Logs

- 31st of October 2018:

  Started reading "'Isolation Forest"' by Liu, Ting and Zhou, a paper where they describe the Isolation Forest algorithm.

- 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 points 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 that P_s 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.

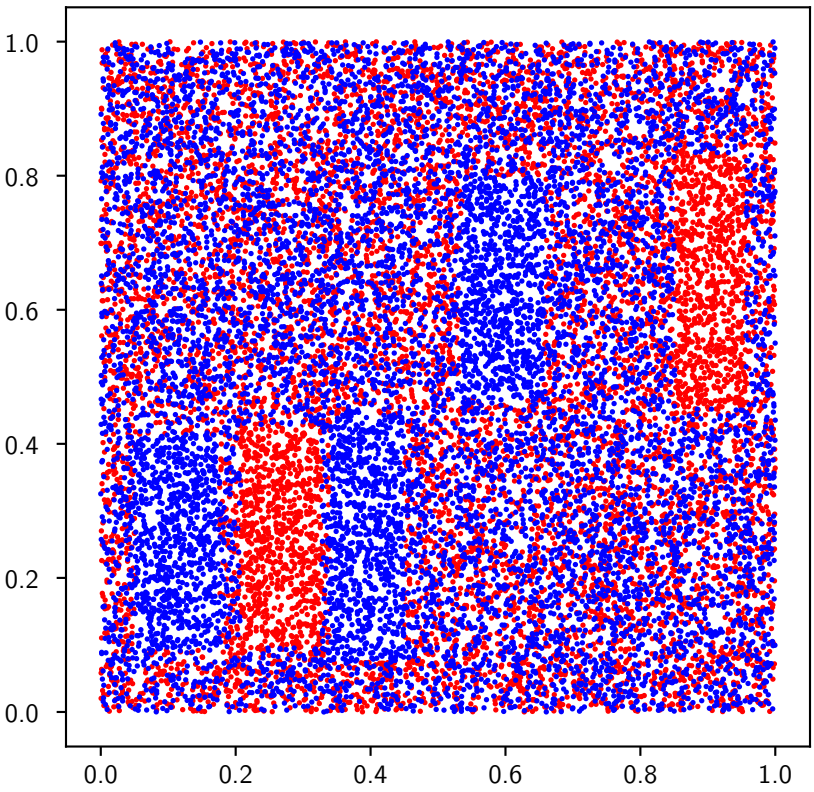

Figure 1: Dataset