Spectroscopic Data in *R* and Validation of Soft Classifiers: Classifying Cells and Tissues by Raman Spectroscopy

Claudia Beleites[1,2,⋆], **Christoph Krafft**[2], **Jürgen Popp**[2,3], **and Valter Sergo**[1]

1. CENMAT and Dept. of Industrial and Information Engineering, University of Trieste, Trieste/Italy
2. Institute of Photonic Technology, Jena/Germany
3. Institute of Physical Chemistry and Abbe Center of Photonics, University Jena/Germany

⋆Contact author: cbeleites@units.it

Keywords: spectroscopy, soft classification, validation, brain tumour diagnosis

Medical diagnosis of cells and tissues is an important aim in biospectroscopy. The data analytical task involved frequently is classification. Classification traditionally assumes both reference and prediction to be *hard*, i. e. stating exactly one of the defined classes. In reality, the reference diagnoses may suffer from substantial uncertainty, or the sample can comprise a mixture of the underlying classes, e.g. if sample heterogeneity is not resolved or if the sample is actually undergoing a transition from one class to another (e. g. rather continuous de-differentiation of tumour tissues). Such samples may be labelled with *partial* or *soft* class memberships.

Many classification methods produce soft output, e. g. posterior probabilities. Methods like logistic regression can also use soft training data. Yet, for medical diagnostic applications it is even more important to include soft samples into the model validation. Excluding ambiguous samples means retaining only clear (i. e. easy) cases. Such a test set is not representative of the original unfiltered population, and creates a risk of obtaining overly optimistic estimates of the model performance.

With **softclassval** (softclassval.r-forge.r-project.org), we introduce a framework to calculate commonly used classifier performance measures like sensitivity and specificity also for samples with soft reference and prediction. Briefly, if the soft class labels are interpreted as uncertainty, best and worst case as well as expected performance are obtained via the weak, strong and product conjunction (and-operators, see e. g. Gottwald, 2010). For the mixture interpretation, weighted versions of well-known regression performance measures like mean absolute and root mean squared errors are derived.

As real world example, we classify 37 015 Raman (thereof 55 % soft) spectra of 80 brain tumour patients into "normal", "low grade", and "high grade" tissue morphologies in order to delineate excision borders during surgical treatment of the tumours. Thus, borderline cases are our actual target samples. We demonstrate spectroscopy-related functionality supplied by **hyperSpec** (hyperspec.r-forge.r-project.org) and its conjoint use with other packages.

Financial support by the Associazione per i Bambini Chirurgici del Burlo (IRCCS Burlo Garofolo Trieste) and of the European Union via the Europäischer Fonds für Regionale Entwicklung (EFRE) and the "Thüringer Ministerium für Bildung, Wissenschaft und Kultur" (Project: B714-07037) is highly acknowledged.

References

Gottwald, S. (2010). Many-valued logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2010 ed.).