Outliers detection as Optimization

Sandipan Dey,
University of Maryland, Baltimore County,
MD, USA
sandip2@umbc.edu

August 10, 2009

1 Assumptions

- Outliers are anomolous points in the sense that they are 'far away'(measured in terms of some distance metric) from all other points in the dataset, i.e., if X denotes the entire dataset and O the set of all outliers, then we have, $\frac{||x^k - x^i||}{||o^l - x^i||} \rightarrow 0$, whenever $o^l \in O, x^k, x^i \in X - O, k \neq i$.
- Number of outliers in the data is small (neglible) compared to the other (properly behaviouring) points in the data, i.e., $\frac{card|O|}{card|X|} \rightarrow 0$, or, $\lim_{n_X \rightarrow \infty} \frac{n_O}{n_X} = 0$.
- Outliers increase the dispersion of data (e.g. variance / deviation).

2 Approach

2.1 Approach - 1

Let $p(x)$ be the p.d.f. of X. Now, the entropy of X, i.e., $H(X)$ is defined by $H(X) = \int_{x \in X} p(x) \log \frac{1}{p(x)} = -\int_{x \in X} p(x) \log p(x)$, the discrete version being $H(X) = -\sum_{x \in X} p(x) \log p(x)$. The contribution of outliers in the entropy of data is very small, according to our second assumption $\left(\sum_{x \in O} p(x) \log p(x) \rightarrow 0 \right)$, although the converse is not true (typically in highly chaotic data like uniformly distributed data with high entropy there may be data points that do not contribute much to the entropy but still not outliers).
Let us assign weight w^i to each individual point x^i. Initially each w^i will be assigned uniformly to $\frac{1}{n_X}$.
Also, let **c** denote the centroid (measured by some central tendency that is not sensitive

to noise or extreme data points, e.g., **median**) of the dataset, then by the first assumption we have, $||o^l - c|| >> ||x^k - c||$, $o^l \in O$, $x^k \in X - O$.
To find outliers, we shall be interested to maximize $\sum_{x^i \in X} w^i ||x^i - c||$. Hence, our optimization problem of outlier detection can be formulated by the following minimization problem:

$$\min - \sum_{x^i \in X} w^i ||x^i - c||$$
$$\text{s.t.} \sum_{x^i \in X} w^i = 1$$

and combining with our second assumption it can be written as:

$$\min \underbrace{- \sum_{x^i \in X} w^i ||x^i - c||}_{1^{st}/3^{rd} \text{ assumption}} + \underbrace{\sum_{x^i \in X} w^i p(x^i)}_{2^{nd} \text{ assumption}}$$
$$\text{subject to} \sum_{x^i \in X} w^i = 1$$

Here $||.||$ denotes 1-norm or 2-norm. The solution weights will have nearly zero value for non-outlier points and high values (> 0.5) corresponding to outlier points. The solution space is convex (?).

2.2 Approach - 2

We can alternatively (predictively) model the outlier detection problem as classifical problem with some training data set containing some extreme outlier points and 2 predefined classes labelled by O (for outlier) and N (for non outlier) where they represent 2 non-intersecting half open spaces deonted by the sets O and $X - O$. Now we can use some robust techniques like SVM for classification of the dataset. Off course, we must be aware of overfitting while estimating the classfication function (e.g., support vectors). The following figure shows these disjoint sets:

2.3 Approach - 3

Similar to PCA, but change the problem into a minimization problem:

$$\min a_i \sum\nolimits_x a_i^T$$
$$\text{s.t. } a_i a_i^T = 1$$

$$\text{and } a_i \sum\nolimits_x a_j^T = 0, j = 1 \ldots i - 1$$

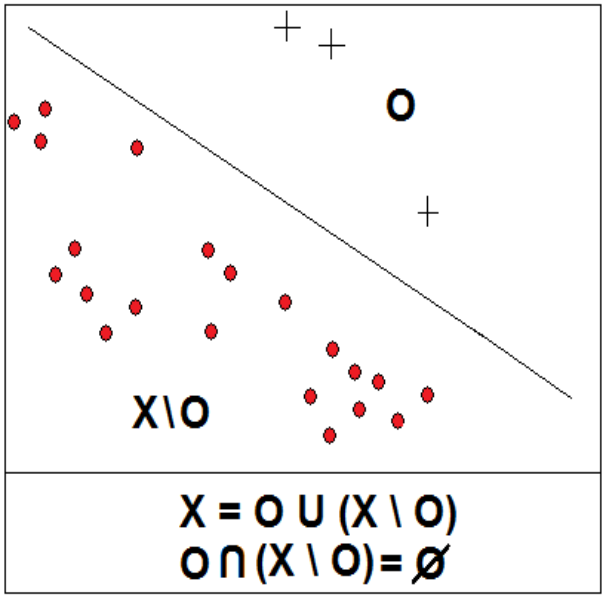

Figure 1: Classification in Outlier and Non-Outlier Classes by SVM