Mutual Information

Alan Chau, Bobby He & Lorenzo Pacchiardi

November 12, 2018

Abstract

Quantifying dependencies between random variables is a key task in Applied Statistics. In this report we will examine the usability of Mutual information estimation and related methods to quantify dependencies.

1 Introduction

When presented with a new data-set one of the first data exploration questions that must be addressed is the question of dependency between the different covariates of the data. A classic approach to this question would be to estimate pairwise Pearson corrleation coefficients, R^2, but by its very definition this measure is only able to quantify linear dependencies. For more complex relationships, the information theoretic concept of Mutual Information has received a lot of attention in the literature recently. Our report will first provide a background on Mutual Information and the recently introduced Maximum Information Criterion [4], before a simulation study on the strengths and weaknesses of Mutual Information related estimates.

2 Background

2.1 Mutual Information

Let X,Y be two random variables with respective supports $\mathcal{X}$, $\mathcal{Y}$, marginal densities p_X, p_Y and joint density $p_{X,Y}$ with respect to the Lebesgue measure. Then, the Mutual Information between X and Y is defined to be:

$$I(X,Y) = \mathrm{KL}(P_{X,Y}||P_X \otimes P_Y) \tag{1}$$

$$= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_{X,Y}(x,y)\log\big(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\big)dxdy \tag{2}$$

where KL is the Kullback-Leibler divergence.

From this definition, it is easy to see that Mutual Information has an alternative characterisation in terms of (differential) Shannon Entropy, H, under regularity conditions:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

These characterisations give us several appealing properties of Mutual Information as a measure of dependence:

- $I(X,Y) = 0$ if and only if X & Y are independent random variables.
- $I(X,Y) = I(f(X), g(Y))$ for any smooth bijective functions f and g that have non-singular Jacobians.
- $I(X,Y)$ can be interpreted as the reduction in uncertainty in X if Y is known. If we choose logarithmic base 2 then $I(X,Y)$ is measured in units of bits.

A further characterisation of $I(X,Y)$ that will be useful to know utilises the Donsker-Varadhan representation of the KL divergence:

$$I(X,Y) = \sup_f \mathbb{E}_{\mathbb{P}_{X,Y}}\big[f\big] - \log\big(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Y}\big[e^f\big]\big)$$

where the supremum is taken over all functions $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ for which the expectations are finite.

3 Dependency measure estimates

Unfortunately, in almost all situations one can only estimate a desired dependency measure using observed data $\{(X_i, Y_i)\}_{i=1}^n$. This section describes some methods that have been suggested in the literature for this problem.

3.1 Maximum Information Criterion

The Maximum Information Criterion (MIC) introduced in [4] is designed to improve on the interpretability of dependency measures between one-dimensional random variables. The method works on the idea that a grid drawn on the scatterplot of the data should partition the observed points in such a way that describes any potential dependency. For various grid sizes (x, y), the authors suggest selecting grids in order to maximise the Mutual Information, $I_{x,y}$ of the distribution on the grid blocks such that each block has mass proportional to the number of observed points lying within it. The authors normalise the Mutual Information scores using the fact that $I_{x,y} \leq \min(log(x), log(y))$, which can be easily deduced from Shannon Entropy characterisation of Mutual Information. The MIC is then calculated to be the maximum over all normalised $I_{x,y}$ values for a range of (x, y) values satisfying $xy \leq B(n)$, for a value $B(n)$ to be chosen. NOT FINISHED.

3.2 Mutual Information Estimation

The most popular method for Mutual Information estimation originated in [3] and uses the well-known Kozachenko-Leonenko estimate for Shannon entropy, H, which relies on k-nearest neighbour methods. Recent work [2] has proven that, under regularity conditions and in dimensions $d \leq 3$, the Kozachenko-Leonenko in fact enjoys a Central Limit type result targeting H. For higher dimensions, a non-trivial bias term typically precludes its efficiency.

3.3 Mutual Information Neural Estimation

The search for a Mutual Information estimator that transcends the curse of dimensionality brings us to the area of deep learning. The estimator introduced in [1], which we shall refer to as MINE, uses the Donsker-Varadhan characterisation of $I(X < Y)$. More specifically, MINE uses the universal approximation theorem of neural networks to define:

$$\hat{I(X, Y)}_n = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{X,Y}}\left[f_\theta\right] - \log\left(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Y}\left[e^{f_\theta}\right]\right)$$

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation. 2018.

[2] Thomas B Berrett, Richard J Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *arXiv preprint arXiv:1606.00304*, 2016.

[3] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69 6 Pt 2:066138, 2004.

[4] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011. doi: 10.1126/science.1205438.