Generating Missing At Random (MAR) data in Images; a Convolutional Approach

TU/e supervisor: dr. W. Duivesteijn
Additional supervision by: R. M. Schouten, TU/e
Additional supervision by: dr. V. Menkovski, TU/e

Introduction

Real-life data often has missing values. If you want to handle missing values appropriately, it helps to understand by what mechanism the data is missing. Traditional literature distinguishes three mechanisms, which can be informally summarized as:

Missing Completely At Random (MCAR): the probability that data is missing does not depend on any values in the dataset (e.g., for any cell in the data table, there is a fixed probability that its value is missing);

Missing At Random (MAR): the probability that data is missing depends on values in other columns of the dataset (e.g., in a fictional society where having big feet is considered shameful, taller people are less likely to report their shoe size; missingness of shoe size values correlates with higher values in the length column);

Missing Not At Random (MNAR): the probability that data is missing depends on the value of the missing data itself (e.g., people who earn a lot of money are less likely to report their income; missing values are likely higher values).

Arguably, MAR is the most interesting mechanism of the three. Under MCAR or MNAR, no information remains in the dataset on why a certain value is missing. Under MAR, this information is present in other columns, so more intelligent conclusions can be drawn.

Modern data mining methods can handle data beyond traditional tables. The current state-of-the-art classifiers are often evaluated by how well they can classify images. One would also like to assess the degree to which these classifiers are robust w.r.t. missing data. However, not all missingness mechanisms can straightforwardly be applied to image data in its traditional pixel-value form:

MCAR data works: one can, for instance, simply apply a fixed probability that any given pixel is missing;

MNAR data also works: one can, for instance, make the probability that a given pixel is missing depend on the intensity of the pixel itself;

MAR data, however, does not work: if one were to make the probability that a given pixel is missing depend on the intensity of some other pixel(s), one could argue that this may be MNAR, since pixel values are not necessarily independent. Hence, information on the intensity of the given pixel may leak into the values of the other pixel(s). In fact, it is entirely unclear how one could make a hard distinction between unobserved and observed information in this case, so the boundary between MAR and MNAR becomes blurry.

The point is that the information we may or may not leak, resides in a higher-level concept than pixel intensities: it resides in a line, or a curve, or a texture. Hence, to properly demarcate MAR missingness in image data, we must look at image data representations that concern such higher-level concepts. Hence, it stands to reason to find MAR missingness in a convolutional layer of a neural network that has been fed this image.

Current Status

Nothing more exists about this project than the idea described in this document. We're in the pioneering stage; boldly go where noone has gone before!

Requirements

Missing data mechanisms have a strong presence in the statistics community; you will need some affinity with statistics while doing the literature review for this project. When done well, we expect this project to lead to a publication at a top-level data mining or machine learning conference.