Chapter 2

PROBABILISTIC AND STATISTICAL MODELS FOR OUTLIER DETECTION

"With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk." – John von Neumann

1. Introduction

The oldest methods for outlier detection are rooted in probabilistic and statistical models, and date back to the nineteenth century [149]. The earliest methods were proposed well before the advent and popularization of computer technology. Therefore, these methods were designed without much focus on practical issues such as data representation or computational efficiency. Nevertheless, the underlying mathematical models are extremely useful, and have eventually been adapted to a variety of computational scenarios.

A popular form of statistical modeling in outlier analysis is that of detecting *extreme univariate values*. In such cases, it is desirable to determine data values at the tails of a univariate distribution, along with a corresponding level of statistical significance. This would seem a rather restrictive case, since most multidimensional outliers do not correspond to extremes in data values. Rather, outliers are typically defined by the *relative* positions of the data values with respect to each other. While extreme univariate values correspond to a very specific kind of outliers, they have numerous applications beyond the univariate case. This is because virtually all outlier detection algorithms perform some kind of numerical scoring, in order to measure the anomalousness of data points. In some cases, the scores may come with a confidence value or probability, though this capability is often not directly built

C.C. Aggarwal, *Outlier Analysis*, DOI 10.1007/978-1-4614-6396-2_2,

into outlier analysis algorithms. Therefore, the final step in all these algorithms is to determine the extreme values from these scores. The determination of statistically extreme values helps in the conversion of outlier scores into binary labels.

Some examples of outlier scoring mechanisms, which are returned by different classes of algorithms, are as follows:

- In probabilistic modeling, the likelihood fit of a data point to the model is the outlier score.

- In proximity-based modeling, the k-nearest neighbor distance, distance to closest cluster centroids, or local density value is the outlier score.

- In linear modeling, the residual distance of a data point to a lower-dimensional representation of the data is the outlier score.

- In temporal modeling, a function of the distance from previous data points (or the deviation from a forecasted value) is used to create the outlier score.

Thus, even when extreme value modeling cannot be performed on the original data, the ability to determine the extreme values effectively from a set of outlier scores forms the cornerstone of all outlier detection algorithms as a final step. Some recent work has been devoted exclusively to the problem of determining such extreme values [179] from outlier scores, by converting these scores into probabilities. Therefore, the issue of extreme value modeling will be studied extensively in this chapter. Extreme value modeling can also be easily extended to multivariate data, and will be discussed in this chapter.

It is also possible to use probabilistic modeling for finding general outliers beyond extreme values. Mixture models can be considered probabilistic versions of clustering algorithms, and can therefore be used for outlier analysis. A significant advantage of these methods is that they are fairly easy to generalize to different data formats, or even heterogenous data attributes, once a generative model for the data has been defined. Most probabilistic models assume a particular form to the underlying distribution, according to which the data is modeled. Subsequently, the parameters of this model are learned, typically with a maximum-likelihood estimation technique [135]. This model then becomes a *generative* model for the data, and the probability of a particular data point being generated can be computed from this model. Data points which have an unusually low probability of being generated from the model are returned as outliers.