Open Thesis Topics

Simon.Kluettermann@cs.tu-dortmund.de

ls9 tu Dortmund

27. September 2022

Simon Kluettermann

Thesis@ls9

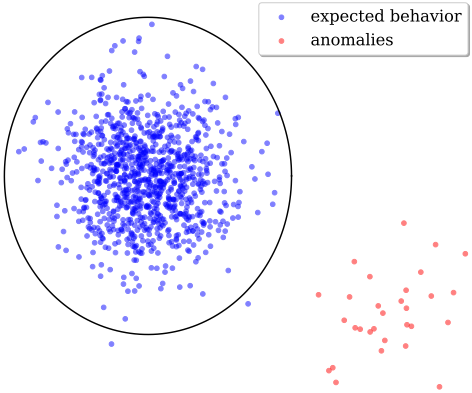
- First: Find a topic and a supervisor
- Work one month on this, to make sure
  - you still like your topic
  - and you are sure you can handle the topic
- then short presentation in front of our chair (15min, relaxed)
  - get some feedback/suggestions
- afterwards register the thesis
  - (different for CS/DS students)
- Problem: We are not able to supervise more than 2 students at the same time (CS faculty rules)

Today

- First: A short summary of each Topic
- Then time for questions/Talk with your supervisor about each topic that sounds interesting
- Your own topics are always welcome;)

Anomaly Detection

- Im working on Anomaly Detection
- That means characterising an often very complex distributions, to find events that dont match the expected distribution



knn

- kNN algorithm can also be used for AD
- if the k closest point is further away, a sample is considered more anomalous
- $r = \frac{k}{2N \cdot pdf}$
- Powerful method, as it can model the pdf directly

Better knn

- The model (mostly) ignores every known sample except one
- So there are extensions
- $avg = \frac{1}{N} \sum_i knn_i(x)$
- $wavg = \frac{1}{N} \sum_i \frac{knn_i(x)}{i}$

Dataset	wavg	avg	1	3	5
vertebral	**0.4506**	**0.4506**	**0.4667**	**0.4667**	**0.45**
...					
thyroid	**0.9138**	**0.9151**	**0.8763**	**0.9086**	**0.914**
Iris_setosa	**0.9333**	**0.9333**	**0.9333**	**0.9**	**0.9**
breastw	**0.9361**	**0.9361**	**0.9211**	**0.9248**	**0.9286**
wine	**0.95**	**0.95**	**0.9**	**0.95**	**0.95**
pendigits	**0.9487**	**0.9487**	**0.9391**	**0.9295**	**0.9359**
segment	**0.9747**	**0.9747**	**0.9495**	**0.9545**	**0.9394**
banknote − authentication	**0.9777**	**0.9776**	**0.9408**	**0.943**	**0.9583**
vowels	**0.9998**	**0.9972**	**0.99**	**0.92**	**0.93**
Ecoli	**1.0**	**1.0**	**0.9**	**1.0**	**1.0**
Average	**0.7528**	**0.7520**	0.7325	0.7229	0.7157

What to do?

- Evaluation as anomaly detector is complicated
  - Requires known anomalies
- $\Rightarrow$So evaluate as density estimator
  - Does not require anomalies
  - Allows generating infinite amounts of training data

What to do?

- Collect Extensions of the oc-knn algorithm
- Define some distance measure to a known pdf
- Generate random datapoints following the pdf
- Evaluate which algorithm finds the pdf the best

Requirements

- Knowledge of python (sum([i for i in range(5) if i%2]))
  - Ideally incl numpy
- Basic university level Math (you could argue that $r_k \propto \frac{k}{pdf}$)
- Ideally some experience working on a ssh server
- $\Rightarrow$Good as a Bachelor Thesis
- For a Master Thesis, I would extend this a bit (Could you also find k?)

Normalising Flows

- Deep Learning Method, in which the output is normalised
- $\int f(x)dx = 1 \; \forall f(x)$
- Can be used to estimate probability density functions
- $\Rightarrow$Thus useful for AD
- $\int f(h(x))\|\frac{\delta x}{\delta h}\|dh = 1 \; \forall h(x)$

Graph Normalising Flows

- How to apply this to graphs?
- One Paper (Liu 2019) uses two NN:
- Autoencoder graph⇒vector
- NF on vector data
- which is fine, but also not really graph specific
- No interaction between encoding and transformation

Graph Normalising Flows

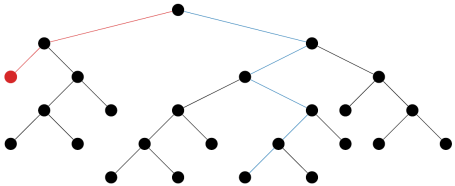
- So why not do this directly?
- $\Rightarrow$Requires differentiating a graph
- Why not use only one Network?
- Graph$\Rightarrow$Vector$\Rightarrow$pdf
- $\Rightarrow$Finds trivial solution, as $< pdf > \propto \frac{1}{\sigma_{Vector}}$
- So regularise the standart deviation of the vector space!
  - Interplay between encoding and NF
  - Could also be useful for highdim data

Requirements

- Proficient in python ([i for i in range(1,N) if not [j for j in range(2,i) if not i%j]])
  - Ideally incl numpy, tensorflow, keras
- Some deep learning experience
- University level math (google Cholesky Decomposition. Why is this useful for NF?)
- Ideally some experience working on a ssh server
- A bit more challenging⇒Better as a Master thesis
- (Still we would start very slowly of course)

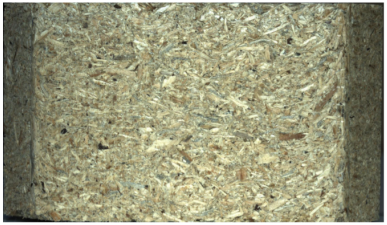
Sina

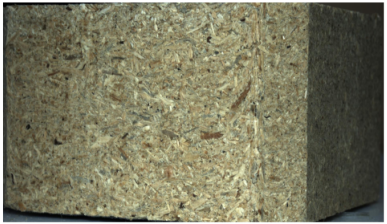
- Isolation Forest: Different Anomaly Detection Algorithm
- Problem: Isolation Forests dont work on categorical data
- ⇒Extend them to categorical data



Britta

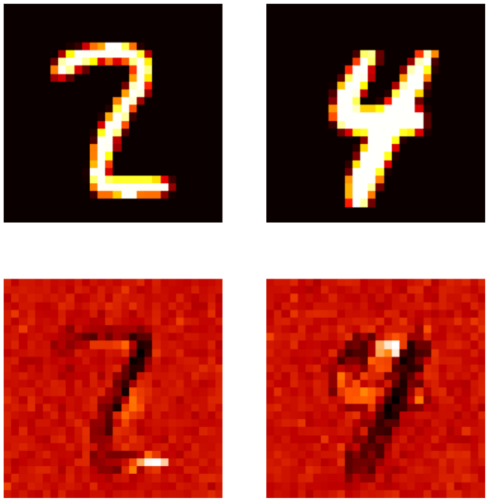
- Reidentification: Find known objects in new images
- Task: Find if two images of pallet blocks are of the same pallet block
- Use AD to represent the pallet blocks





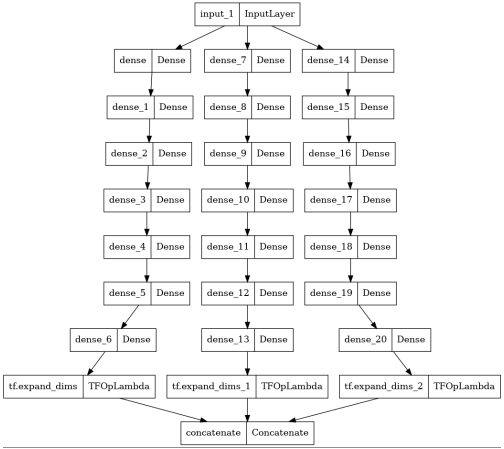
Hsin Ping

- Ensemble: Combination of multiple models
- Task: Explain the prediction of a model using the ensemble structure



Nikitha

- Task: Explore a new kind of ensemble
- Instead of many uncorrelated models, let the models interact during training



Questions?