Outlier Analysis -Charu Aggarwal

Chapter 1-2: *Probabilistic and Statistical Models for Outlier Detection*

CSE 537: Artificial Intelligence

Conor Kelton, Tim Barron

GROUP 10

Introduction: Outliers

- One definition:

  “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”

  -D. Hawkins. Identification of Outliers, Chapman and Hall, 1980.

- Also referred to as abnormalities, discordants, deviants, or anomalies
- Many applications
- How do we classify a data point as an outlier?

Introduction: Applications

- Intrusion Detection Systems
- Credit Card Fraud
- Interesting Sensor Events
- Medical Diagnosis
- Law Enforcement
- Earth Science

Introduction:
Extreme Univariate Value Analysis

- The process of defining outliers as points on tails of distributions
- Univariate models date back to the 19th century
- Most current datasets contain multiple dimensions, so Univariate Extreme Values are not as useful for anomaly detection in these cases
- However many multidimensional outlier detection algorithms return a single outlier score for each data point

Introduction:
Outlier Scores

- When data is multidimensional methods are used to get outlier scores for each data point (discussed later)

- Outlier scores can then be treated as a single variable

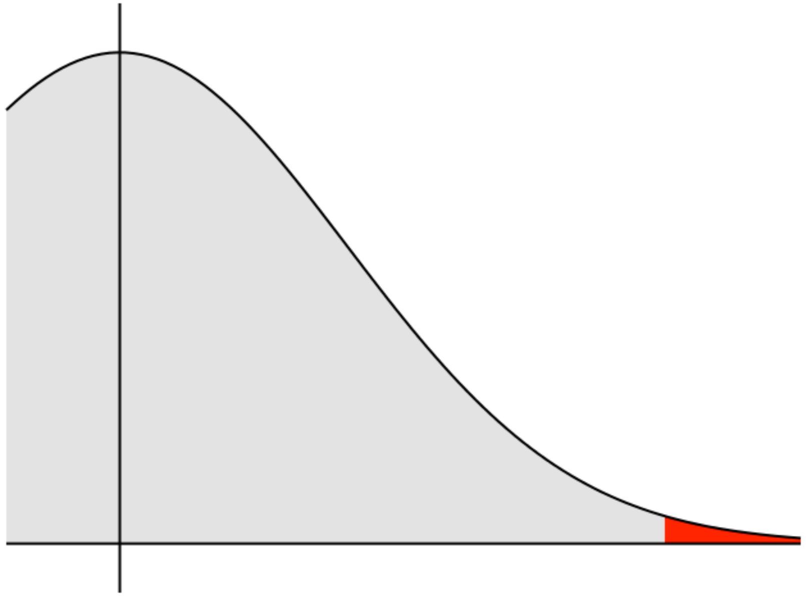
- Extreme Univariate Analysis can be performed on the outlier scores as a last step in the overall anomaly detection

Statistical Methods for EUVA: Probabilistic Tail Inequalities

- A *Statistical Tail* is the set of extreme values in a distribution (low likelihood)

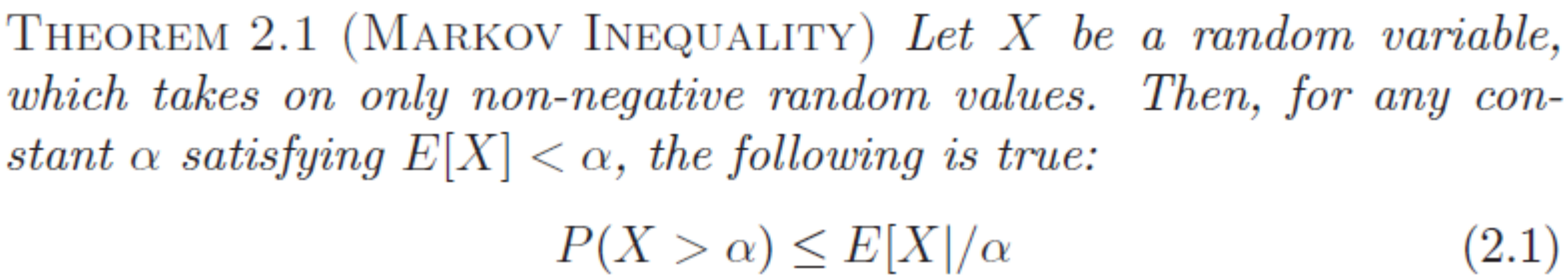
- Tail inequalities are used to determine whether points in the tails of a distribution are truly anomalous

- The following Tail Inequalities are covered:
  1. Markov Inequality
  2. Chebyshev Inequality
  3. Chernoff Bound
  4. Hoeffding Bound
  5. The Central Limit Theorem



Tail Inequalities: Markov Inequality

- The Markov inequality provides inference about the upper tails of any positive probability distribution



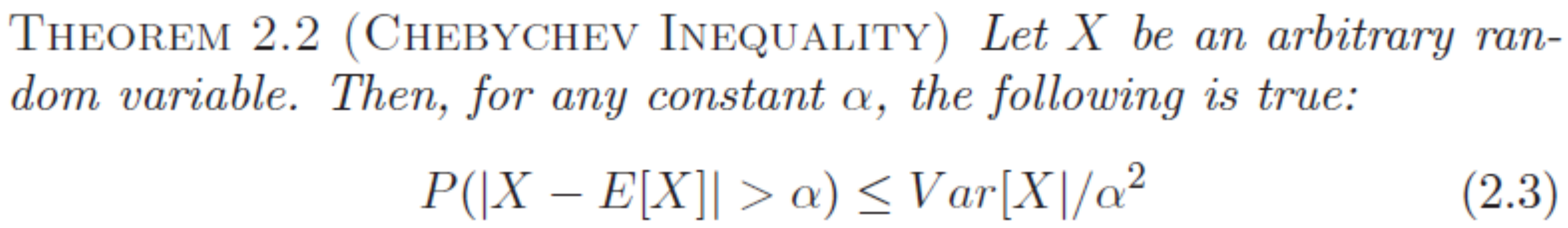
THEOREM 2.1 (MARKOV INEQUALITY) *Let X be a random variable, which takes on only non-negative random values. Then, for any constant α satisfying $E[X] < \alpha$, the following is true:*

$$P(X > \alpha) \leq E[X|/\alpha \tag{2.1}$$

- Useful for relating Expectations to probabilities, bounds the right tail based on the mean
- Also can be used to prove Chebyshev's Inequality

Tail Inequalities: Chebyshev's Inequality

- Gives insight as to both tails of any probability distribution based on mean and standard deviation, or variance



THEOREM 2.2 (CHEBYCHEV INEQUALITY) *Let X be an arbitrary random variable. Then, for any constant α, the following is true:*

$$P(|X - E[X]| > \alpha) \leq Var[X|/\alpha^2 \tag{2.3}$$

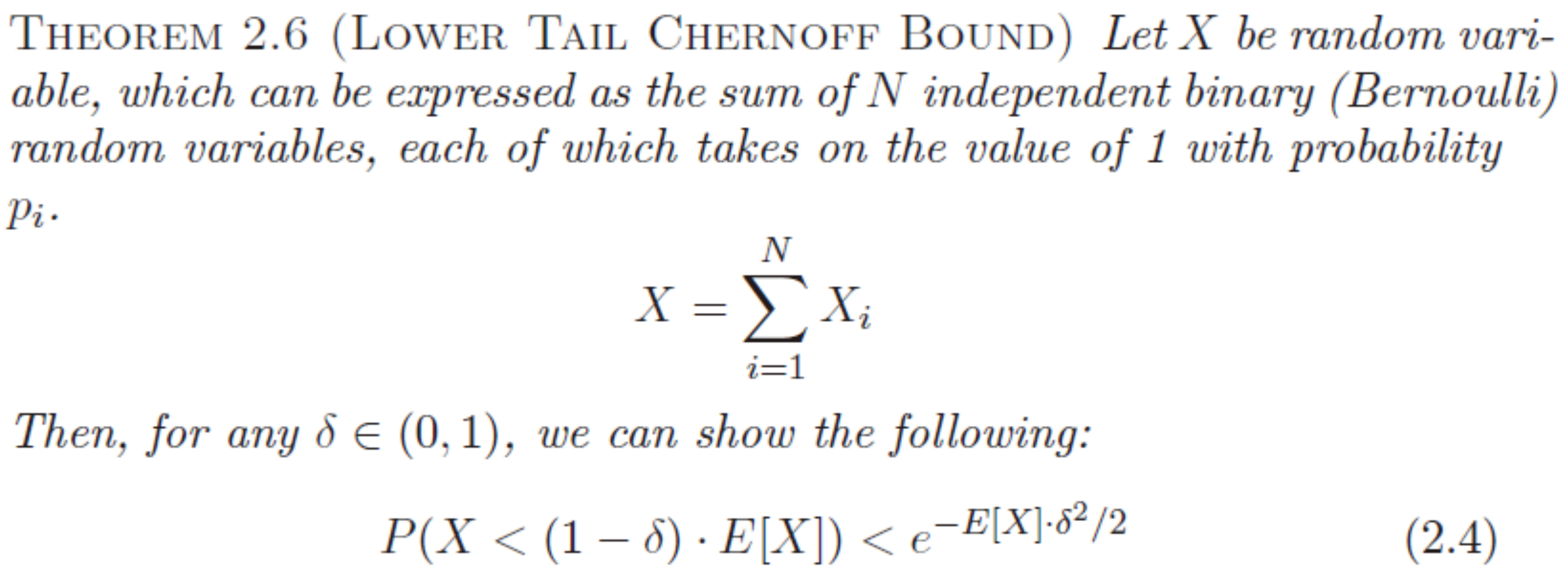
- Restated: $P(|X\ -\ \mu| > \alpha\sigma) \leq 1/\alpha^2$, or at most $(1/\alpha^2)$*100% lie in the tails of the distribution

- Similar in nature to (68-95-97.5) rule for a Gaussian

- Tighter bounds can usually be found as this inequality assumes nothing about the distribution, including shape and symmetry.

Tail Inequalities: Chernoff Bounds

- Gives bounds on a Random Variable that can be expressed as the sum of independent Bernoulli variables (non identical)



THEOREM 2.6 (LOWER TAIL CHERNOFF BOUND) *Let X be random variable, which can be expressed as the sum of N independent binary (Bernoulli) random variables, each of which takes on the value of 1 with probability p_i.*

$$X = \sum_{i=1}^{N} X_i$$

Then, for any $\delta \in (0,1)$, we can show the following:

$$P(X < (1 - \delta) \cdot E[X]) < e^{-E[X] \cdot \delta^2 / 2} \tag{2.4}$$

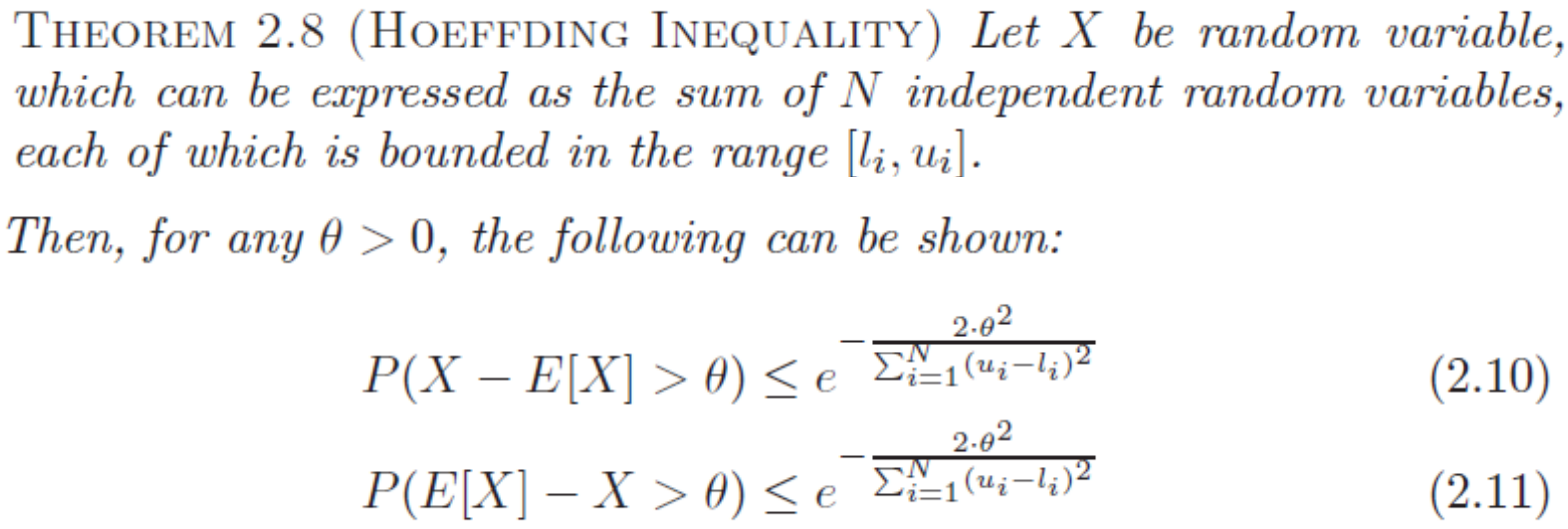
Tail Inequalities: Chernoff Bounds (cont'd)

- Gives tighter bounds (exponential decay) on Lower tail than inequalities above
- Upper tail has similar (also exponential) bound (for different δ values)
- Chernoff gives tighter tails (good for anomaly detection) but can be used much more infrequently (large assumptions)
- Example usage: Grocery Shoppers (Sum of i.d. Bernoulli)

Tail Inequalities:
Hoeffding Bounds

- Gives bounds on a distribution that can be expressed as the sum of any bounded independent Random Variables (also non identical)

- General case to the Chernoff Bound



THEOREM 2.8 (HOEFFDING INEQUALITY) *Let X be random variable, which can be expressed as the sum of N independent random variables, each of which is bounded in the range $[l_i, u_i]$.*

Then, for any $\theta > 0$, the following can be shown:

$$P(X - E[X] > \theta) \leq e^{-\frac{2\cdot\theta^2}{\sum_{i=1}^{N}(u_i - l_i)^2}} \tag{2.10}$$

$$P(E[X] - X > \theta) \leq e^{-\frac{2\cdot\theta^2}{\sum_{i=1}^{N}(u_i - l_i)^2}} \tag{2.11}$$

Tail Inequalities:
Hoeffding Bounds

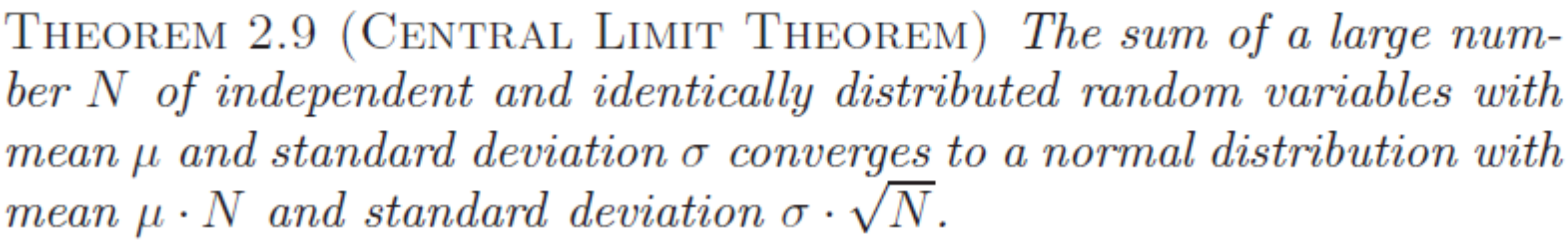
- Gives tighter bounds than Markov or Chebyshev, like Chernoff with different assumptions

- Lower tail has exact same bound

- Example usage: Sports Statistics (Sum of i.d. bounded)

Tail Inequalities:
Central Limit Theorem

- Gives an exact distribution for a Random Variable that can be expressed as a sufficiently large sum of any independent, identically distributed (i.i.d.) Random Variables



THEOREM 2.9 (CENTRAL LIMIT THEOREM) *The sum of a large number N of independent and identically distributed random variables with mean μ and standard deviation σ converges to a normal distribution with mean $\mu \cdot N$ and standard deviation $\sigma \cdot \sqrt{N}$.*

- Generally introduced for the Sampling Distribution of the Sample Mean

Tail Inequalities:
Central Limit Theorem (cont’d)

- Since an exact distribution is Normal, the Tail Inequality is that of the “68-95-99.7” rule

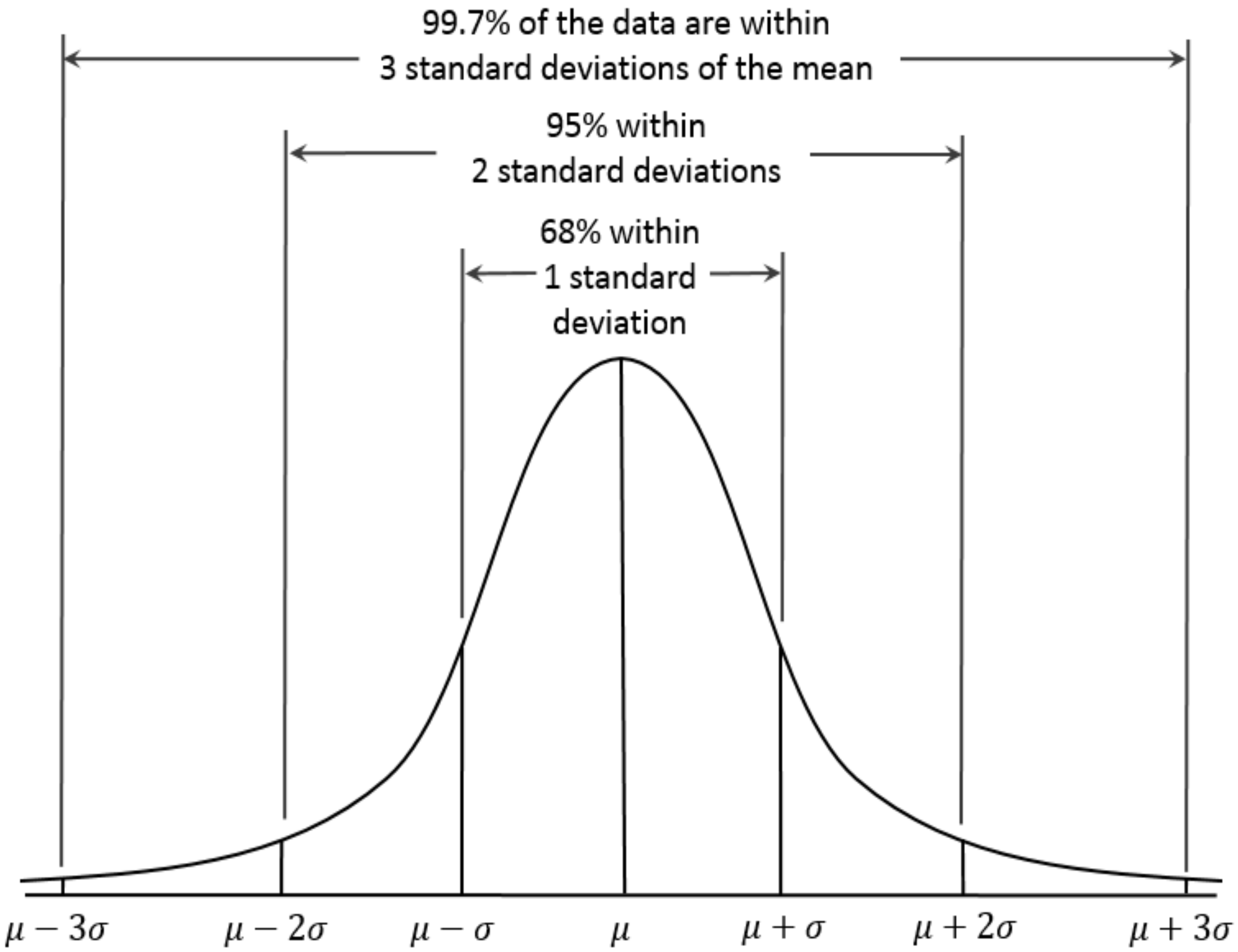
- Example usage: Quality Control, is a machine an outlier in terms of failures? (Sum of i.i.d. Bernoulli)

- Also a generalized CLT for non identical, independent, Random Variables: Lyupanov CLT

Tail Inequalities: "68-95-99.7" rule

- Common rule of thumb for normal distribution

- The probability of a value lying beyond 3 standard deviations is 0.3%



By Dan Kernler - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36506025

Statistical Tail Confidence Testing: Normal Distribution

- Used for when data values are assumed to come from a Normal Distribution (i.e. means, summations, or from domains proven to be Normal)

- Described by the following density function:

$$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2 \cdot \sigma^2}} \qquad (2.16)$$

- To find the probability of the tails, $\mathrm{P}(|\mathrm{X}| > \theta)$, need to integrate the above density

- However this has no closed form

Statistical Tail Confidence Testing: Standard Normal Distribution

- In order to evaluate Tail Probabilities of points from a Normal, we generally standardize values to fit a Normal Distribution with μ = 0, σ = 1

- Standardization is done by the following formula, result is called a Z-score:

$$z_i = (x_i - \mu)/\sigma \tag{2.17}$$

- Can use a Z-table to evaluate tail probabilities of all standardized data points

- Probabilities then used to determine anomalies

Statistical Tail Confidence Testing: Student’s T-Distribution

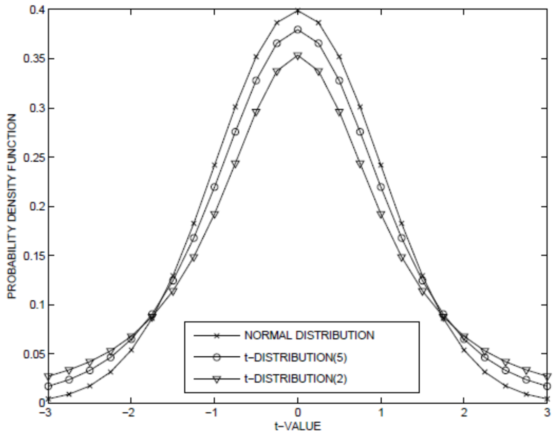
- Z distribution has strict assumption that the population mean and standard deviation are known (Or estimated well from large sample sizes)

- T-distribution is hence used when the mean and standard deviation of the underlying distribution of the data are unknown

$$T(\nu) = \frac{U_0}{\sqrt{(\sum_{i=1}^{\nu} U_i^2)/\nu}} \tag{2.18}$$

- Ratio of a Normal and Chi-Squared Distribution, with parameter ν, degrees of freedom

Statistical Tail Confidence Testing: Student’s T-Distribution (cont’d)



- Above is the T-distribution for varying degrees of freedom it generally has wider tails than the normal (more conservative)

- Data values standardized same way as Z-score (but with sample mean and STD)
- A T-Table then used to evaluate tail probabilities

Extreme Multivariate Value Analysis: Overview

- Designed to determine data points at the boundaries of multivariate data
- Useful when a vector of outlier scores is given for each data point
- Types To Be Discussed:
    1. Depth Based Methods
    2. Deviation Based Methods
    3. Angle Based Methods
    4. Distance-Distribution Based Methods

Extreme Multivariate Value Analysis: Depth Based Methods

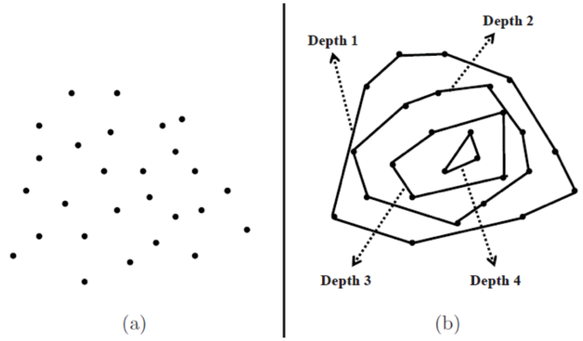


Figure 2.3. Depth-based outlier detection

- An Iterative process: Constructs a convex hull around outer most data points, assigns a depth based on iteration, removes them, and repeats until all points are removed.

- All points before a depth of *r* are outliers.

- Not statistical in nature, Not scalable to high dimensions

- Also not effective in high dimensions as a convex hull in d dimensions has at least 2^d points

Extreme Multivariate Value Analysis: Deviation Based Methods

- These methods find the change in variance of the data when a point is removed

- Removing outliers should reduce the variance significantly, the amount of reduction when points are removed is called the *Smoothing Factor* of these points.

- Outliers are defined as the subset of all points, *E*, for which has the greatest smoothing factor of any other subset of points, *R*.

- Determination of this *E* is difficult since there exist 2^Npossible subsets of *N* points

- A Random Sampling of points is usually employed to find a good *E*. (Smallest set with largest smoothing factor)

Extreme Multivariate Value Analysis: Angle Based Methods

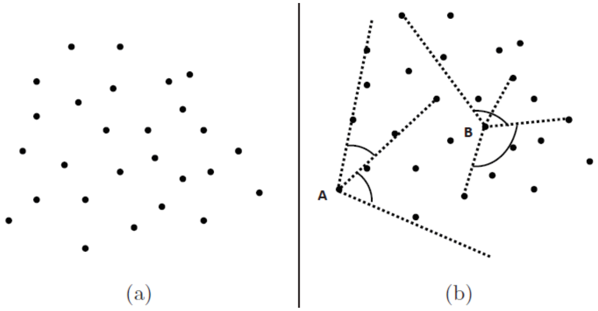


Figure 2.4. Angle-based outlier detection

- The idea is that extreme data points on the boundaries should be able to enclose all the data points with the smallest angle, while data points in the interior should have multiple points around them at multiple angles

- In order to find the angle between two sets of points, a cosine measure is used where $\bar{X} - \bar{Y}$ denotes the vector formed between data point X and Y

$$WCos(\overline{Y}-\overline{X},\overline{Z}-\overline{X}) = \frac{<(\overline{Y}-\overline{X}),(\overline{Z}-\overline{X})>}{||\overline{Y}-\overline{X}||_2^2 \cdot ||\overline{Z}-\overline{X}||_2^2}$$

Extreme Multivariate Value Analysis: Angle Based Methods (cont'd)

- Note that the Cosine measure is weighted, as the magnitudes of the vector are squared, further reducing the angle (and thus penalizing) far away points
- One of the points (*X*) is held constant while the other two are varied to obtain the *variance in spectrum* of *X*
- The *Angle Based Outlier Factor* of X, is thus the set containing weighted Cosine measure of X and varying points from the set of all points, *D*

$$ABOF(\overline{X}) = Var_{\{Y,Z \in \mathcal{D}\}} WCos(\overline{Y} - \overline{X}, \overline{Z} - \overline{X})$$

Extreme Multivariate Value Analysis: Angle Based Methods (cont’d)

- ABOF can be computed a variety of different ways (naive, nearest neighbor)
- Low values of ABOF denote outliers (far away points, non varying angles)
- ABOF is not safe from the ‘curse of dimensionality’ (not guaranteed to be robust for very high dimensions) due to sparsity of the data points, and thus narrower variation in angles for all points.

Extreme Multivariate Value Analysis: Distance Distribution Based

- A distance distribution approach is to model the entire multidimensional dataset as Normally Distributed, given by the following density:

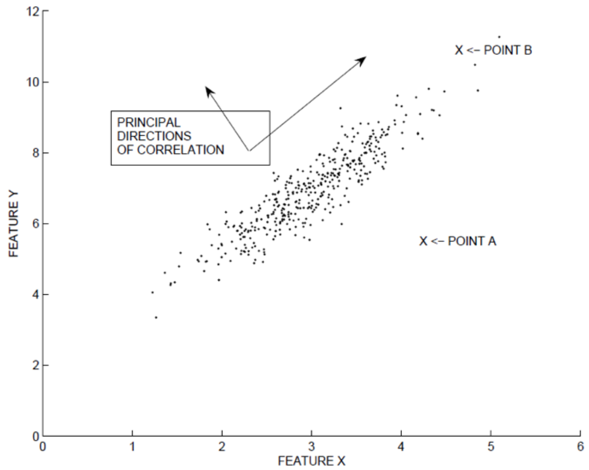
$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot \exp(-\frac{1}{2} \cdot (\overline{X} - \overline{\mu}) \cdot \Sigma^{-1} \cdot (\overline{X} - \overline{\mu})^T)$$

- Here $\bar{\mu}$ is a *dx1* vector of means and Σ is a *dxd* matrix of covariances

- The exponential term is known as the *Mahalanobis Distance* from the center of the density

- The *Mahalanobis Distance* allows computation of probabilities based on the *Principal Directions of Correlation* rather than Euclidian distance

Extreme Multivariate Value Analysis: Distance Distribution Based (cont'd)



- Utilizing Distance Distribution methods, Point A is much more of an outlier than point B. (This would not be the case for angle or deviation based measures)

- Limited by assumptions that all variables are Normally Distributed (one cluster)

- Parameters can be learned through EM

Probabilistic Mixture Modeling
Generative Distributions

- Outliers are usually determined by their relative positions to the data, not just by being on the data's boundaries

- To do this a *Generative* Distribution is assumed for the data

- This Generative Distribution can be a single distribution (as introduced previously) or a mixture of distributions $M = \{G_1, G_2, G_3, \dots\}$, where M is then called the Mixture Model

- These Distributions have the same dimensions as the data and have initially unknown (but assumed) parameters

Probabilistic Mixture Modeling *Generative Distributions* (cont'd)

- In order to obtain the best fit Mixture Model to the data, the parameters of each distribution, as well as the proportion of each distribution in the model, need to be estimated.
- This is done via *Expectation Maximization*
- This process requires us to define the fit of the data to the model

Probabilistic Mixture Modeling
Model Fit

- First we define the fit of a data point, $\bar{X}_j$ to the model, where f^{i} is the joint density of the *ith* distribution, and α_i is the mixture probability of the *ith* distribution:

$$f^{point}(\overline{X_j}|\mathcal{M}) = \sum_{i=1}^{k} \alpha_i \cdot f^i(\overline{X_j}) \tag{2.19}$$

- For the fit of the data to the model we then compute the following known as the *Likelihood Function:* (often Log-Likelihood is used for efficiency)

$$f^{data}(\mathcal{D}|\mathcal{M}) = \prod_{j=1}^{N} f^{point}(\overline{X_j}|\mathcal{M})$$

- The goal of EM is to obtain a Model that maximizes the Likelihood Function

Probabilistic Mixture Modeling
Expectation Maximization

- EM is an iterative procedure

1) Expectation

- Let θ represent the current state of all parameters and mixture probabilities

- Then the *Bayes Probability* that data point $\bar{X}_j$ was generated by $\boldsymbol{G}_i$ is:

$$P(\overline{X_j} \in \mathcal{G}_i | \Theta) = \frac{\alpha_i \cdot f^{i,\Theta}(\overline{X_j})}{\sum_{r=1}^{k} \alpha_r \cdot f^{r,\Theta}(\overline{X_j})} \qquad (2.21)$$

- We assume for the next step $\bar{X}_j$ belongs to the $\boldsymbol{G}_i$ with the highest Bayes Probability

Probabilistic Mixture Modeling
Expectation Maximization (cont'd)

2) Maximization

- First the α values are maximized using *Laplacian Smoothing*

$$(1 + \sum_{j=1}^{N} P(X_j \in \mathcal{G}_i | \Theta))/(k + N)$$

- The parameter values of each G_i are then computed, and updated, using the *Maximum Likelihood Estimator* with the points that were assigned to it in the Expectation step
- I.e. For each distribution G_i, take the Partial Derivative of the (Log) Likelihood Function with respect to each parameter, set it to 0 and solve for the parameter.
- For example, if G_2 is a Multivariate-Gaussian, the MLE's of the parameters $\bar{\mu}_2$ and Σ_2 are the means and co-variances of the data points assigned to G_2 in the E step

Probabilistic Mixture Modeling *Expectation Maximization* (cont’d)

- EM iterates for a defined number of steps until convergence of the parameters is reached
- The result is a Generative Mixture model, where the probabilistic fit of each point can be calculated using f^{point}(not true probabilities)
- These values can be ranked to produce outlier scores for each point
- EM very often used to Mixture Model univariate outlier scores from a variety of methods

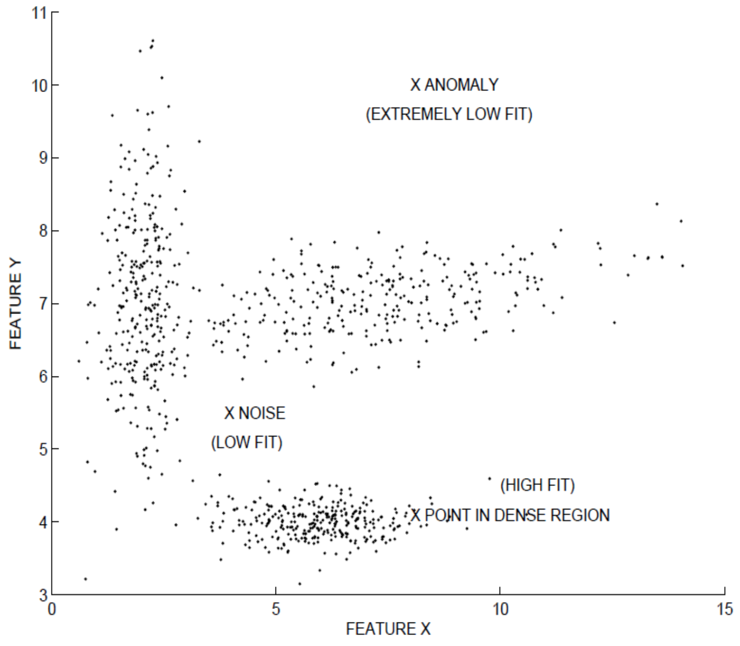


Figure 2.6. Relating Fit Probabilities to the Anomalous Behavior

Probabilistic Mixture Modeling *Limitations*

- Susceptible to overfitting if the chosen distribution is too restrictive.
- Can be too general, for example, often Mixture of Independent Gaussians use naïve independence.
- Does not scale well to high dimensions due to increasing number of parameters to estimate
- Note: EM is one of the most widely used algorithms in all of Data Mining/ Statistics. It was coined by Dempster, Laird, and Rubin in 1977 in the *Journal of the Royal Statistical Society*. (paper has over 34k citations to date!)