Anomaly Detection

CAS RPM 2017

Prashant De

Topics

1) Problems anomaly detection solve
2) The curse of dimensionality
3) Global and Local Outlier Detection + Demo
4) High Dimensional Subspace methods + Demo
5) Actuarial Applications

Problem statement for Actuaries

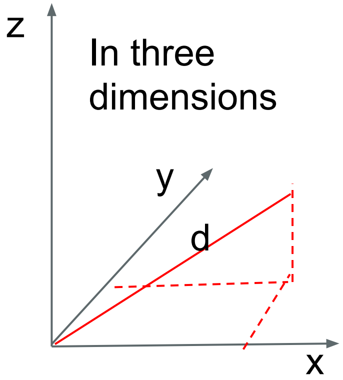
- A modern insurance company in 2017 has seen three great shifts over the last two decades
  - Waves of automation through the 90s and continuing through the 2010s in policy and claims systems leading to better data capture
  - Larger datasets and deeper information. Unstructured data making strides, first in text tagging and mining and now in image and sound processing
  - Supportive software (Apache) and hardware (Nvidia GPUs) making the theoretical practical
- Data generated from new systems now comes from several sources(read potentially different “mechanisms” or “data generating functions”).
- Actuaries need to make decisions on these sources for pricing, claims and fraud detection.

➢ How do we know that the **a) rules and models we use, works for most of the data** and b) **are there other effects in the data that we are missing and c) what doesn’t fit well to the general data distribution**

Curse of dimensionality part 1/2

➢ **Before we get to answer that question, we have a problem.**

- Outlier detection depends on distances between datapoints
- Distances are easy to calculate in a few dimensions, but increasingly difficult in multiple dimensions

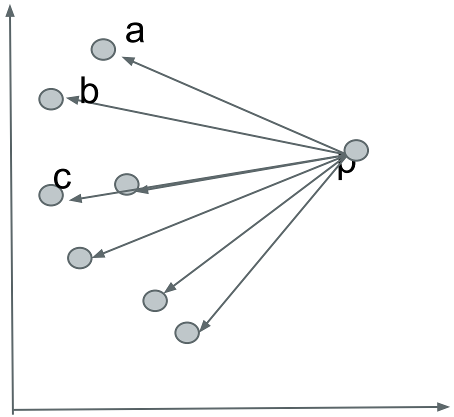


Euclidean distance
d^3 = x^3 + y^3 + z^3

Distance measures are a research topic themselves. Here are two of the most popular

- Euclidean Distance (left): measures the linear distance between the point and in this case the origin. We measure if a point is far away

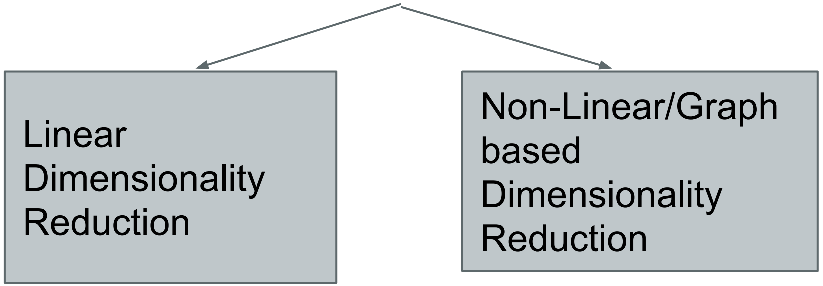
- Angular distance(right): measures the angle from one point to the remaining and claims outliers if the angles are similar



Angle based outlier = Var(ap,cp)/(ap^2 * cp^2)

Curse of dimensionality part 2/2

- **Before we get to answer that question, we have another problem.**
  - Distance and Angle approaches work well in low dimensions and one can compare points well enough
  - In high dimensions the distance from one point to another reaches equity. Therefore as dimensions increase, data needs to be added: This is the curse of dimensionality. Since we may not have more data, we observe sparsity.
- You guessed it, we have dimensionality reduction techniques at hand to help



Dimensionality Reduction

Dimensionality Reduction is a large research topic but the goal is to reduce a set of points in high dimension to a lower dimension for and before analysis

Common linear approaches include: PCA/rPCA, Linear Discriminant Analysis,

Non-linear approaches include: ISOMAP, t-SNE, Diffusion Maps, Neural Net Autoencoders

In business context: simplifying the number of drivers one explores for an outcome (risk, severity, suspicion of fraud) can save time and money

Let’s look at code and demos for each in R

Global and Local Outlier Detection 1/2

Goals:

Outliers detection: *“We want to smooth the data for analysis ” (an older perspective)*

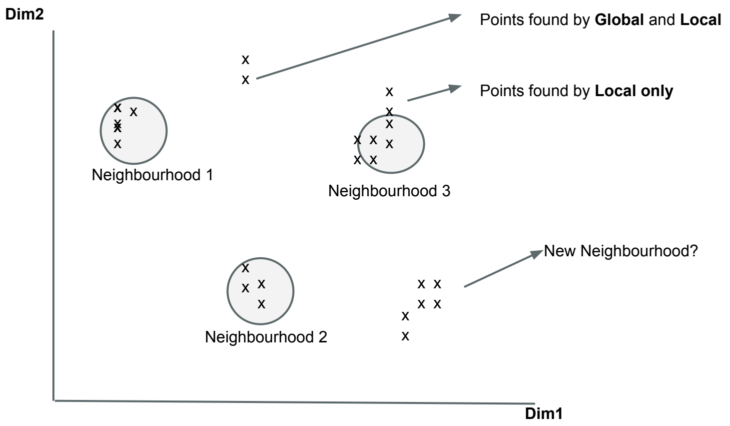
Anomaly detection : *“We want to classify data that is rare and different from the expected data generating process or identify issues with the process”*

Applied Fraud/Intrusion Detection: *“We want to identify individuals or groups that are behaving suspiciously”*

K-NN Detection R Demo

Global and Local Outlier Detection 2/2

Goals of Local Outlier Factor detection: “Use the k neighbours from initial analysis and find distance measures from neighbourhood to local outlier; finding points that differ to neighbourhood”



1) Each point is compared to their neighbourhood
2) A local Distance is calculated with respect to the neighbourhood
3)

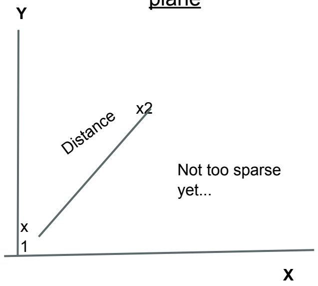
LOF Detection R Demo

Detection in High Dimensional Space

“Sparse dataset”: A dataset where the distances between points are roughly equal, data is not clustered in high dimensions, instead usually in isolation with *space* in between.

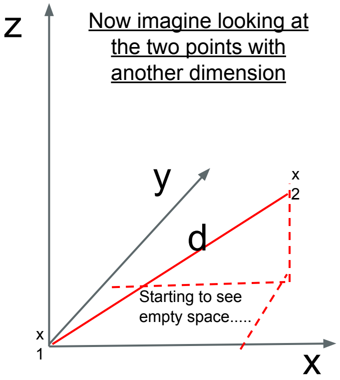
Goal: Sparsity in insurance data increases as we add dimensions and importantly, unstructured data: especially text. *How can an insurance application such as Claims or UW leakage adjust to this growing trend?*

Imagine looking **down on points** on a 2D plane



Euclidean distance
d^2 = (x2-x1)^2 + (y2-y1)^2

Now imagine looking at the two points with another dimension



Euclidean distance
d^3 = x^3 + y^3 + z^3

R Demo - Subspace Outlier techniques

Demo using Apache Spark and H2o

R, Spark and H20 Demo for anomaly detection

References

Papers

Hans Peter Kriegel : “Angle-Based Outlier Detection in High-dimensional Data “ , “Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data”

Hodge, V.J. and Austin, J. orcid.org/0000-0001-5762-8614 (2004) A survey of outlier detection methodologies. Artificial Intelligence Review. pp. 85-126. ISSN 1573-7462.

Gustavo H. Orair Carlos H. C. Teixeira Wagner Meira Jr et al Distance-Based Outlier Detection: Consolidation and Renewed Bearing

Arthur Zimek : Ensembles for Unsupervised Outlier Detection: Challenges and Research Question

Charu C. Aggarwal et all: Outlier Detection for High Dimensional Data