", usually one is seeking for some specific object or group of objects, but it is hard or tedious to find in the enormous amount of data. And as in real mining, one might stumble over surprising findings or unknown issues while looking for the object of interest. ''Intelligent data analysis" (IDA) finally is mostly used as a synonym for data mining, but – again my personal view - implies the inclusion of prior information into the data learning and data analysis processes. This prior information may be either knowledge about the data structure like distributional information or some particular expert knowledge regarding the underlying, potentially complex, process which has generated the observed data.

In observational seismology (as probably in other applied sciences as well) we encounter two opposing situations with respect to data availability in which the use of ML- and IDA-techniques may not only prove to be beneficial but may be even regarded as essential or indispensable. One of those areas may be called "pattern extraction" or "data screening" and is related to that part of observational seismology which is drowning in data, i.e. the observation and analysis of raw waveform data. Today, the seismological community records and archives several hundreds of GigaBytes/day of continuous waveforms just counting the data, which is made immediately available to the interested researcher and is served conveniently through various established network request and/or data interchange mechanisms. Although automatic procedures for waveform data analysis have been pursued since decades and are admittedly quite successful in detecting and locating seismic events, there is still need for many improvements. The densification and augmentation of the global seismometer network has not yet come to an end (if it will at all) and thus, any attempt related to harvesting of information from the mass of data recordings via ML- or IDA- techniques is worthwhile for exploring the spatially ever more densely sampled seismic wavefield. Opposed to this situation seismologists also face similar problems like other natural science disciplines which cannot re-create their objects of study by themselves in the laboratory. The Earth as natural laboratory does of course not generate on demand the data we would like to obtain for confirming or rejecting our physical model hypothesis regarding the earthquake source process. Thus, the 2nd area of high interest for the application of machine ML-, DM- and IDA- techniques are those areas of observational seismology where one has to deal with the problem of making inferences from sparse and partially missing data. A prime area of interest is the wide field of seismic hazard analysis where the number of instrumental recordings of relevant / damaging earthquakes are small when it comes down to estimate ground motions attenuation models.

In this presentation I will show examples of IDA-techniques applied to both of the above described situations: data screening and inference from sparse data. In the context of data reduction I will provide examples from the classical context of automated detection of seismic signals of interest from continuous data streams. Both unsupervised learning of data structures via self-organizing maps as well as supervised learning of a generative type classifier based on a graphical model description will be demonstrated. As examples for the inference from sparse data I will show recent attempts how to use the power of the graphical model formalism for describing complex generative models with applications to creating a decision support system for early tsunami warning and learning ground motion models.