ALGONQUIN
COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 5

Outlier Detection

Assignment 2

Final Project

- Project selections must be submitted by Sunday June 20 at 11:59 PM.

Decision Trees - Recap

- Type of attributes
- Parameters
    - Minimum number of objects
    - Pruning
- Numbers at leaf: The first number is the total number of instances reaching the leaf. The second number is the number of misclassified instances

Introduction

- What is outlier detection?
- Outlier Detection using Statistical Methods
- Outlier Detection using Machine Learning techniques

What is an outlier?

- A data object that deviates significantly from the Majority of normal objects
- Ex.: unusual credit card purchase

Applications of Outlier Detection

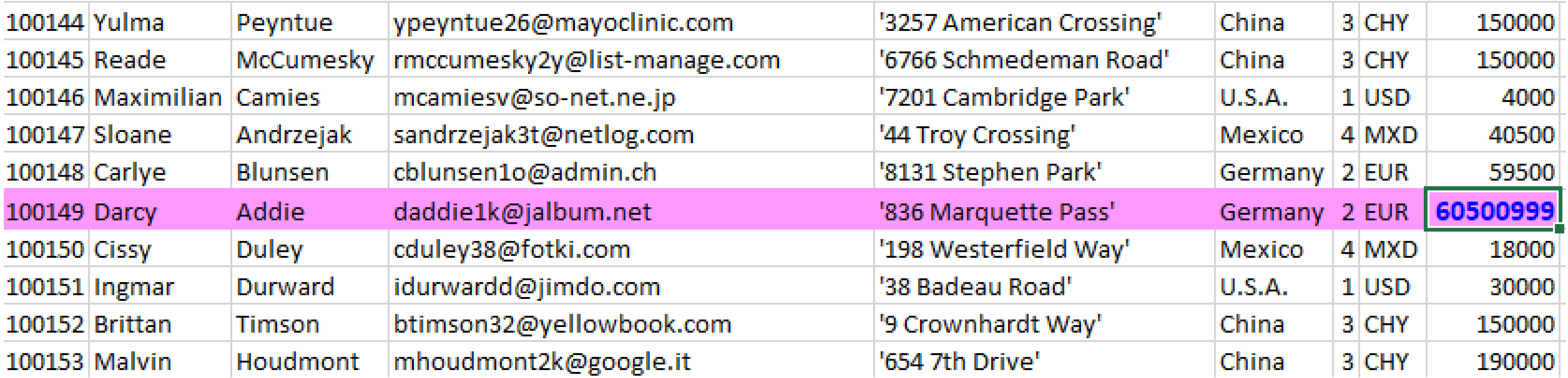
- Financial fraud detection (banking, credit card etc.)
- Telecom fraud detection
- Medical Diagnosis
- Web Analytics

Types of Outliers

- Three types:
  - ➢Global Outlier (point anomalies)
  - ➢Contextual outlier (conditional outlier)
  - ➢Collective Outliers

Global Outlier

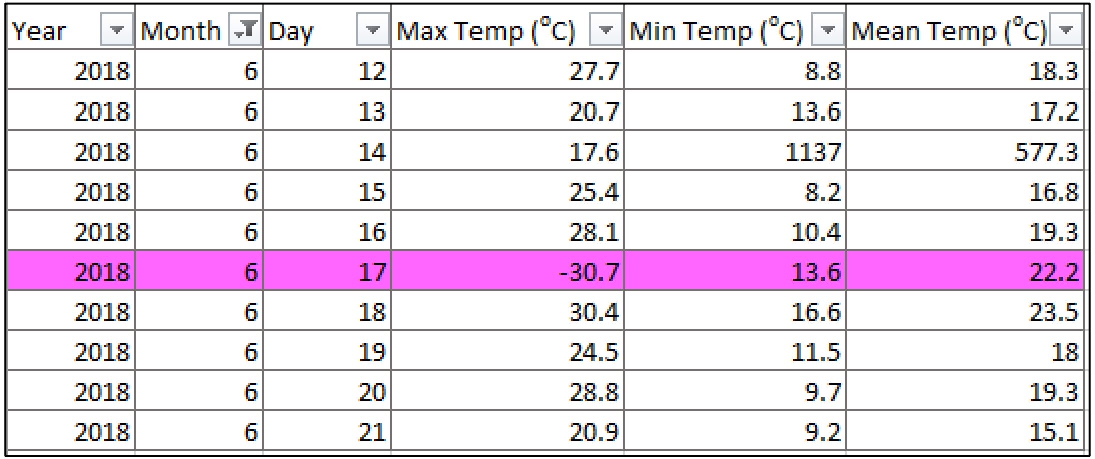
- A data point is considered a global outlier if its value is far outside the entirety of the data set in which it is found



100144	Yulma	Peyntue	ypeyntue26@mayoclinic.com	'3257 American Crossing'	China	3	CHY	150000
100145	Reade	McCumesky	rmccumesky2y@list-manage.com	'6766 Schmedeman Road'	China	3	CHY	150000
100146	Maximilian	Camies	mcamiesv@so-net.ne.jp	'7201 Cambridge Park'	U.S.A.	1	USD	4000
100147	Sloane	Andrzejak	sandrzejak3t@netlog.com	'44 Troy Crossing'	Mexico	4	MXD	40500
100148	Carlye	Blunsen	cblunsen1o@admin.ch	'8131 Stephen Park'	Germany	2	EUR	59500
100149	Darcy	Addie	daddie1k@jalbum.net	'836 Marquette Pass'	Germany	2	EUR	60500999
100150	Cissy	Duley	cduley38@fotki.com	'198 Westerfield Way'	Mexico	4	MXD	18000
100151	Ingmar	Durward	idurwardd@jimdo.com	'38 Badeau Road'	U.S.A.	1	USD	30000
100152	Brittan	Timson	btimson32@yellowbook.com	'9 Crownhardt Way'	China	3	CHY	150000
100153	Malvin	Houdmont	mhoudmont2k@google.it	'654 7th Drive'	China	3	CHY	190000

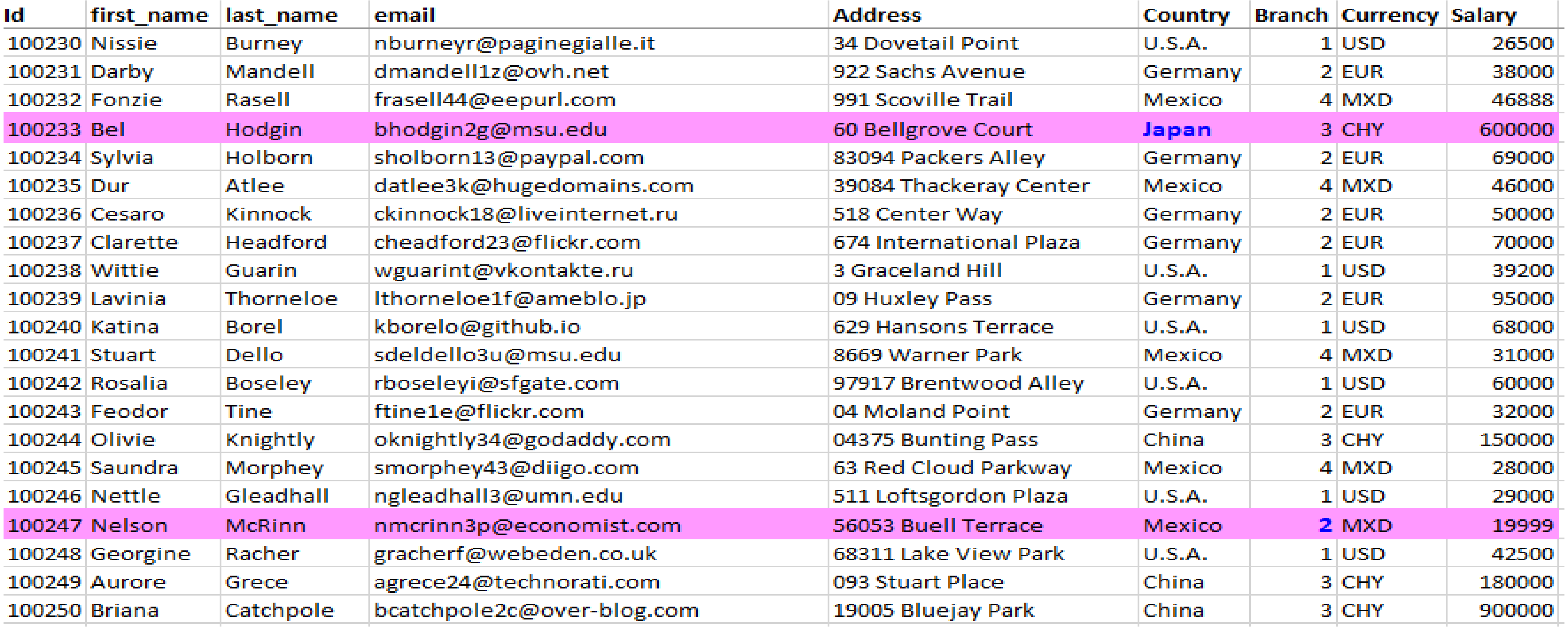
Contextual Outlier

- If the value deviates significantly based on a selected context
- Ex: a temp of -30.7 degree Celsius during the month of June in Ottawa.



Year	Month	Day	Max Temp (°C)	Min Temp (°C)	Mean Temp (°C)
2018	6	12	27.7	8.8	18.3
2018	6	13	20.7	13.6	17.2
2018	6	14	17.6	1137	577.3
2018	6	15	25.4	8.2	16.8
2018	6	16	28.1	10.4	19.3
2018	6	17	-30.7	13.6	22.2
2018	6	18	30.4	16.6	23.5
2018	6	19	24.5	11.5	18
2018	6	20	28.8	9.7	19.3
2018	6	21	20.9	9.2	15.1

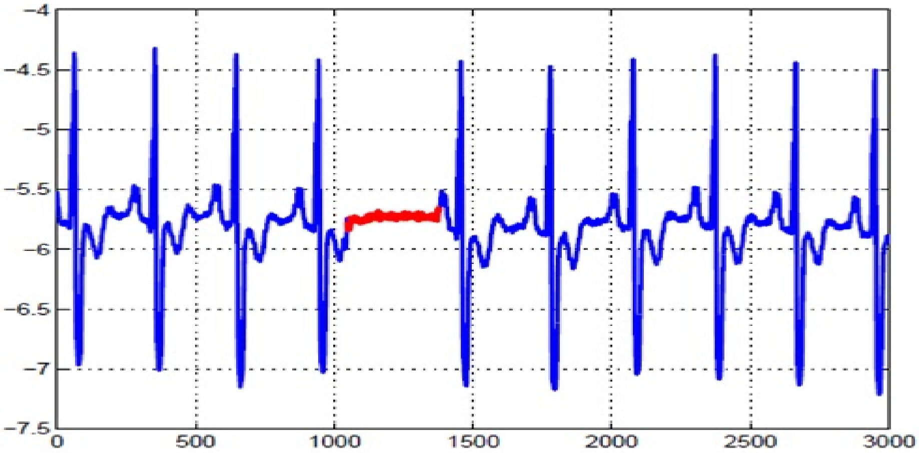
Contextual Outlier - Example



Id	first_name	last_name	email	Address	Country	Branch	Currency	Salary
100230	Nissie	Burney	nburneyr@paginegialle.it	34 Dovetail Point	U.S.A.	1	USD	26500
100231	Darby	Mandell	dmandell1z@ovh.net	922 Sachs Avenue	Germany	2	EUR	38000
100232	Fonzie	Rasell	frasell44@eepurl.com	991 Scoville Trail	Mexico	4	MXD	46888
100233	Bel	Hodgin	bhodgin2g@msu.edu	60 Bellgrove Court	Japan	3	CHY	600000
100234	Sylvia	Holborn	sholborn13@paypal.com	83094 Packers Alley	Germany	2	EUR	69000
100235	Dur	Atlee	datlee3k@hugedomains.com	39084 Thackeray Center	Mexico	4	MXD	46000
100236	Cesaro	Kinnock	ckinnock18@liveinternet.ru	518 Center Way	Germany	2	EUR	50000
100237	Clarette	Headford	cheadford23@flickr.com	674 International Plaza	Germany	2	EUR	70000
100238	Wittie	Guarin	wguarint@vkontakte.ru	3 Graceland Hill	U.S.A.	1	USD	39200
100239	Lavinia	Thorneloe	lthorneloe1f@ameblo.jp	09 Huxley Pass	Germany	2	EUR	95000
100240	Katina	Borel	kborelo@github.io	629 Hansons Terrace	U.S.A.	1	USD	68000
100241	Stuart	Dello	sdeldello3u@msu.edu	8669 Warner Park	Mexico	4	MXD	31000
100242	Rosalia	Boseley	rboseleyi@sfgate.com	97917 Brentwood Alley	U.S.A.	1	USD	60000
100243	Feodor	Tine	ftine1e@flickr.com	04 Moland Point	Germany	2	EUR	32000
100244	Olivie	Knightly	oknightly34@godaddy.com	04375 Bunting Pass	China	3	CHY	150000
100245	Saundra	Morphey	smorphey43@diigo.com	63 Red Cloud Parkway	Mexico	4	MXD	28000
100246	Nettle	Gleadhall	ngleadhall3@umn.edu	511 Loftsgordon Plaza	U.S.A.	1	USD	29000
100247	Nelson	McRinn	nmcrinn3p@economist.com	56053 Buell Terrace	Mexico	2	MXD	19999
100248	Georgine	Racher	gracherf@webeden.co.uk	68311 Lake View Park	U.S.A.	1	USD	42500
100249	Aurore	Grece	agrece24@technorati.com	093 Stuart Place	China	3	CHY	180000
100250	Briana	Catchpole	bcatchpole2c@over-blog.com	19005 Bluejay Park	China	3	CHY	900000

Collective Outliers

- A subset of data objects *collectively* deviate significantly from the whole data set, even if the individual data objects may not be outliers



The highlighted region denotes an outlier because the same low value exists for an abnormally long time. The low value by itself is not an outlier but its successive occurrence for long time is an outlier.

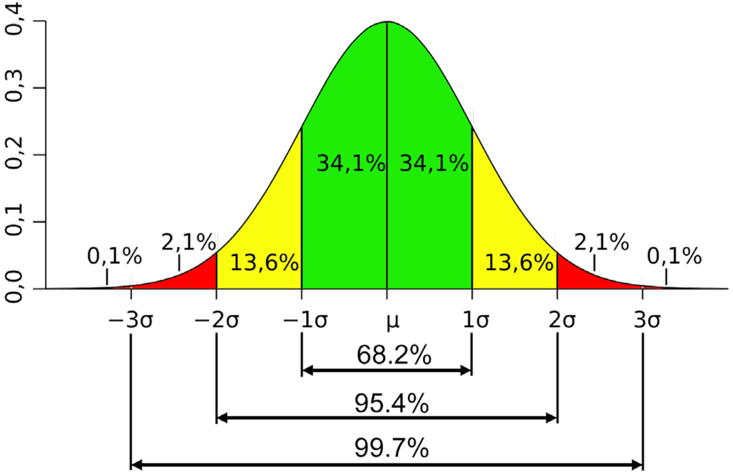
Methods for Outlier Detection

- Statistical Methods
- Proximity-based methods
  - Distance-based
  - Density-based (Ex. Local Outlier Factor - LOF)
- Clustering-based methods

Gaussian distribution

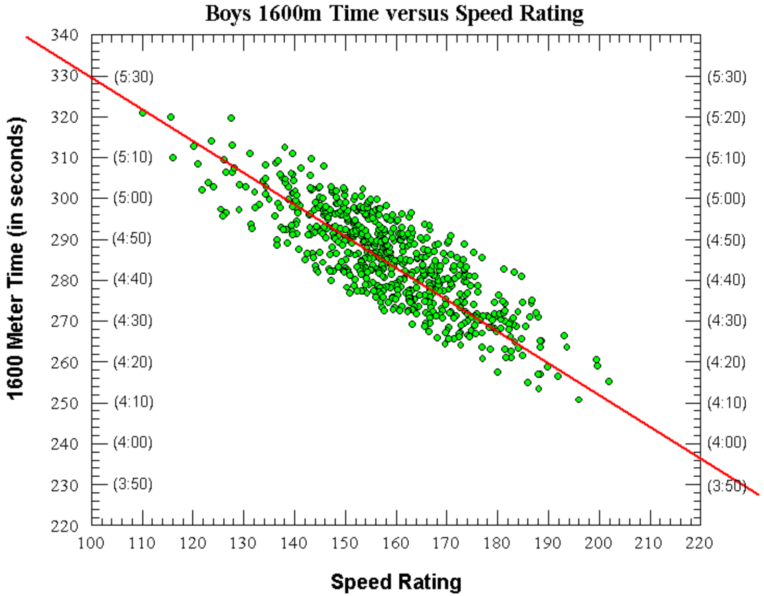
- Use an error margin "ε" to set the limit of what is an outlier. It is a probability at which everything beyond will be categorized as an outlier.

For instance, use 1% as the limit. This means that everything that has a less than 1% chance of happening is an outlier.



Outlier detection

- Calculate the mean and standard deviation. Then calculate the 99% limit. Then use the range as your classification.
- This works for each attribute independently but not when the data have a correlation.

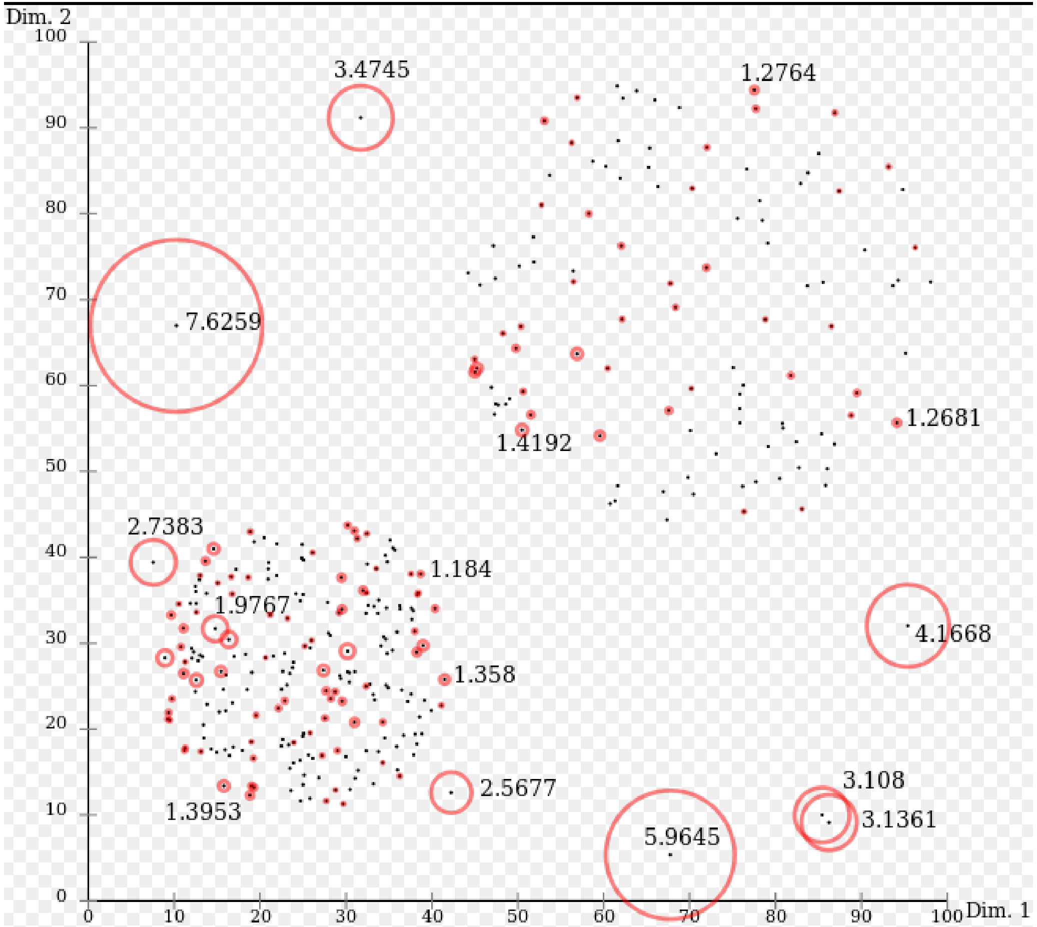


Local Outlier Factor - LOF

- Local outliers: Outliers comparing to their local neighborhood, instead of the global data distribution
- LOF: finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbors

LOF

Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set. For example, a point at a "small" distance to a very dense cluster is an outlier, while a point within a sparse cluster might exhibit similar distances to its neighbors.



Weka Demo

- Interquartile range
- LOF

Isolation Forest

- explicitly identifies anomalies instead of profiling normal data points
- built on the basis of decision trees
- partitions are created by
  - first randomly selecting a feature and
  - then selecting a random split value between the minimum and maximum value of the selected feature.
- should be identified closer to the root of the tree with fewer splits necessary.

Excel Demo

Weka Demo

References

- http://researchmining.blogspot.com/2012/10/types-of-outliers.html
- http://scikit-learn.org/stable/modules/outlier_detection.html