# Logs

- 31st of October 2018:

  Started reading "Isolation Forest" by Liu, Ting and Zhou, a paper where they describe the Isolation Forest algorithm. The way the algorithm uses an ensemble of "iTrees" which randomly partition the feature space to isolate outliers is very interesting for me because I think I could use the concept not to isolate outliers but partitions which are part of $P_s$.[LTZ08] Instead of the "iTree" structure I want to implement a structure based on a kd-tree.

- 1st of November 2018:

  Started working on a random dataset generator that generates a two dimensional dataset containing "clean" patches, meaning areas where data points all belong to the same class.

- 6th of November 2018:

  Finished the random dataset generator generating $D$. The generator generates a two dimensional normalized set of points (normalized meaning: $\forall P(x, y) \in D : x, y \in [0, 1)$). The points are distributed uniformally, so every possible value on the plane has the same chance of being selected.

  The generator takes as input:

  - The amount of points $D$ should contain ($|D|$).
  - How much percent of the area should be "clean". The idea I have is that, while a uniformally distributed dataset is impossible to classify correctly since it is random, the dataset the generator builds contains "patches" or areas which are clean, meaning every point inside one of these patches has the same label. My goal is to find a classifier that produces a $P_s$ so the area $P_s$ covers comes as close to the percent of clean points as possible.
  - The amount of patches. The clean points are further separated in different patches, all randomly set into the plane. $P_s$ should be coextensive with all patches.
  - The seed for the random number generator, so results can be compared.

  Figure 1 shows a dataset generated by the random dataset generator.

- 8th of November 2018:

  Started implementing the kd-tree.

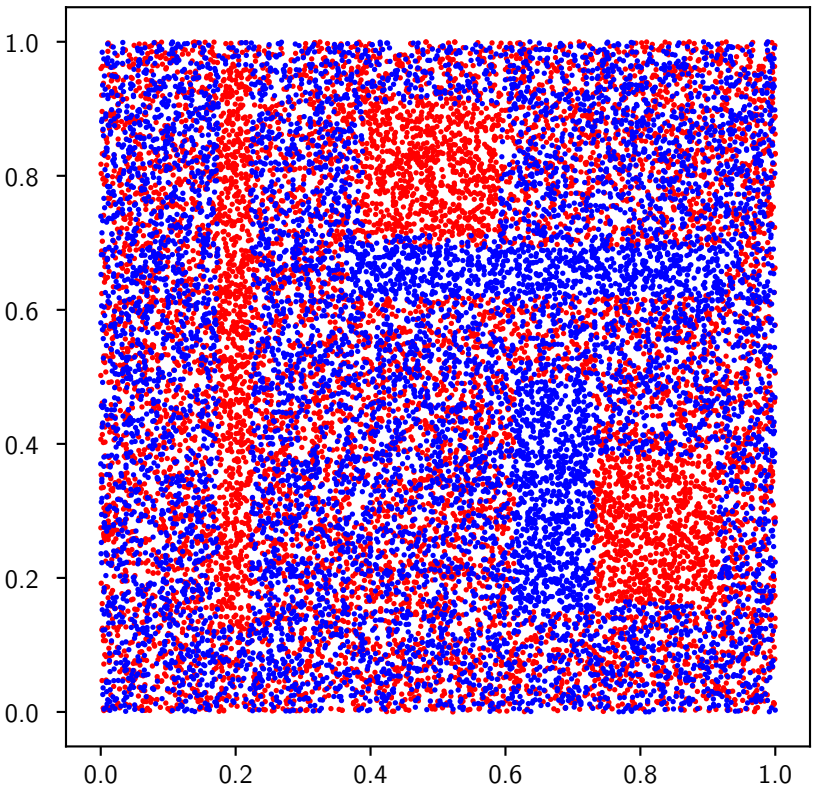


Figure 1: A dataset generated by the random dataset generator. It contains 20000 data points from which 20 percent are clean. The clean points are distributed in 5 patches. The seed was 42.

# References

[LTZ08] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08, pages 413–422, Washington, DC, USA, 2008. IEEE Computer Society.