There are no translations available.

V okviru biostatističnega centra bo **v ponedeljek, 7. junija 2010, ob 13. uri predavanje na IBMI** . Predavala bo **Giovanna Menardi** z univerze v Padovi.

Statistical issues emerging in training and evaluating classification models in presence of rare events

The problem of modeling binary responses by using cross sectional data has found a number of satisfying solutions extending throughout both parametric and nonparametric methods. Examples are traditional classification models like logistic regression, discriminant analysis, classification trees or procedures at the forefront as neural networks or combinations of classifiers (bagging, boosting, random forests). These models are based on the implicit assumption that the distribution of the responses is well balanced over the sample. However, there exist many real situations where it is a priori known that one of the two responses (usually the most interesting for the analysis) is rare. This class imbalance occurs in several domains as for example finance (detection of defaulter credit applicants), epidemiology (diagnosis of rare diseases), social sciences (analysis of anomalous behaviors), computer sciences (identification of some features of interest in image data). The class imbalance heavily afiects both the model estimation and the evaluation of its accuracy. Classification methods are in fact conceived to estimate the model that best fits the data according to some criterion of global accuracy. When data are unbalanced the model tends to focus on the prevalent class and ignore the rare events (Japkowicz, and Stephen, 2002). Moreover, when evaluating the quality of the classification, the same measures of global accuracy may lead to misleading results or even if alternative error measures are used, the scarcity of data conducts to high variance estimates of the error rate, especially for the rare class. In this work an unified and systematic framework for dealing with both the problems is proposed, based on a smoothed bootstrap form of re-sampling from data. The proposed technique includes some of the existing solutions as a special case, it is supported by a theoretical framework and reduces the risk of model overfitting. The presented talk is based on joint work with prof. Nicola Torelli from University of Trieste.