

data imputation

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Olioumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Synonyms

imputation, missing-value imputation

Abstract

Data imputation replaces the missing feature or label values of data points by estimates computed from the values that were recorded. It is one response to missing data, alongside discarding the incomplete data points. The filling becomes a machine learning (ML) problem once it is read as a prediction task: the quantity to estimate is treated as a label and the values that survive around it as a feature vector, so the intact data points form a training set on which a hypothesis is learned. For a photograph with lost pixels, each pixel is a data point whose feature vector holds the colors of its neighbors and whose label is its own color. Where a lost value sits among other lost values, the hypothesis is applied repeatedly until the estimates stop changing, which is a fixed point of one sweep. An imputed value is an estimate and not a measurement, so a method trained on imputed data points reports a training error that flatters it.

Definition

An aerial photograph of the vineyards at Rossatz on the Danube comes back from the survey with a rectangular patch of its pixels lost, because the sensor dropped them. The photograph is still wanted, so those pixels need values. Data imputation is that filling in: it replaces the missing feature or label values of data points by estimates computed from the values that were recorded (Abayomi et al., 2008). It is one response to missing data, alongside discarding the incomplete data points or using a method that tolerates the gaps.

The filling becomes a machine learning (ML) problem once it is read as a prediction task (Abayomi et al., 2008). For the photograph, each pixel is a data point. Its feature vector holds the red, green and blue values of the eight pixels surrounding it, which is 24 numbers, and its label is its own triple of red, green and blue values (Fig. 1). The intact pixels then form a training set: each one supplies both a feature vector and the label that belongs with it, so a hypothesis can be learned from them by empirical risk minimization (ERM). Applying that hypothesis to a lost pixel yields the estimate that fills it.

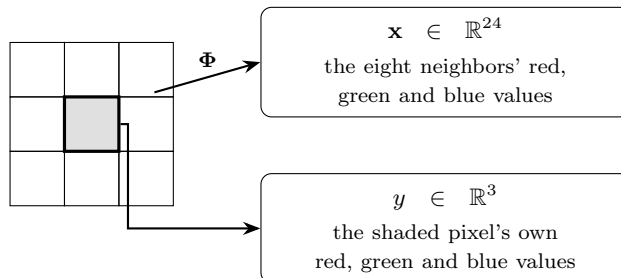


Figure 1: One pixel of the photograph as a data point. Its feature vector collects the red, green and blue values of the eight surrounding pixels, and its label is its own triple of those values. An intact pixel supplies both, and so belongs in the training set

A lost pixel usually has lost pixels among its neighbors, so its feature vector is not available and the hypothesis cannot be applied once and be done. The lost pixels instead start at the average color of the training set and are re-predicted from their current neighbors, sweep after sweep. Writing $\mathbf{u}$ for the values currently held by the lost pixels, one sweep applies an operator T to them,

$$\mathbf{u}^{(t+1)} = T(\mathbf{u}^{(t)}),$$

and the filling stops at a fixed point of T , where another sweep leaves the values unchanged. The experiment `pythondemos/dataimputation.py` reaches one after 413 sweeps on a view of 64×96 pixels with 6% of them lost.

What imputation can and cannot recover is visible in the numbers of that experiment (Fig. 2). A linear regression hypothesis learned from the intact pixels halves the average squared error loss on the lost ones against the baseline that fills them all with the average color. The gain is not spread evenly: inside a region of nearly constant color the error is a quarter of that baseline, while on the pixels where the river meets the meadow it is as large. The filled river runs through the gap, but its banks come out blurred.

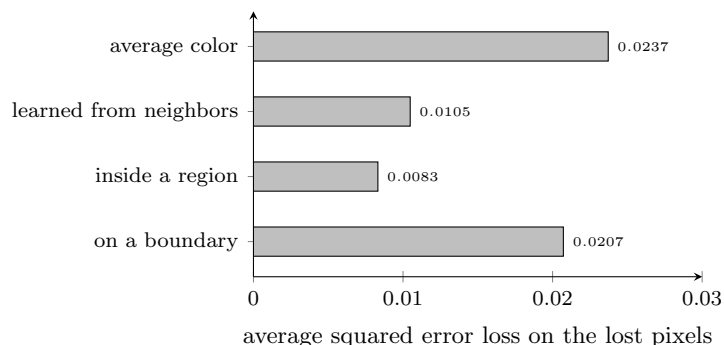


Figure 2: Average squared error loss on the lost pixels. The first two bars compare filling them all with the average color against the linear regression hypothesis learned from the intact pixels. The last two split that learned result by where the lost pixel sits, inside a region of nearly constant color or on the boundary between two regions. Data generated by `pythondemos/dataimputation.py`

An imputed value is an estimate and not a measurement, and a method trained on imputed data points cannot tell the two apart. The training error it reports therefore flatters it, since part of the training set was produced by a hypothesis rather than observed. Recording which values were imputed, and keeping them out of the test set, is what preserves an honest estimate of the risk.

Cross References

missing data, data point, feature, label, feature vector, training set, ERM, linear regression, fixed point, baseline, test set

References

1. Abayomi et al. (2008). Diagnostics for Multivariate Imputations. J. Roy. Statist. Soc.: Ser. C (Appl. Statist.), 57(3), pp. 273–291.