CSE 590
Data Science Fundamentals

Outlier Analysis

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Data Science components and tasks	
3	Data types	Project #1 out
4	Introduction to R, statistics foundations	
5	Introduction to D3, visual analytics	
6	Data preparation and reduction	
7	Data preparation and reduction	Project #1 due
8	Similarity and distances	Project #2 out
9	Similarity and distances	
10	Cluster analysis	
11	Cluster analysis	
12	Pattern mining	Project #2 due
13	Pattern mining	
14	Outlier analysis	
15	Outlier analysis	Final Project proposal due
16	Classifiers	
17	Midterm	
18	Classifiers	
19	Optimization and model fitting	
20	Optimization and model fitting	
21	Causal modeling	
22	Streaming data	Final Project preliminary report due
23	Text data	
24	Time series data	
25	Graph data	
26	Scalability and data engineering	
27	Data journalism	
	Final project presentation	Final Project slides and final report due

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

Also called:

- abnormalities
- discordants
- deviants
- anomalies

Tough problem

- often not easy to see

Applications

Credit card fraud detection

- is it you using the credit card?
- typically an unusual combinations of attributes, such as high frequency transactions in a particular location
- but there are many attributes which makes this difficult

Financial fraud

- use temporal detection methods to detect an anomalous crash

Web log analytics

- unusual sequence may indicate hacking attempt
- use sequence analysis

Other applications

- intrusion detection applications
- biological and medical applications
- earth science applications
- quality control and fault detection in manufacturing

Basics

Working assumption:

- there are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

Challenges

- how many outliers are there in the data?
- detection is usually unsupervised
- this makes validation quite challenging
- it's often like "finding a needle in a haystack"

Outlier models

- extreme value
- cluster
- distance-based
- density-based
- probabilistic
- information-theoretic

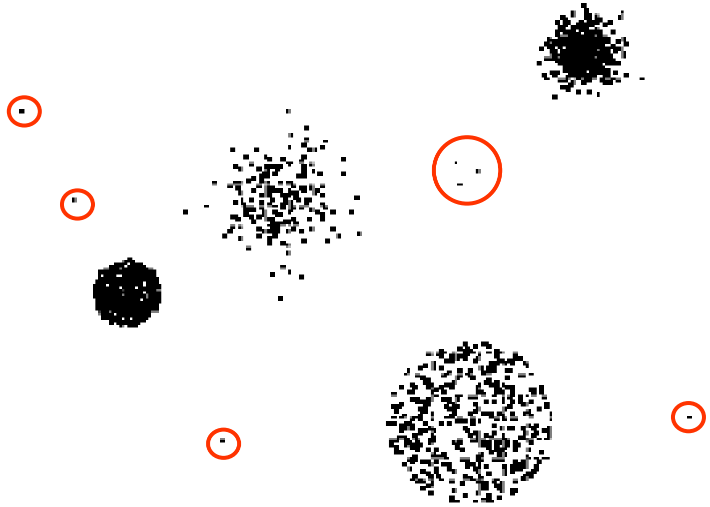
Anomaly Detection Schemes

General Steps

- build a profile of the "normal" behavior
  - profile can be patterns or summary statistics for the overall population
- use the "normal" profile to detect anomalies
  - anomalies are observations whose characteristics differ significantly from the normal profile

Types of anomaly detection schemes:

- graphical & statistical-based
- distance-based
- model-based

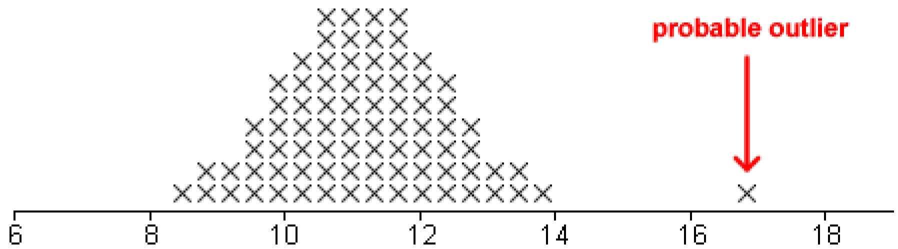


Extreme Value Analysis

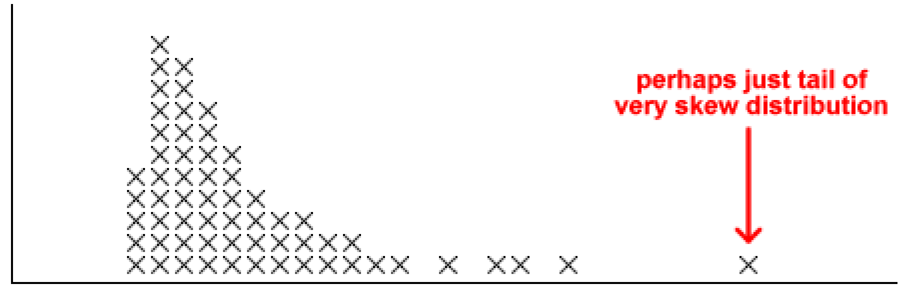
Assume normal distribution

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

- then a data point k standard deviations away from the mean is an outlier



- but beware of skewed distributions



- skew can be measured

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}},$$

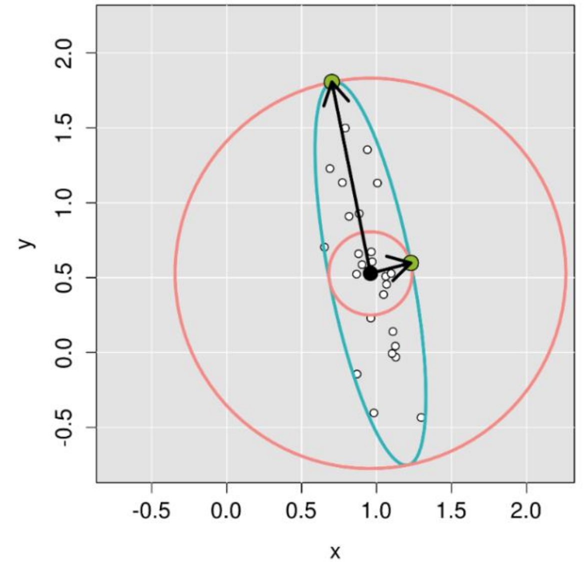
Multivariate Extreme Values

Now the normal distribution becomes multivariate

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

- mean vector μ and covariance matrix Σ
- $|\Sigma|$ is the determinant of the covariance matrix
- expressing the distance in terms of the Mahalanobis distance

$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma| \cdot (2 \cdot \pi)^{(d/2)}}} \cdot e^{-\frac{1}{2} \cdot Maha(\overline{X}, \overline{\mu}, \Sigma)^2}.$$



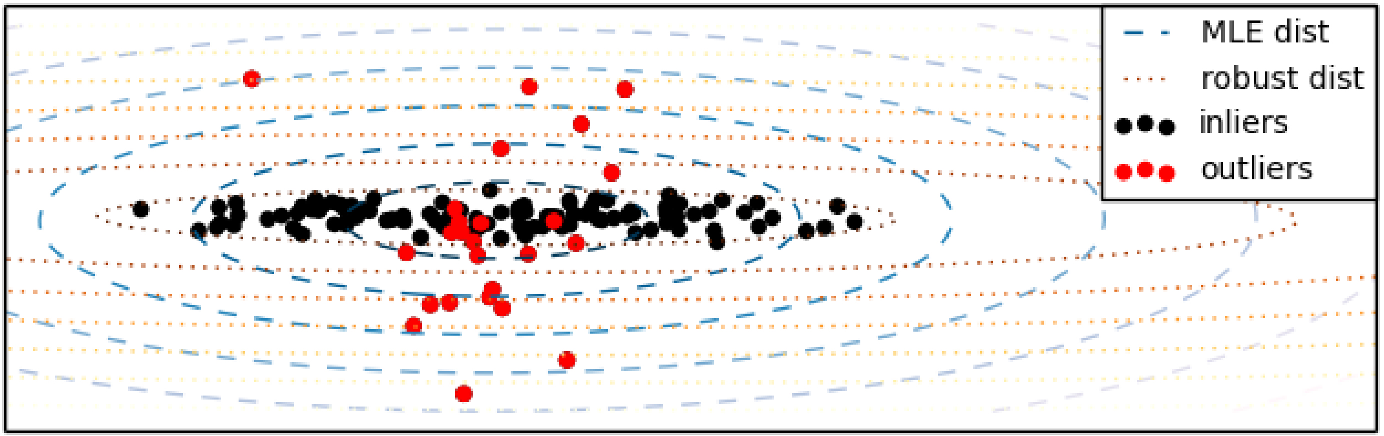
- both green points have the same Mahalanobis distance

Multivariate Extreme Values

Use the Mahanalobis distance to measure outlierness

- smaller values of this probabilistic measure imply greater likelihood of being an extreme value
- more distribution-aware than the Euclidian distance

Mahalanobis distances of a contaminated data set:

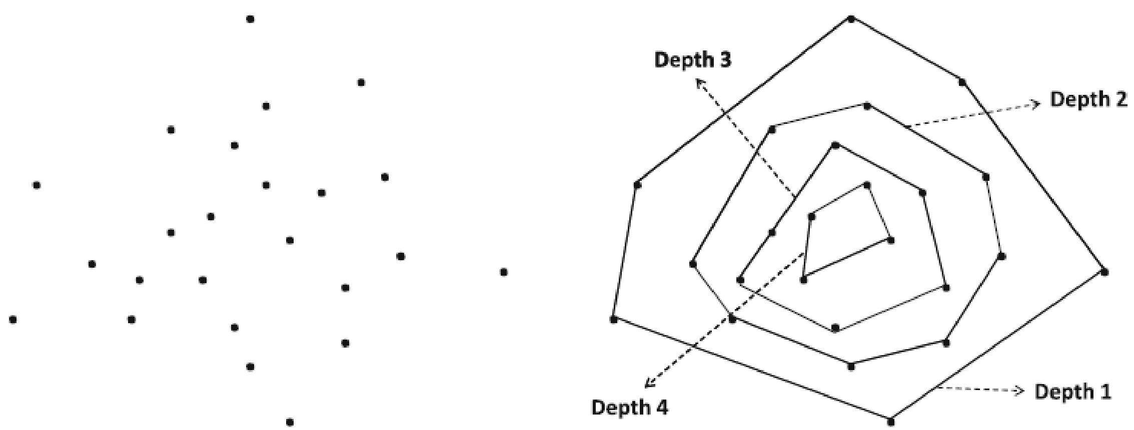


Depth-Based Method

Proceeds in an iterative fashion

For each iteration k

- remove points at the corners of the convex hull of the data
- use k as the outlier score
- a smaller values indicate a greater tendency for a data point to be an outlier
- repeat until the data set is empty



Depth-Based vs. Mahalanobis

Generally depth-based method compares less favorably

Can you imagine why?

Downsides of the depth-based method

- more computationally intensive. especially for high dimensions
- not as accurate since it is does not normalize to the characteristics of the data distribution
- fraction of data points at the corners of the convex hull generally increases with dimensionality
- for very high dimensionality, it may not be uncommon for the majority of the data points to be located at the corners of the outermost convex hull
- as a result, it is no longer possible to distinguish the outlier scores of different data points

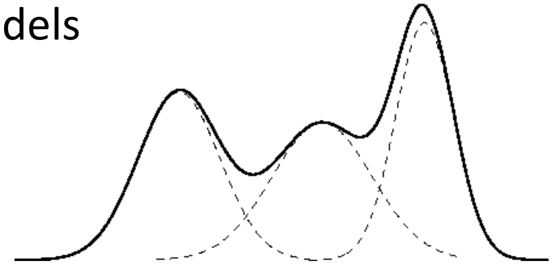
Extension to Mixture Models

Outlier score (density) for point X_j given mixture model *M* of k distributions

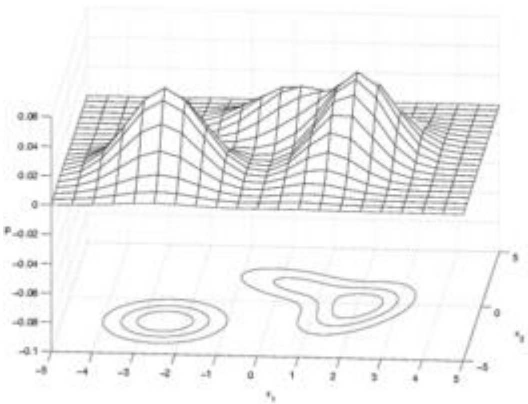
$$f^{point}(\overline{X_j}|\mathcal{M}) = \sum_{i=1}^{k} \alpha_i \cdot f^i(\overline{X_j}).$$

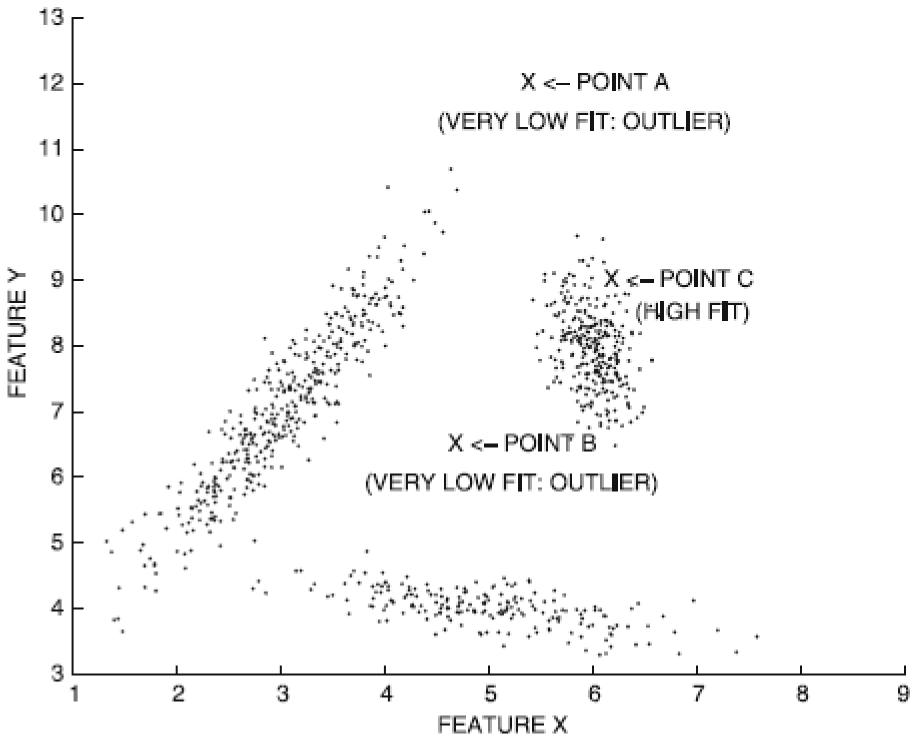
Mixture models

1D



2D





Finding the Mixture model

The model parameters M need to be computed first

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

Convert to the log-likelihood

$$\mathcal{L}(\mathcal{D}|\mathcal{M}) = \log(\prod_{j=1}^{n} f^{point}(\overline{X_j}|\mathcal{M})) = \sum_{j=1}^{n} \log(\sum_{i=1}^{k} \alpha_i \cdot f^i(\overline{X_j})).$$

and apply the EM algorithm to find the model parameters

- mean vector μ and covariance matrix Σ for each of the k distributions, given the n points
- see clustering lecture on EM

Clustering

First cluster the dataset and then use the raw distance of the data point to its closest cluster centroid

- use Mahanalobis to define the outlier score of a data point

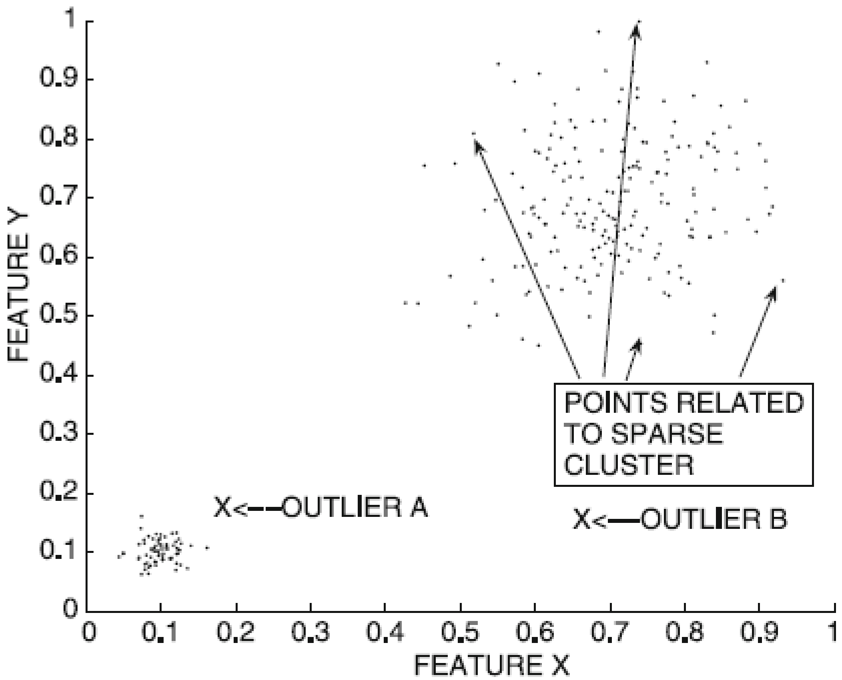
Algorithm

- cluster data
- compute statistical parameters for each cluster
- compute *global* Mahanalobis score using

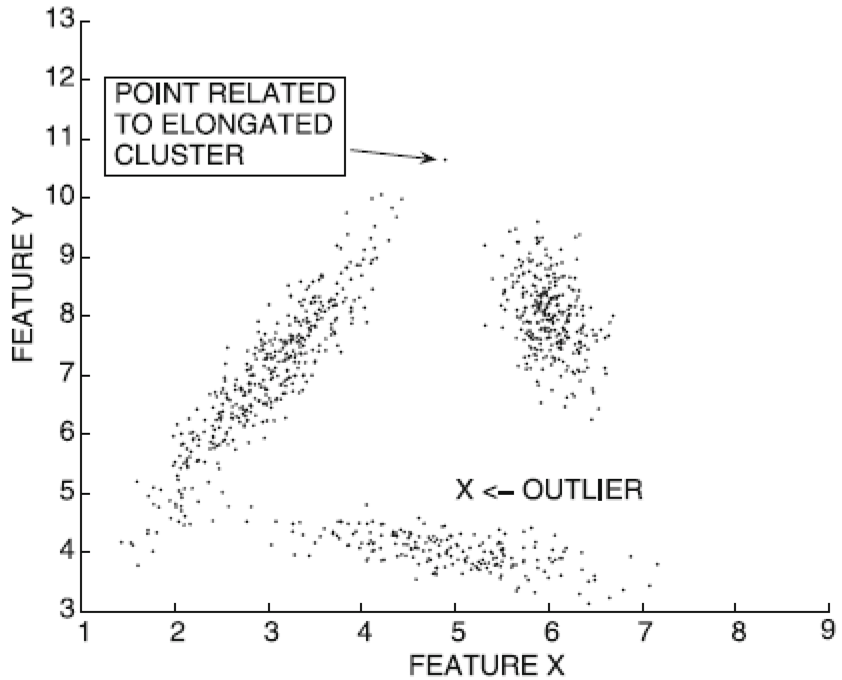
$$f^{point}(\overline{X_j}|\mathcal{M}) = \sum_{i=1}^{k} \alpha_i \cdot f^i(\overline{X_j}).$$

- if global score is high, compute *local* Mahanalobis score for the closest cluster

Realistic Conditions



(a) Varying cluster density

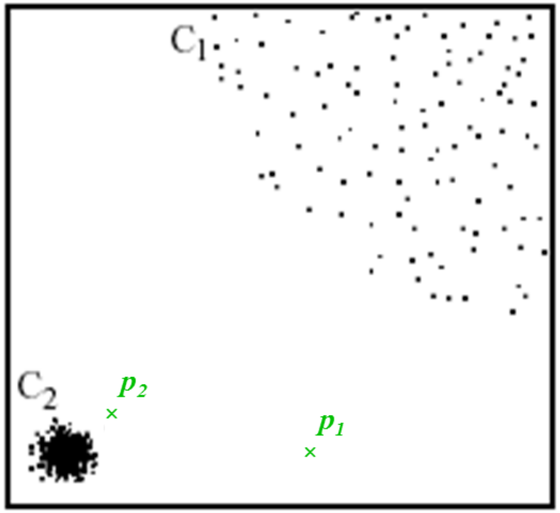


(b) Varying cluster shape

Local Outlier Factor (LOF)

Adjusts for local variations in cluster density by normalizing distances with the average point-specific distances in a data locality

- for each point, compute the density of its local neighborhood
- compute local outlier factor (LOF) of a sample p as the average of the ratios of the density (e.g., number of close neighbors) of sample p and the density of its nearest neighbors
- outliers are points with largest LOF value



In the distance-based approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Outlier Validity

Need to evaluate the tradeoffs in outlier detection

- picking the right threshold to decide outlier or not can be tricky
- threshold too restrictive increases false negatives (misses outliers)
- threshold too relaxed increases false positives (counts too many)

Use Receiver Operating Characteristic (ROC) to compare the effect of different thresholds

- see next slide for an explanation of ROC

Receiver Operating Characteristic (ROC)

TPR: True positive rate
FPR: False positive rate
P: positives
N: negatives
N: total data

TP	FP
FN	TN

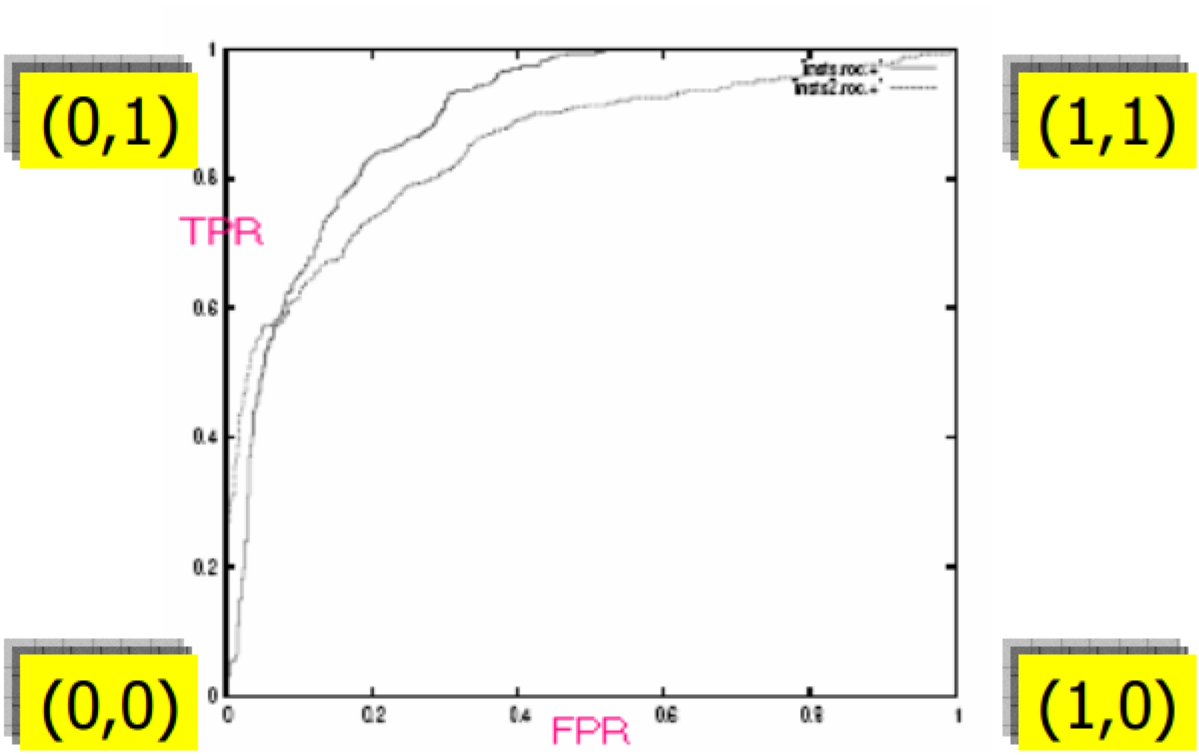
$$TPR = \frac{TP}{P} = Recall, \; FPR = \frac{FP}{N}$$

$$Precision = \frac{TP}{TP + FP}, \; Accuracy = \frac{TP + TN}{P + N}$$

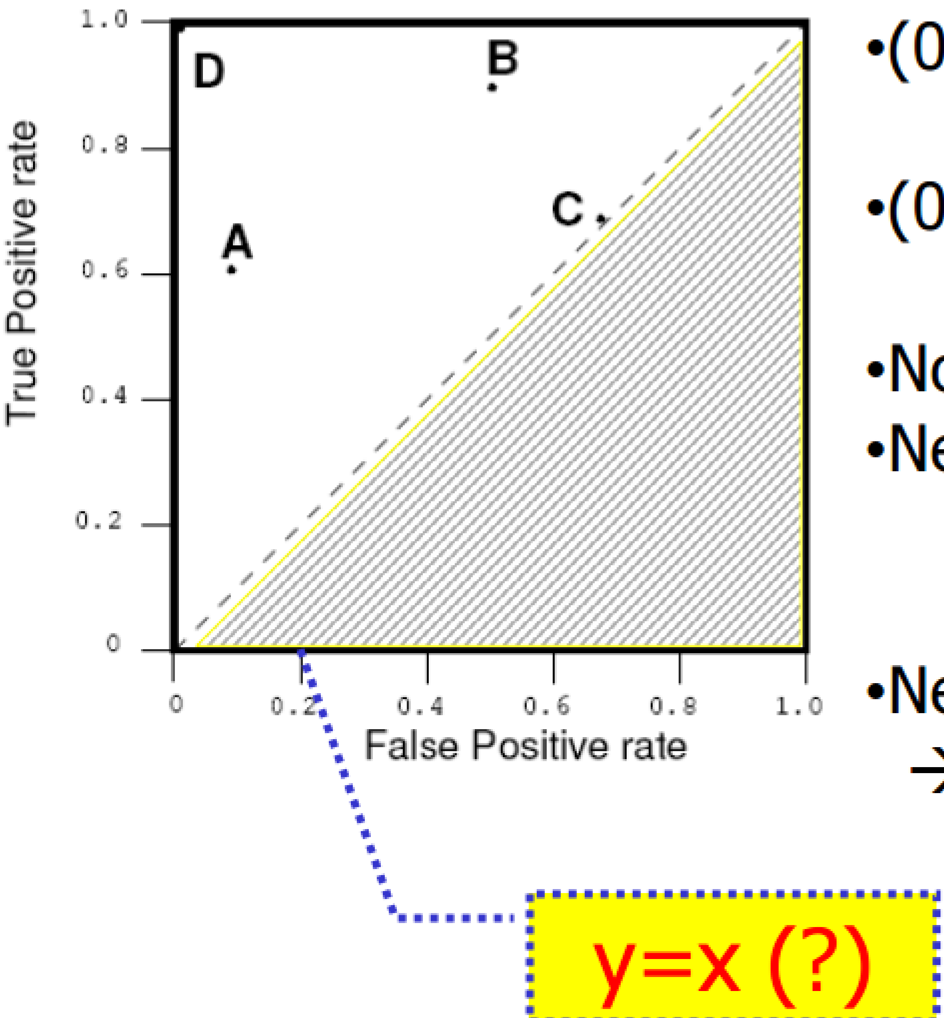
$$Sensitivity = Recall, \; Specificity = 1 - FPR$$

Receiver Operating Characteristic (ROC)

- Y axis: TPR
  X axis: FPR



Receiver Operating Characteristic (ROC)

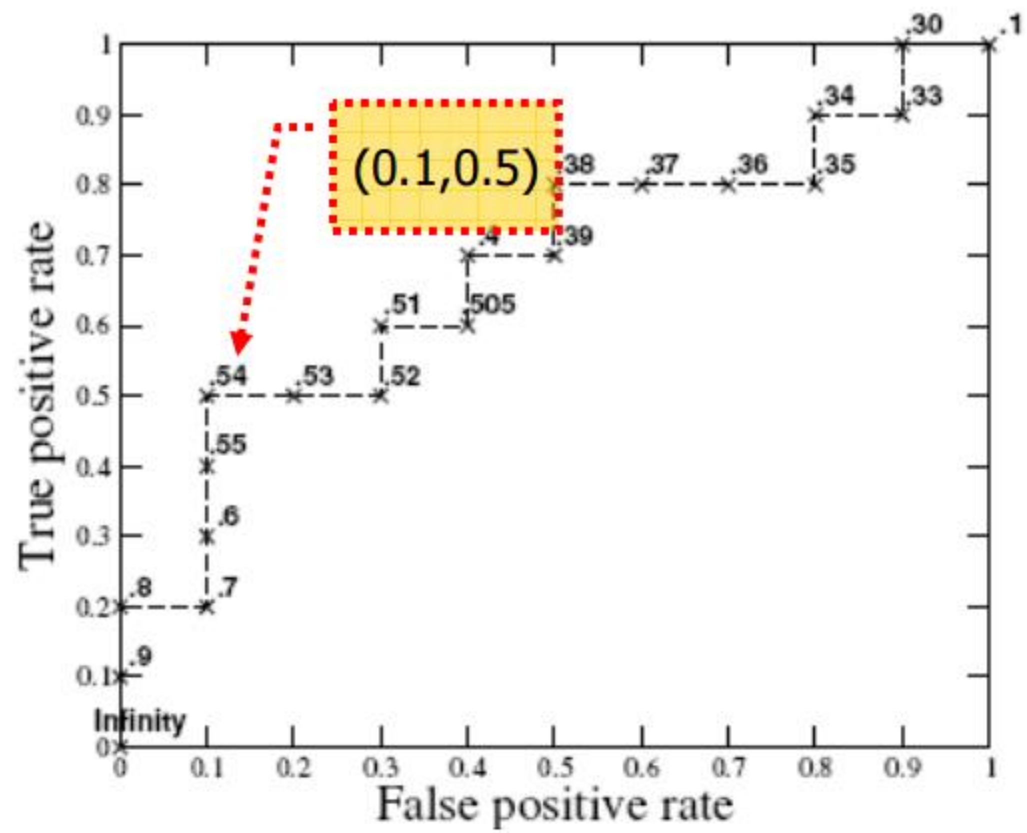


- (0,0) Numbers of P =0,
  → No FP error, No TP
- (0,1) perfect
  → D classifiers
- Northwest location is better.
- Near x axis and on the left side
  → Conservative
  e.g. A vs. B
- Near upper right-hand side
  → Liberal

Receiver Operating Characteristic (ROC)

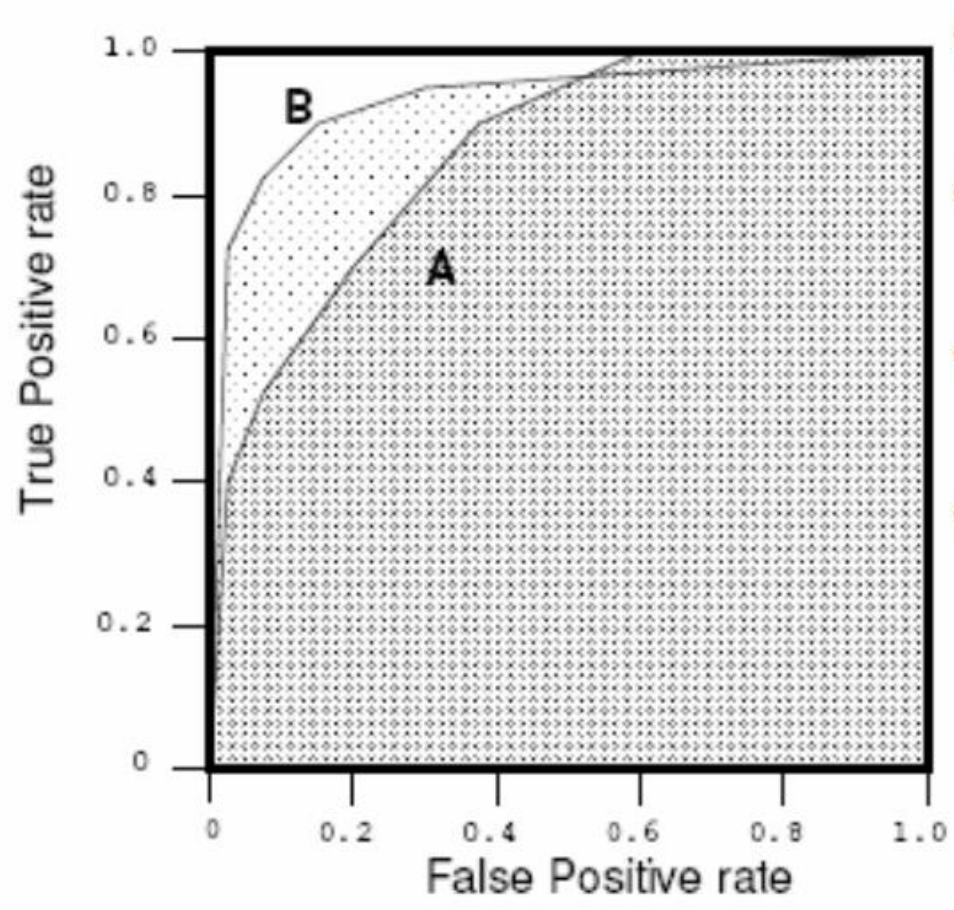
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

If threshold =0.54 →
Numbers of Score ≧0.54 → 6



Area Under The ROC Curve

Important metric for the quality of a classifier



- AUC (Bradley, 1997)
- Wilcoxon test of ranks
- Area : Classifier B > A
- Average performance
  → B > A

Apply ROC to Outlier Classifier

Use Receiver Operating Characteristic (ROC) to compare the effect of different thresholds

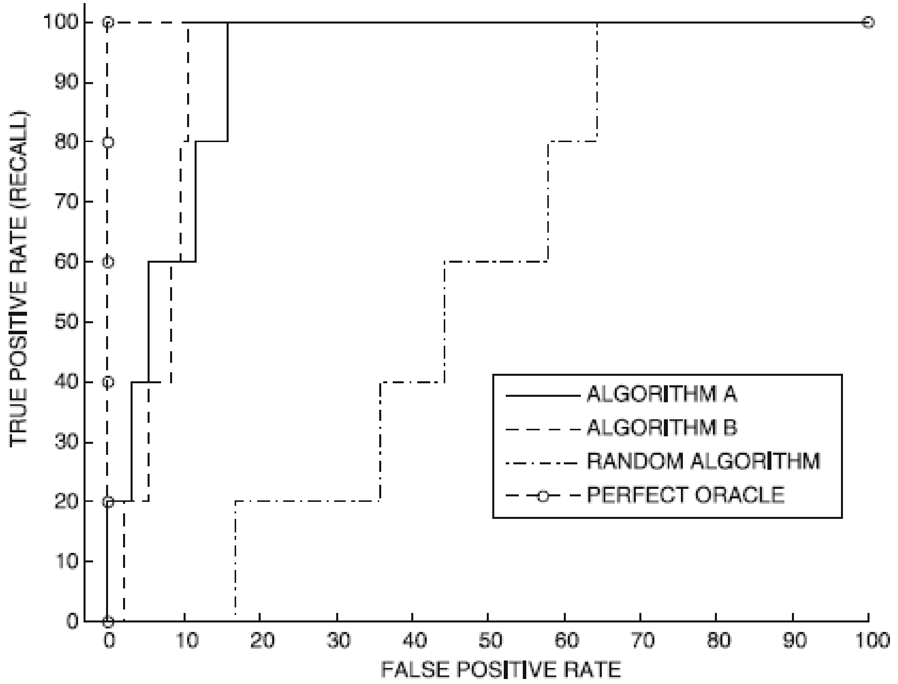
- G = the true set (ground-truth set) of outliers in the data set
- S = set of declared outliers, D = overall dataset
- True positive rate: $TPR(t) = Recall(t) = 100 * \frac{|\mathcal{S}(t) \cap \mathcal{G}|}{|\mathcal{G}|}$
- False positive rate: $FPR(t) = 100 * \frac{|\mathcal{S}(t) - \mathcal{G}|}{|\mathcal{D} - \mathcal{G}|}.$

ROC – Example

Rank 100 data points

- outlierness in increasing numerical order
- A does well early on, but then picks lower-ranked outliers later
- B does worse early on, but then gets more accurate later
- so overall B is better

  has a better ROC-area



Algorithm	Rank of ground-truth outliers
Algorithm A	1, 5, 8, 15, 20
Algorithm B	3, 7, 11, 13, 15
Random Algorithm	17, 36, 45, 59, 66
Perfect Oracle	1, 2, 3, 4, 5

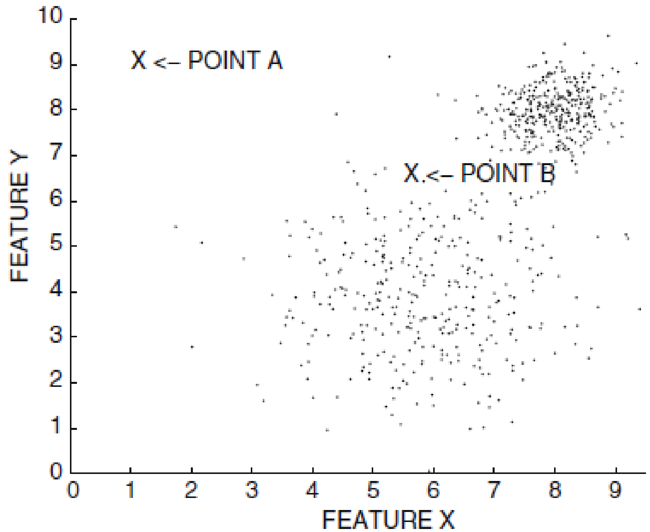
High-Dimensional Data

Often an anomaly can be typically perceived in only a small subset of the dimensions

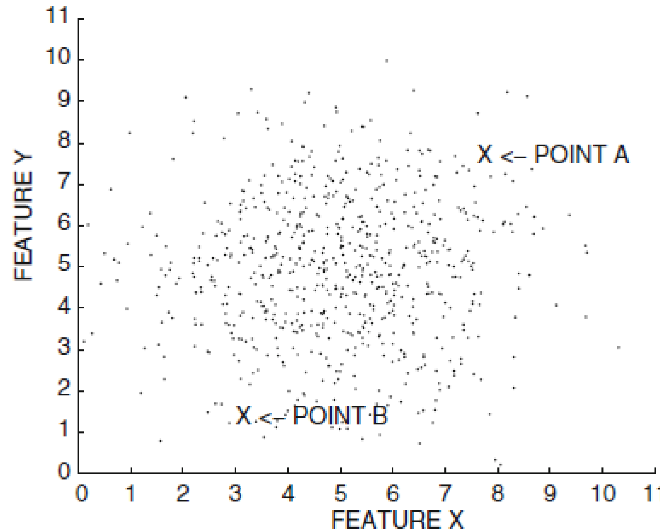
- the remaining dimensions are irrelevant and only add noise to the anomaly-detection process
- furthermore, different subsets of dimensions may be relevant to different anomalies.
- As a result, full-dimensional analysis often does not properly expose the outliers in high-dimensional data

The next slide shows this for four 2D projections

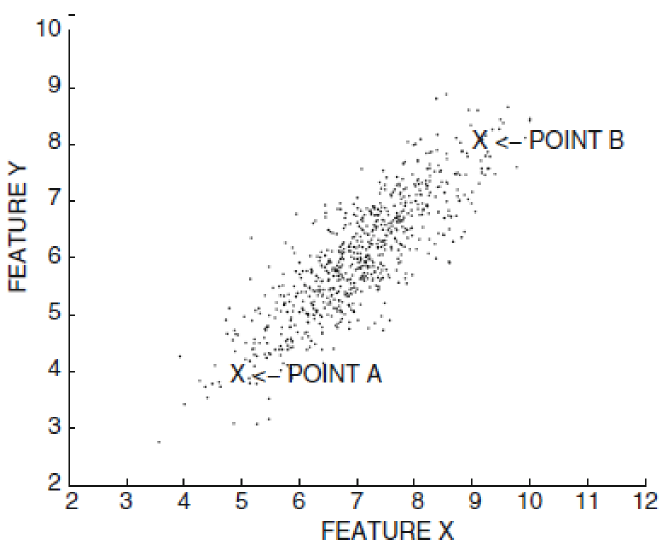
- outliers A and B are exposed in different subspaces but not in others
- so attribute selection is crucial → subspaces
- need to find the one or more subspaces crucial to the outlier



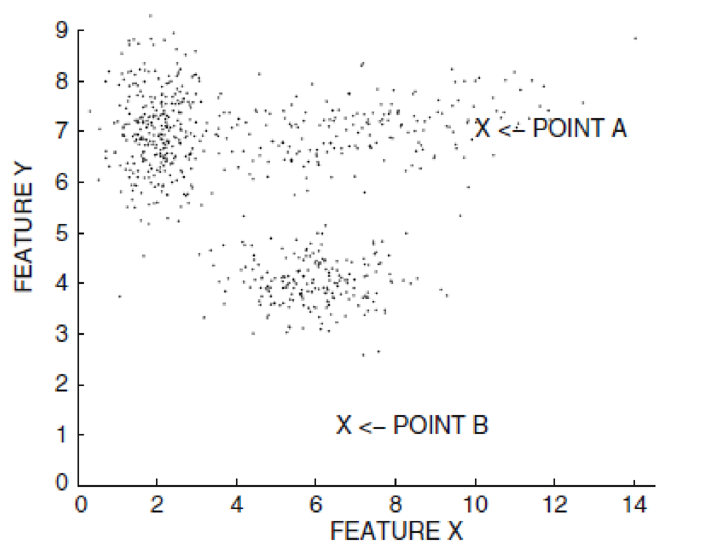
(a) View 1
Point A is outlier



(b) View 2
No outliers



(c) View 3
No outliers



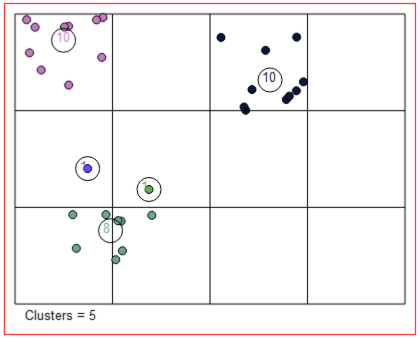
(d) View 4
Point B is outlier

Relevant Subspace Selection

Number of possible subspaces of a d-dimensional data point is 2^d

- daunting problem when d is large

Two strategies to deal with this



Grid-based rare subspace exploration

- discretize the data into a grid-like structure
- outliers are in cells where the density is lower than expected assuming uniform distributions

Random subspace sampling

- subspaces of the data are sampled to discover the most relevant outliers
- perform usual tests on these subsets