

Brian Tusi
November 13, 2008
Comp 578 – Data Mining

Assignment 11
Chapter 10, 1, 8, 9

1. Compare and contrast the different techniques for anomaly detection that were presented in Section 10.1.2. In particular, try to identify circumstances in which the definitions of anomalies used in different techniques might be equivalent or situations in which one might make sense, but another would not. Be sure to consider different types of data.

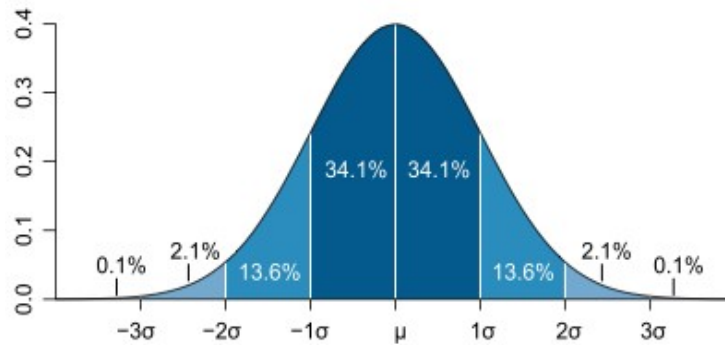
Boiling things down, I'd venture to say that there's two types of techniques displayed, grouped as model-based and proximity-/density-based. This is due to how each of these techniques works – model-based attempts to fit a known model to a set of data, and hopes it works. Proximity- and Density-based, however, don't require any pre-set notions of a model, and let the data define it for itself. Sure, the procedure is a bit different (comparing proximity to density), but the concept is the same – find the clusters, and label anything not in the cluster an anomaly.

There are various applications where one method is preferred over another. Since model-based techniques work best when a known model already exists, it's useful in scenarios such as in a census – for example, communities will generally have a close to 1:1 ratio of men:women (though not exactly), and deviations from this will be easy to spot because of the known model. In another example, it's well-known when high and low tide will be – deviations from this known formula will be considered anomalies.

Just as how model-based techniques are useful in some circumstances, proximity- and density-based techniques are preferred in other scenarios. For example, on an exam given by a professor, it's not very easy to predict how the results will turn out. However, once the results have been collected, these techniques are easy to apply. Plotting all of the scores on a line, for example, will show students with score anomalies – such as those who truly understand the material (or cheated), and those who are in danger of not passing the class. Cheating itself is another scenario where these techniques can be used; a student who normally gets a C+ or B- average on everything and suddenly gets an A on a large test may deserve a second look, even if just precautionary. Such is the nature of anomalies; sometimes they're perfectly valid points that just don't fit expectations, and other times they're erroneous.

8. Many statistical tests for outliers were developed in an environment in which a few hundred observations was a large data set. We explore the limitations of such approaches.

a) For a set of 1M values, how likely are we to have outliers according to the test that says a value is an outlier if it is more than 3 standard deviations from the average? (Assume a normal distribution.)



The above chart is from Wikipedia (http://en.wikipedia.org/wiki/Image:Standard_deviation_diagram.svg), showing the standard deviations in a normal distribution. The chart supports the textbook's claims (p. 659) that there is a probability of 0.0027 of an object laying outside 3 deviations from the mean.

b) Does the approach that states an outlier is an object of unusually low probability need to be adjusted with large data sets? If so, how?

No, the approach does not appear to say that it requires adjustment for large data sets. This makes sense, as probability is expressed as a percentage; it does not define the number of items in a set, but how likely it is to be a part of that set. Sure, the actual quantity of items classified as outliers will go up, but so will the total number of objects too. Statistically speaking, the same ratio can be used regardless of the data set. (This may not be true for all data sets due to the nature of the set, but holds in a purely mathematical situation.)

9. The probability density of a point x with respect to a multivariate normal distribution having a mean μ and covariance matrix Σ is given by equation 10.8. Using the sample mean \bar{x} and covariance matrix S as estimates of the mean μ and covariance matrix Σ , respectively, show that the $\log \text{prob}(x)$ is equal to the Mahalanobis distance between a data point x and a sample mean \bar{x} plus a constant that does not depend on x .

Typing the exact equation for this could take quite some time. However, the concept is fairly straightforward.

Our goal is to show that $\ln(\text{prob}(x)) = \text{mahalanobis}(x, \bar{x}) + T$.

Therefore, viewing equation 10.8, we can see something obvious: except for a constant, the exponent of e can be brought down and reduced out of the equation, as it is $-1/2$ times the Mahalanobis distance. This leaves us with a value which, importantly, does not contain x . Therefore, we can show that the log of $\text{prob}(x)$ is equal to the Mahalanobis distance between x and \bar{x} plus an additional constant.