Mutual Information: an Adequate Tool for Feature Selection

Benoît Frénay BENOIT.FRENAY@UCLOUVAIN.BE
Gauthier Doquire GAUTHIER.DOQUIRE@UCLOUVAIN.BE
Michel Verleysen MICHEL.VERLEYSEN@UCLOUVAIN.BE
Machine Learning Group - ICTEAM, Université catholique de Louvain, Place du Levant 3, B-1348 Louvain-la-Neuve

Keywords: mutual information, probability of misclassification, Hellman-Raviv and Fano bounds, MSE, MAE

Because of the curse of dimensionality (Bellman, 1961), it is often necessary to reduce the dimensionality of data before learning. For example, micro-array datasets contain only a few tens of instances with thousands of features. Such preprocessing can be achieved by feature selection.

In the field of machine learning, mutual information (MI) has been widely used as a multivariate criterion of non-linear feature relevance (Kojadinovic, 2005; Rossi et al., 2006; Doquire & Verleysen, 2011). Indeed, it is well-known in information theory that $I(X;Y)$ measures the reduction of uncertainty about a target Y when a set of features X are observed. However, other criteria are commonly used for classification and regression to assess the quality of models, like e.g. accuracy or mean square error (MSE). This presentation reviews several works (Frénay et al., 2012b; Frénay et al., 2013; Frénay et al., 2012a; Doquire et al., 2013) which address the relationships between MI and these criteria in a feature selection context.

Bounds exist for classification between MI and the probability of error, like e.g. the Hellman-Raviv inequality and the Fano inequalities. These bounds are sometimes used in the literature to justify the use of MI, claiming that selecting the feature subset which maximises MI corresponds to minimising the probability of error. However, this is not necessarily true (Frénay et al., 2012b; Frénay et al., 2013) and it is easy to design counterexamples where a suboptimal feature subset has an higher MI than the optimal feature subset. Hopefully, such failures have limited impact in practice (Frénay et al., 2013; Doquire et al., 2013).

There exists a deterministic link for regression between MSE and MI, as well as the mean absolute error (MAE), when the estimation error has a Gaussian, Laplacian or uniform distribution (Frénay et al., 2012a). In these realistic cases, MI can be safely used to perform feature selection. It is also possible to design counterexamples for regression, like e.g. when the estimation error has a Student distribution with a variable number of degrees of freedom, but the impact of such failures remains limited in practice.

In conclusion, this presentation provides both theoretical and empirical evidences that MI is not a perfect feature selection criterion in all situations, but is still a valuable criterion for feature selection, which is supported by the large number of successful applications in the literature.

Acknowledgments

Gauthier Doquire is funded by a Belgian F.R.I.A. grant.

References

Bellman, R. E. (1961). *Adaptive control processes - a guided tour*. Princeton University Press.

Doquire, G., Frénay, B., & Verleysen, M. (2013). Risk estimation and feature selection. *Proc. ESANN*.

Doquire, G., & Verleysen, M. (2011). Feature selection with mutual information for uncertain data. In *Data warehousing and knowledge discovery*, vol. 6862 of *Lecture Notes in Computer Science*, 330–341.

Frénay, B., Doquire, G., & Verleysen, M. (2012a). Is mutual information adequate for feature selection in regression ? Neural Networks, Submitted.

Frénay, B., Doquire, G., & Verleysen, M. (2012b). On the potential inadequacy of mutual information for feature selection. *Proc. ESANN* (pp. 501–506).

Frénay, B., Doquire, G., & Verleysen, M. (2013). Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. *Neurocomputing, Accepted for publication*.

Kojadinovic, I. (2005). On the use of mutual information in data analysis : an overview. *Proc. ASMDA*, 738–747.

Rossi, F., Lendasse, A., Francois, D., Wertz, V., & Verleysen, M. (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometr. Intell. Lab.*, *80*, 215–226.