**Ruiz's Notes on Anomaly Detection. WPI.**

These notes are based on “Anomaly Detection”: Chapter 10 + Slides from the textbook. Use the notes below together with the figures from the textbook (provided on the book webpage) and some of the slides that come with the textbook, as described below.

- List of topics to covered on these notes: Taken from http://web.cs.wpi.edu/~cs4445/b14/Exams/topics\_exams.html#exam2

\* What's an anomaly (or outlier)?

\* Give an example of a situation in which an anomaly should be removed

during pre-processing of the dataset, and another example of a situation

in which an anomaly is an interesting data instance worth keeping and/or

studying in more detail.

\* Define each of the following approaches to anomaly detection, and

describe the differences between each pair: Model-based, Proximity-based,

and Density-based techniques.

\* Can visualization be used to detect outliers? If so, how? Give

specific examples of visualization techniques that can be used for

anomaly detection. For each one, explain whether or not the visualization

technique can be considered a Model-based (which includes Statistical),

Proximity-based, or Density-based technique for anomaly detection.

\* Define each of the following modes to anomaly detection, and

describe the differences between each pair: supervised, unsupervised,

and semi-supervised.

\* Consider the case of a dataset that has labels identifying the anomalies

and the task is to learn how to detect similar anomalies in unlabelled

data. Is that supervised or unsupervised anomaly detection? Explain.

\* Consider the case of a dataset that doesn't have labels identifying

the anomalies and the task is to find how to assign a sound anomaly

score, f(x), to each instance x in the dataset. Is that supervised or

unsupervised anomaly detection? Explain.

\* Precision, recall, and false positive rate are mentioned in your

textbook as appropriate metrics to evaluate anomaly detection algorithms

(see p. 657). What are those metrics (see Section 5.7.1, p. 295) and

how can they be used to evaluate anomaly detection?

\* For each of the anomaly detection approaches (statistical-based,

proximity-based, density-based, and clustering-based) do

1. State the definition(s) of outlier used by the approach

2. How can this be definition used to assign an anomaly score to each

data instance?

3. How does this anomaly detection approach work in general? Give an

example to illustrate your description.

-----------------------------------------------------------------------

**Approach: Statistical-based**

**Definition of Outlier:**

Probabilistic definition of outlier:

An outlier is an object that has a low probability wrt a probability

distribution model of the data.

**Anomaly score function:**

Given a data instance x from a dataset D,

f(x) = 1/P(x|D)

**How does the approach work?** (in general)

Calculate the anomaly score, f(x), for each data point in the dataset.

Use a threshold t on this score to determine outliers.

That is, x is an outlier iff f(x) > t.

- In order to figure out a good value for the threshold, one can repeat

the same idea used in clustering of sorting all data points

according to their score value, and then finding a good "elbow"

in that plot. See example on the last page of this document.

**Example**

For instance, use normal (Gaussian) distribution. Outliers are those

in the right or left tail of the distribution.

See Slides 6-8 (graphical approaches ... statistical approaches)

from Textbook slides

Remember that for normal distributions, z\_N is a constant that

tells how many standard deviations from the mean on both directions

(i.e., mean +- z\_N \* sigma) contain N% of the area under the curve.

z\_N can be found in statistical tables.

-----------------------------------------------------------------------

-----------------------------------------------------------------------

**Approach: Proximity-based**

**Definition of Outlier:**

Proximity-based definition of outlier using distance to k-nearest neighbor

**Anomaly score function:**

Given a data instance x from a dataset D and a value k,

Alternate definitions:

1. f(x) = Distance between x and its k-nearest neighbor

2. f(x) = Avg distance between x and its k-nearest neighbors

**How does the approach work?** (in general):

Calculate the anomaly score, f(x), for each data point in the dataset.

Use a threshold t on this score to determine outliers.

That is, x is an outlier iff f(x) > t.

- In order to figure out a good value for k, one can repeat the same

Idea used in clustering: Run experiments with different values of k

- In order to figure out a good value for the threshold, one can repeat

the same idea used in clustering of sorting all data points

according to their score value, and then finding a good "elbow" in

that plot.

**Example:**

See Figures 10.4-10.7 from the textbook.

-----------------------------------------------------------------------

-----------------------------------------------------------------------

**Approach: Density-based**

**Definition of Outlier:**

Outliers are instances that are in regions of low density.

Alternate definitions of Density:

1. Inverse distance: (see p.668)

Inverse of the average distance to the k nearest neighbors:

where N(x,k) is the set containing the k-nearest neighbors of x

|N(x,k)| is the size of that set

y is a nearest neighbor

2. Count of points within radius: (like in DBSCAN)

density(x,epsilon)= number of objects within epsilon distance to x.

3. Average relative density:

**Anomaly score function:**

Given a data instance x from a dataset D,

f(x) = 1/density(x,k), or

f(x) = 1/avg\_rel\_density(x,k)

**How does the approach work?** (in general):

Calculate the anomaly score, f(x), for each data point in the dataset.

Use a threshold t on this score to determine outliers.

That is, x is an outlier iff f(x) > t.

Same comments on how to determine good values for k and the threshold

as discussed above.

**Example:**

See Figure 10.8 from textbook. It uses the avg\_rel\_density.

LOF: Local Outlier Factor

Points A, C, and D have the largest anomaly scores:

C: the most extreme outlier

D: the most extreme point wrt the compact set of points

A: the most extreme point wrt the loose set of points

-----------------------------------------------------------------------

-----------------------------------------------------------------------

**Approach: Clustering-based**

**Definition of Outlier:**

Clustering-based definition of outlier:

A data instance is a cluster-based outlier if the instance does not

strongly belong to any cluster.

**Anomaly score function:**

Given a data instance x from a dataset D,

Alternate definitions:

1. f(x) = distance between x and its closest centroid

2. f(x) : (called relative distance)

= ratio between the point's distance from the centroid to the

median distance of all points in the cluster from the

centroid.

3. f(x) = improvement in the goodness of a cluster (as

measured by an objective function) when x is removed.

**How does the approach work?** (in general):

Calculate the anomaly score, f(x), for each data point in the dataset.

Use a threshold t on this score to determine outliers.

That is, x is an outlier iff f(x) > t.

Same comments on how to determine good values for k and the threshold

as discussed above.

**Example:**

See Figures 10.9-10.10 from textbook, using K-means with 2 clusters.

Fig. 10.9 uses distance of point from closest centroids

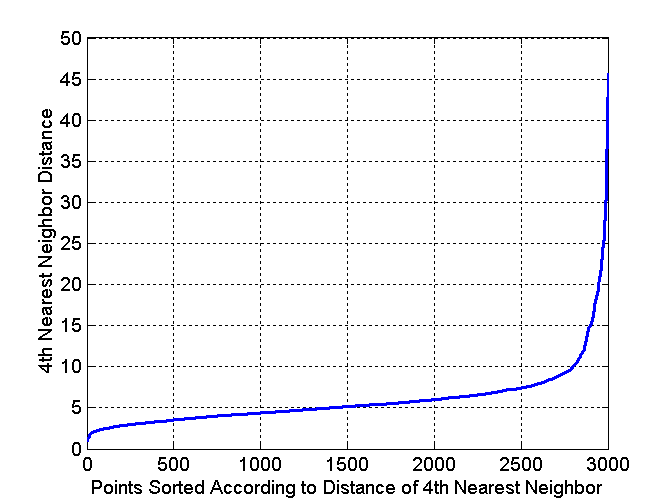
(D is not considered outlier)

Fig. 10.10 uses relative distance of point from closest centroids

to adjust for the difference of densities among the clusters.

-----------------------------------------------------------------------

Assume that we plot below depicts the anomaly scores of the data instances, sorted in increasing order.



Data instances sorted in increasing order of *f(x)*

Anomaly score *f(x)*

1. What would be a natural choice for the value of this threshold *t* based on the plot above? Explain your answer. Mark your chosen threshold value on the y-axis of the plot and label it “*t1*”. In this case, how many data instances (more or less) would be classified as anomalies?
2. This question is unrelated to question 1 above. Assume that we want to classify 20% of the dataset instances as anomalies. In this case, what threshold value would you pickbased on the plot above? Explain your answer. Mark your chosen threshold value on the y-axis of the plot and label it “*t2*”.