Information Theory and Classification Error in Probabilistic Classifiers

Aritz Pérez, Iñaki Inza, Pedro Larrañaga

Department of Computer Science and Artificial Intelligence
University of the Basque Country - P.O. Box 649
20018 Donostia - San Sebastián, Spain

ISG - Intelligent Systems Group
http://www.sc.ehu.es/isg
{aritz,inza,ccplamup}@si.ehu.es

1 Abstract

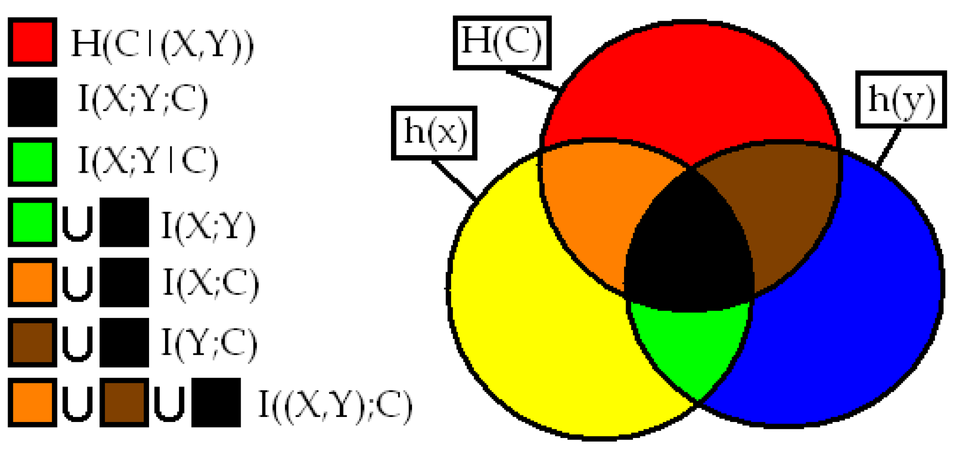
This work shows, using continuous real-world data [4] and artificial bivariate domains, the relation that seems to exist between the mutual information $I(\boldsymbol{X};C)$ [1] and the expected classification error ϵ_M. Besides, it shows that maximizing $I(\boldsymbol{X};C)$ is equivalent to maximize the conditional log likelihood $CLL(M|D)$ [3].

2 Introduction

- Classification expected error

$$\epsilon_M = \sum_{c=1}^{r} \int p(c) f(\boldsymbol{x}|c)(1 - p_M(c|\boldsymbol{x}))d\boldsymbol{x}$$

- Multivariate and univariate models and errors
  - $p_{mul}(c|\boldsymbol{x}) = p(c|\boldsymbol{x}) \propto p(c)\prod_{i=1}^{n} p(x_i|\pi_i, c) \rightarrow \epsilon_{mul}$
  - $p_{uni}(c|\boldsymbol{x}) \propto p(c)\prod_{i=1}^{n} p(x_i|c) \rightarrow \epsilon_{uni}$
  - $\epsilon_{dif} = \epsilon_{uni} - \epsilon_{mul}$
- Information theory (*IT*) based measures [1].

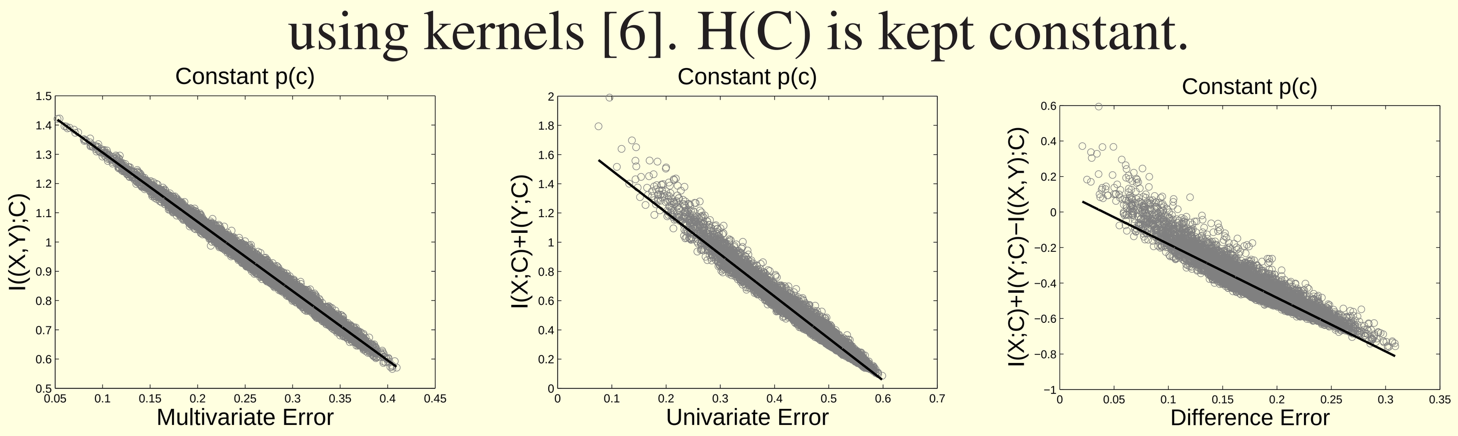


Relation between different IT based measures. Each region specifies a part of the uncertainty that surrounds the variables. Information theory measures are estimated using kernel based densities [6].

- Questions/motivations:
  - What kind of relation exists between the uncertainty that surrounds the class variable $H(C|\boldsymbol{X})$ and the classification errors ϵ_{uni} and ϵ_{mul}?
  - When is more advisable to use $p_{mul}(c|(x,y))$ instead of $p_{uni}(c|(x,y))$ for classification?
  - How are related the information theory based measures and the $CLL(M|D)$ [3]?

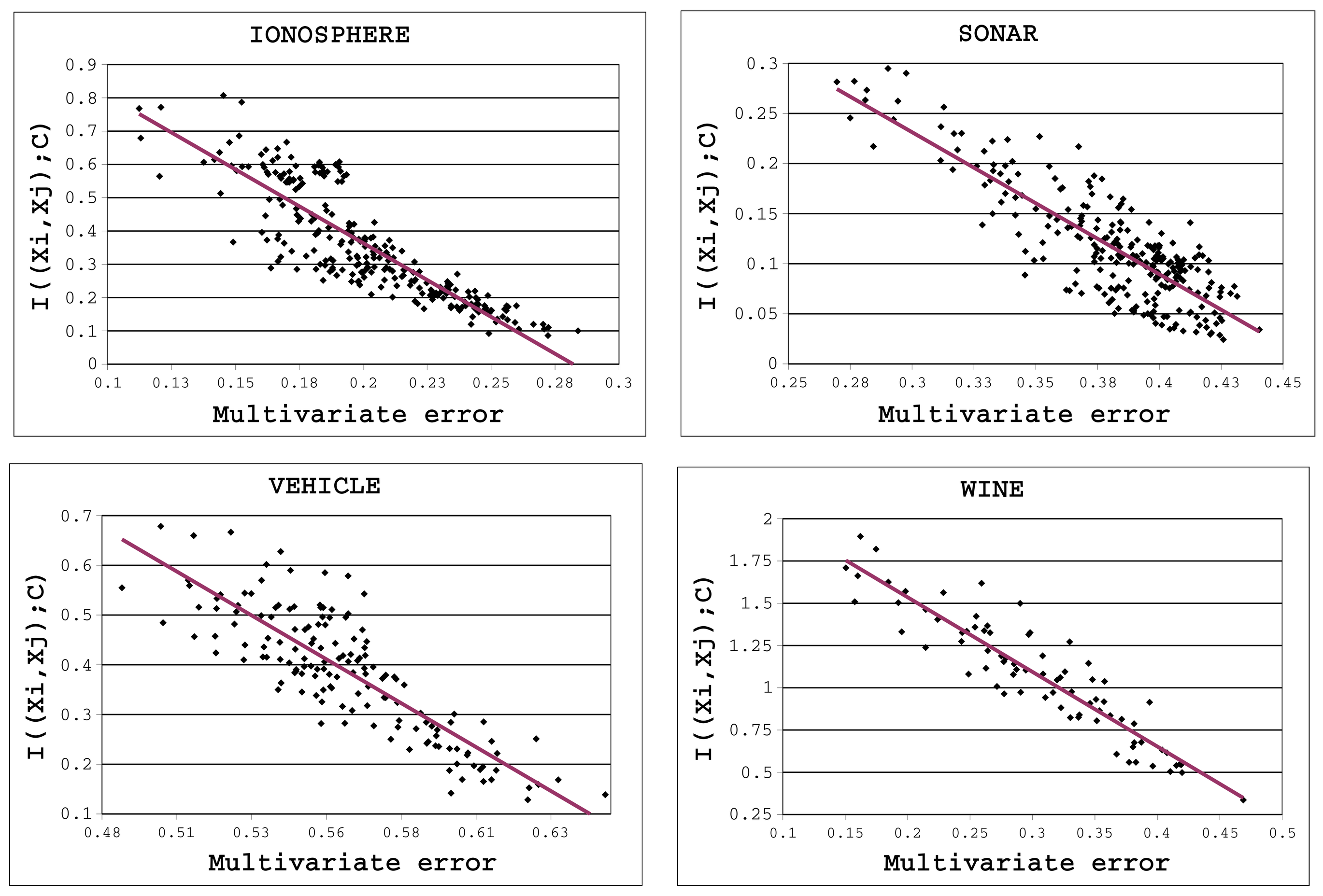
3 Artificial data

10000 artificial 2D-domains with arbitrary density shapes modelled using kernels [6]. H(C) is kept constant.



ϵ_{mul}, ϵ_{uni} and ϵ_{dif} versus IT based measures.

4 Real-world data



ϵ_{mul} versus $I((Xi,Xj);C)$ in four data sets of the UCI repository [4].The variables have been normalized (same variance).

5 $CLL(M|D)$ and IT

- $CLL(M|D)$ [3] is a more relevant score than $LL(M|D)$ for classification purposes [2, 3].
- Conditional log likelihood $CLL(M|D)$ for $p_M(c|\boldsymbol{x})$ can be written as:

$$\begin{aligned} CLL(M|D) &= \sum_{\boldsymbol{x},c} p_M(c|\boldsymbol{x}) = -N^{-1} H_{\hat{p}(\boldsymbol{x},c)}(C|\boldsymbol{X}) \\ &\propto -H(C) + I(\boldsymbol{X};C) \\ &= -H(C) + \sum_{1=1}^{n} I(X_i;C) - \sum_{1=1}^{n} I(X_i;\Pi_i;C) \end{aligned}$$

- Maximize the $CLL(M|D)$ is equivalent to maximize $I(\boldsymbol{X};C)$. Besides, when all predictors are included in the model, maximize $I(\boldsymbol{X};C)$ is equivalent to minimize $\sum_{1=1}^{n} I(X_i;\Pi_i;C)$.
- $I_{mul}((X,Y);C) = I(X;C) + I(Y;C) - I(X;Y;C)$; $I_{uni}((X,Y);C) = I(X;C) + I(Y;C)$.

6 Conclusions

- Bivariate models:
  - $I_M((X,Y);C)$ is directly proportional to the $CLL(M|D)$ and it seems to be inversely proportional to the error ϵ_{mul}.
  - $I(X;Y;C)$ seems to be inversely proportional to the error ϵ_{dif}. Therefore $I(X;Y;C)$ can be used in order to decide when is advisable to model the correlation between two variables.
- n-variate models: $CLL(M|D)$ is directly proportional to $I(\boldsymbol{X};C)$. Maximizing $CLL(M|D)$ is equivalent to minimize $\sum_{1=1}^{n} I(X_i;\Pi_i;C)$). $I(X_i;X_j;C)$ is known as explaining away residual (EAR) and is used in order to learn Bayesian network structures in a discriminative way [5].

[1] Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley and Sons (1991)
[2] Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning. (1997) 29:131–163
[3] Jebara, T.: Machine Learning: Discriminative and Generative. Kluwer Academic Publishers. (2004)
[4] Murphy, P.M., Aha, D.W.: UCI repository of machine learning databases. University of California at Irvine. http://www.ics.uci.edu/∼mlearn. (1995)
[5] Pernkopf, F., Bilmes, J.: Discriminative versus generative parameter and structure learning of Bayesian network classifiers. Proceedings of the 22nd International Conference in Machine Learning. (2005)
[6] Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall: London (1986)

DS/06 - Discovery Science 2006. Barcelona - Spain