Anomaly Detection Challenge

Chair of IT Security, TU München

# Homework 1

One important part of anomaly detection is to be able to estimate a probability distribution around a set of points. This is called "density estimation". If we get a new point, we should be able to estimate the probability that it belongs to the same distribution. A simple method is to draw a (multivariate) Gaussian distribution around each point, add those Gaussians and normalize by the number of points (see picture in slides). In essence this is the "Kernel Density Estimation" with a Gaussian kernel. However, is this the optimal solution? As we have seen in class, if data is on a lower dimensional manifold, we should be able to do better and look at the number of neighbours a new point has.

Your assignment is to try to find a way to utilize the neighbouring information, implement it with numpy and demonstrate it using a dataset of your choice. Some proposals for datasets are below:

> http://www.cnrl.colostate.edu/Projects/VCS/Dataset_Spiral.htm

> https://cs.joensuu.fi/sipu/datasets/spiral.txt (ignore labels)

> http://homepage.tudelft.nl/n9d04/occ/index.html (ignore labels)

Send solutions (code) to kolosnjaji@sec.in.tum.de and describe in your e-mail shortly how you did it. Maybe add matplotlib plots to make it more convincing.

Have fun!