Chapter 2 - Probabilistic and Statistical Models for Outlier Detection

1 Introduction

Extreme value analysis is good for univariate data. Generative models are also popular.

2 Statistical Methods for Extreme-Value Analysis

The Markov Inequality says that, for nonnegative random variable X, we have for any constant α where $E[X] < \alpha$ that

$$P(X > \alpha) \leq E[X]/\alpha \tag{1}$$

The Chebyshev Inequality says that, for random variable X, we have for any constant α that

$$P(|X - E[X]| > \alpha) \leq Var(X)/\alpha^2 \tag{2}$$

The Chernoff Bound says that if X is the sum of N independent Bernoulli random variables (where the i^{th} variable has parameter p_i) then for any $\delta \in (0, 1)$ we have

$$P(X < (1 - \delta)E[X]) < \exp(-\frac{1}{2}E[X]\delta^2) \tag{3}$$

and for $\delta \in (0, 2e - 1)$ we have

$$P(X > (1 + \delta)E[X]) < \exp(-\frac{1}{4}E[X]\delta^2) \tag{4}$$

The Hoeffding Inequality says that if X is the sum of N independent random variables (where the i^{th} variable lies between $[l_i, u_i]$), that we have, for any $\theta > 0$

$$P(X - E[X] > \theta) \leq \exp(-\frac{2\theta^2}{\sum_{i=1}^{N}(u_i - l_i)^2}) \tag{5}$$

$$P(E[X] - X > \theta) \leq \exp(-\frac{2\theta^2}{\sum_{i=1}^{N}(u_i - l_i)^2}) \tag{6}$$

The Central Limit Theorem says the sum of N i.i.d random variables with mean μ and standard deviation σ converges to a normal distribution with mean $N\mu$ and standard deviation $\sigma\sqrt{N}$.

For univariate data, we can fit a Gaussian distribution and compute the Z-scores and PDF values. If the dataset is small (under 1000 examples), we fit a Student's t distribution where the degrees of freedom are $\nu = N - 1$.

We can also make a box-and-whiskers plot. The box is drawn between the first and third quartiles. The lower whisker is 1.5 times the interquartlie range (IQR) below the bottom of the box and the upper whisker is 1.5 times the IQR above the top of the box. Points outside the whiskers are plotted as points.

3 Extreme-Value Analysis in Multivariate Data

Extreme value analysis finds outliers at the boundaries of the data, not in the middle.

Depth based methods find a convex hull around the data and assign the points on the hull a depth of 1. We then remove those points and repeat to find points at depth 2. We do this r times and consider those points outliers. The convex hull algorithm scales exponentially with the dimensionality of the data.

In deviation based methods, we identify a subset of the points whose removal reduces dataset variance the most. There are hueristics for picking the subsets.

In angle based methods, we use the intuition that outlier points should be able to close the rest of the data by casting out two rays with a small angle. So, given a test point $\bar{X}$ and random points $\bar{Y}$ and $\bar{Z}$, we can compute the angle as

$$W\cos(\bar{Y} - \bar{X}, \bar{Z} - \bar{X}) = \frac{(\bar{Y} - \bar{X})^T(\bar{Z} - \bar{X})}{||\bar{Y} - \bar{X}||_2^2 ||\bar{Z} - \bar{X}||_2^2} \tag{7}$$

If we vary $\bar{Y}$ and $\bar{Z}$ we can compute the angle based outlier factor

$$ABOF(\bar{X}) = Var_{Y,Z \in D} W\cos(\bar{Y} - \bar{X}, \bar{Z} - \bar{X}) \tag{8}$$

A naive implementation takes $O(N^3)$ time, but we can speed this up by only considering the k nearest neighbors of $\bar{X}$ and ignoring it altogether if its first ABOF is large. Note that this method suffers from the curse of dimensionality because of the cosine.

We could also fit a Gaussian with mean μ and covariance Σ and compute the PDF (the term in the exponential is $-1/2$ times the squared Mahalanobis distance).

$$f(\bar{X}) = \frac{1}{\sqrt{\Sigma}(2\pi)^{d/2}} \exp(-\frac{1}{2}(\bar{X} - \mu)\Sigma^{-1}(\bar{X} - \mu)^T) \tag{9}$$

If Σ is not invertible, we can replace it with $\Sigma + \lambda I$ for small λ (this is regularization). This nice thing about the above Mahalanobis technique is that is it is parameter-free, accounts for variance in each dimension, and runs in linear time (to compute mean and covariance).

4 Probabilistic Mixture Modeling for Outlier Analysis

We can model the data with a mixture of k distributions. That is, the probability density of $\bar{X}_j$ is

$$f^{point}(\bar{X}_j | \mathcal{M}) = \sum_{i=1}^{k} \alpha_i f_i(\bar{X}_j) \tag{10}$$

where α_i and f_i are the mixing proportion and PDF of the i^{th} distribution, respectively. We can infer these parameters (call them Θ) by maximizing the log-likelihood of the data using the Expectation Maximization algorithm. This is an iterative algorithm where we first fix the parameters and compute $P(\bar{X}_j|\mathcal{G}_r, \Theta)$ for all (point, distribution) pairs $(\bar{X}_j, \mathcal{G}_r)$. This is the E-step. Then we fix the probabilities above and infer the parameters again. This is the M-step.

The E-step computes:

$$P(\mathcal{G}_r|\bar{X}_j, \Theta) = \frac{\alpha_r f^{r,\Theta}(\bar{X}_j)}{\sum_{i=1}^{k} \alpha_i f^{i,\Theta}(\bar{X}_j)} \tag{11}$$

The M-step computes (where the α_i expression uses Laplacing smoothing to push the estimate towards $1/k$):

$$\alpha_i = \frac{1 + \sum_{j=1}^{N} P(\mathcal{G}_r|\bar{X}_j, \Theta)}{k + N} \tag{12}$$

To estimate f_r, we treat $P(\mathcal{G}_r|\bar{X}_j, \Theta)$ as the weight of the point in component r and compute the maximum likelihood estimate of the model parameters.

Mixture models are a kind of stochastic clustering or soft clustering.

If you use a single component, you get a soft version of Principal Components Analysis (PCA). It becomes even more powerful when you combine it with kernel methods.

If you have a score for each point computed by another outlier detection algorithm, then you can use a mixture of an exponential distribution (for outlier points) and Gaussian distribution (for inlier points). Then you can compute the posterior probability that each point belongs to the outlier component. Thus, we have turned scores into probabilities.

5 Limitations of Probabilistic Modeling

Parametric methods like mixture models can fail to fit the data if it does not obey the assumed distributions. If you have too many components in your mixture, you can overfit. Note that the E-step and M-step each take $O(Nk)$ time. Mixture models are not easily interpretable.

6 Conclusions and Summary

Extreme value analysis is good for low dimensional data. Mixture models are powerful for fitting data or converting scores into probabilities.