Anomaly Detection

A bare minimal description

Outlier detection

When you want to detect deviant observations in your training data.

Aplications

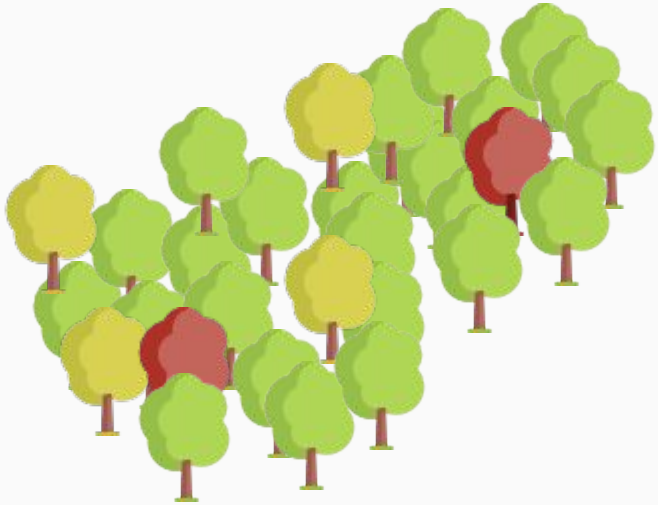
- You have a bunch of data points and want to identify sceptional data points.
- You want to remove strange data points as a preprocessing steps to avoid overfitting

Outlier detection

When you want to detect deviant observations in your training data.

Aplications

- You have a bunch of data points and want to identify sceptional data points.
- You want to remove strange data points as a preprocessing steps to avoid overfitting



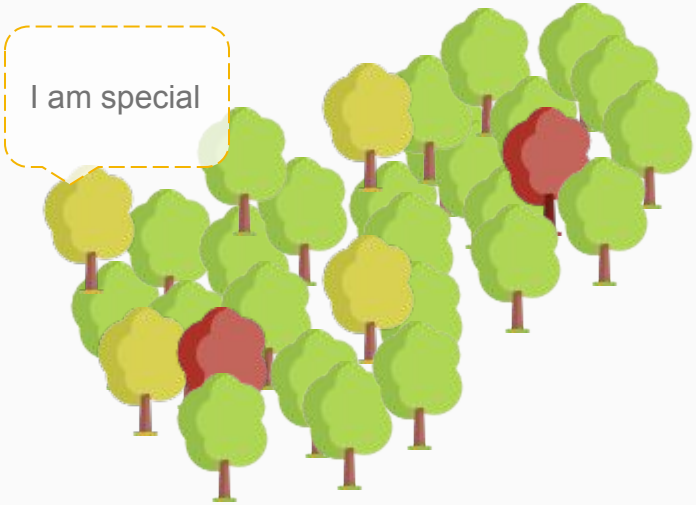
I need to find the special trees!

Outlier detection

When you want to detect deviant observations in your training data.

Aplications

- You have a bunch of data points and want to identify sceptional data points.
- You want to remove strange data points as a preprocessing steps to avoid overfitting



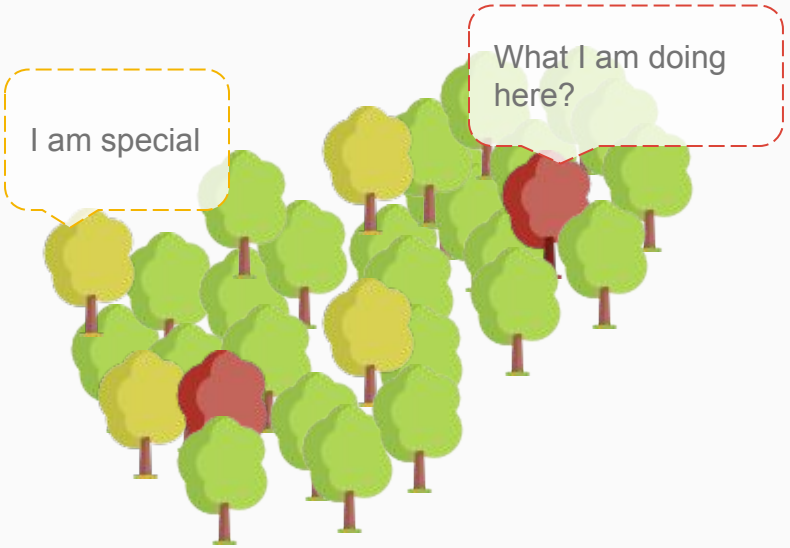
I need to find the special trees!

Outlier detection

When you want to detect deviant observations in your training data.

Aplications

- You have a bunch of data points and want to identify sceptional data points.
- You want to remove strange data points as a preprocessing steps to avoid overfitting



I need to find the special trees!

Novelty detection

You have a bunch of regular observations and you want to detect new observations if they come.

Aplications

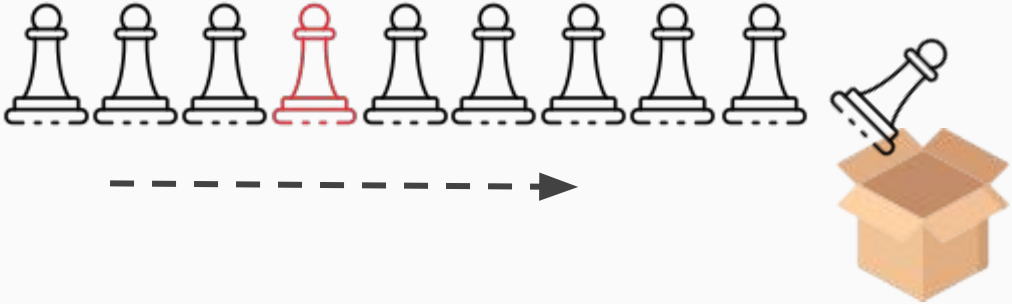
- Health monitoring

Novelty detection

You have a bunch of regular observations and you want to detect new observations if they come.

Aplications

- Health monitoring



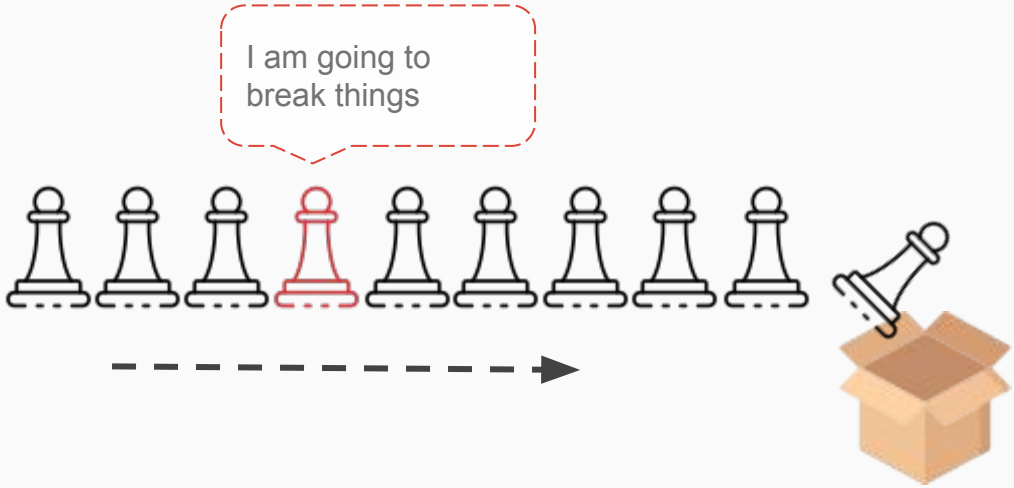
Beware of whatever strange is coming

Novelty detection

You have a bunch of regular observations and you want to detect new observations if they come.

Aplications

- Health monitoring



Beware of whatever strange is coming

Do you trust you have a dataset of **clean** data points? → **Novelty detection**

Do you want to find **exceptional** data points in your current dataset? → **Outlier detection**

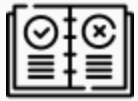
Do you want to check if anything is wrong with **new** data points? → **Novelty detection**

Outlier detection vs novelty detection

Methods

As many as you want, really.

If you already know how anomalous examples look like, feel free to hard code that.

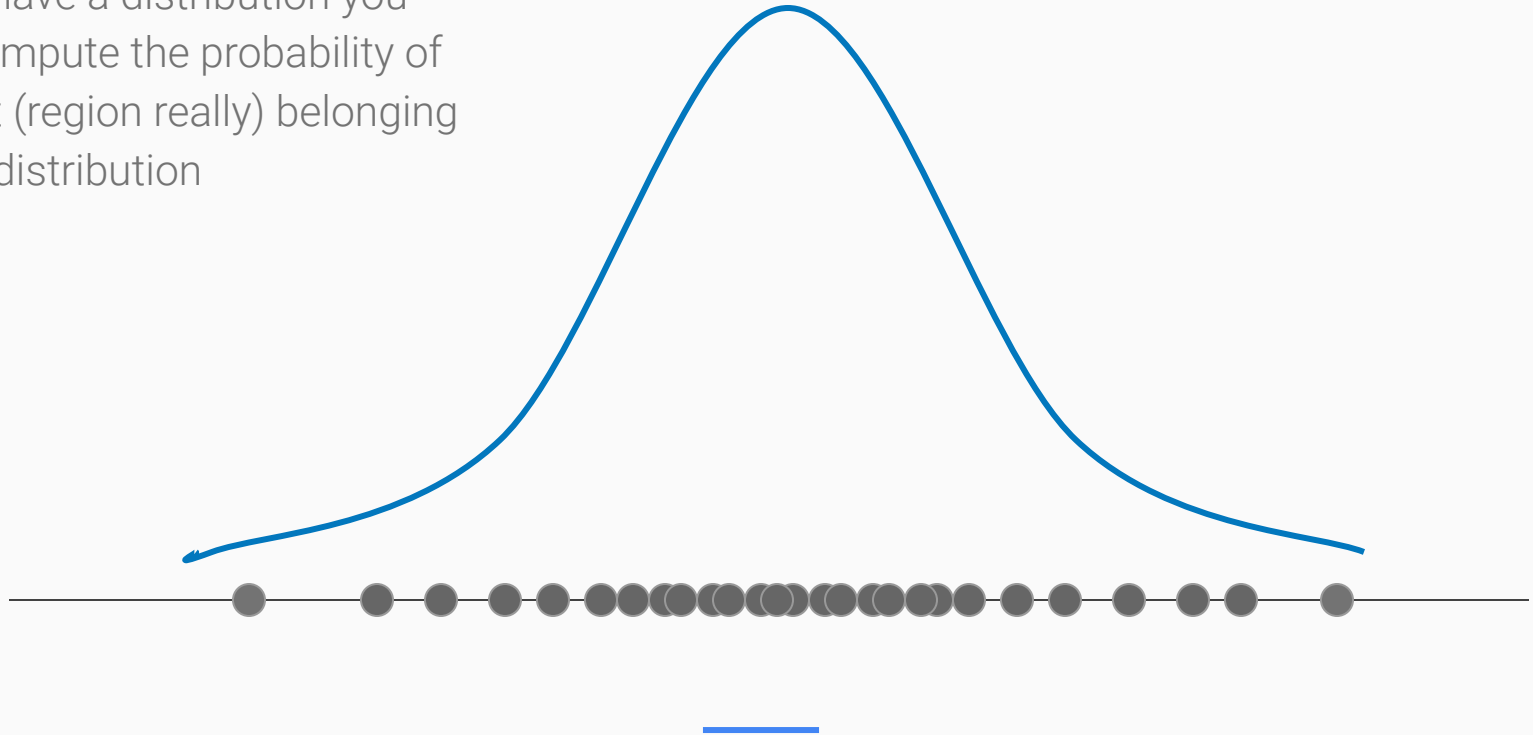


Rule based (no ML, but may still work)

If you have a distribution you can compute the probability of a point (region really) belonging to the distribution

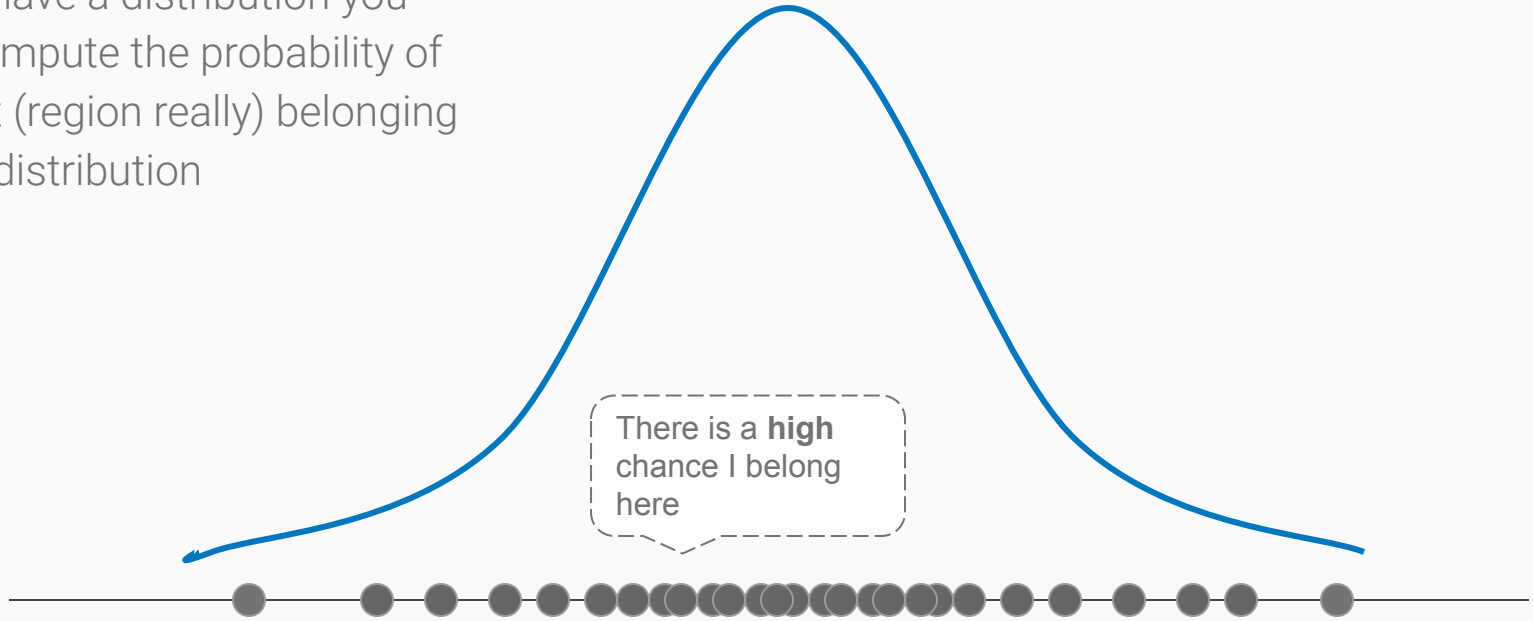
Distribution Based

If you have a distribution you can compute the probability of a point (region really) belonging to the distribution



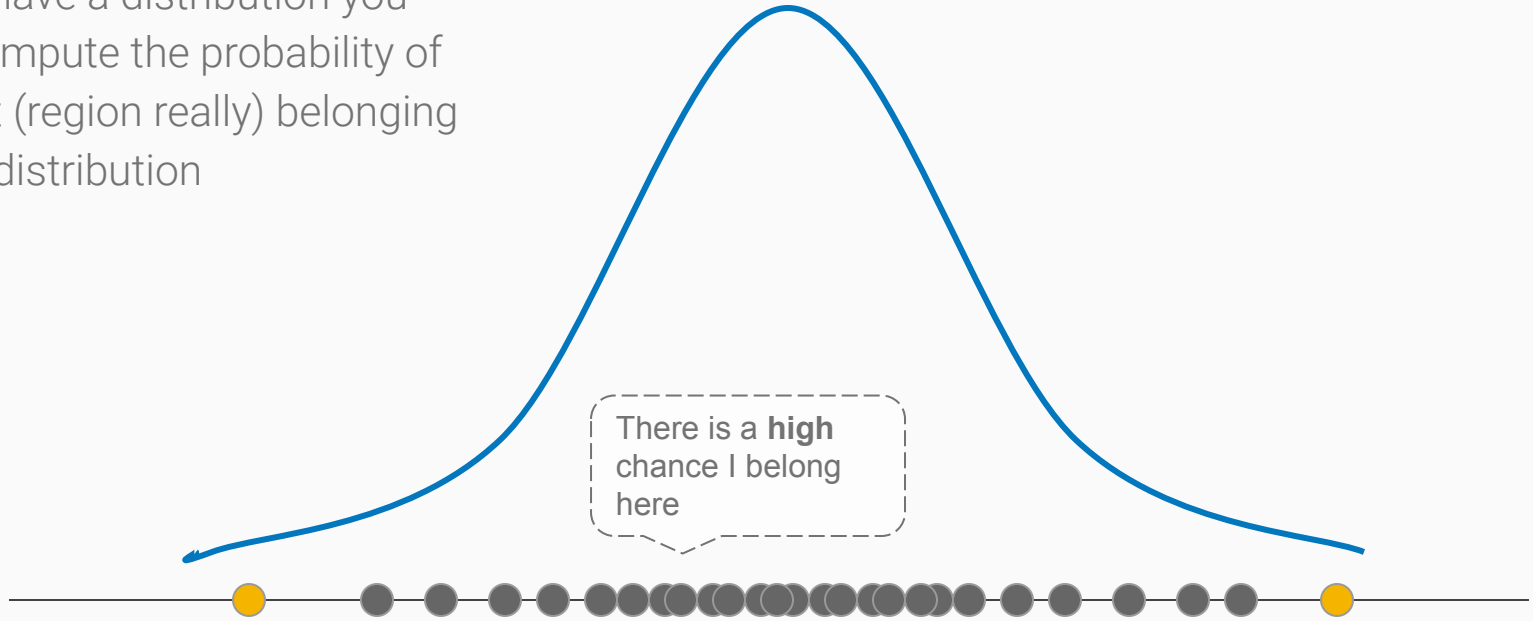
Distribution Based

If you have a distribution you can compute the probability of a point (region really) belonging to the distribution



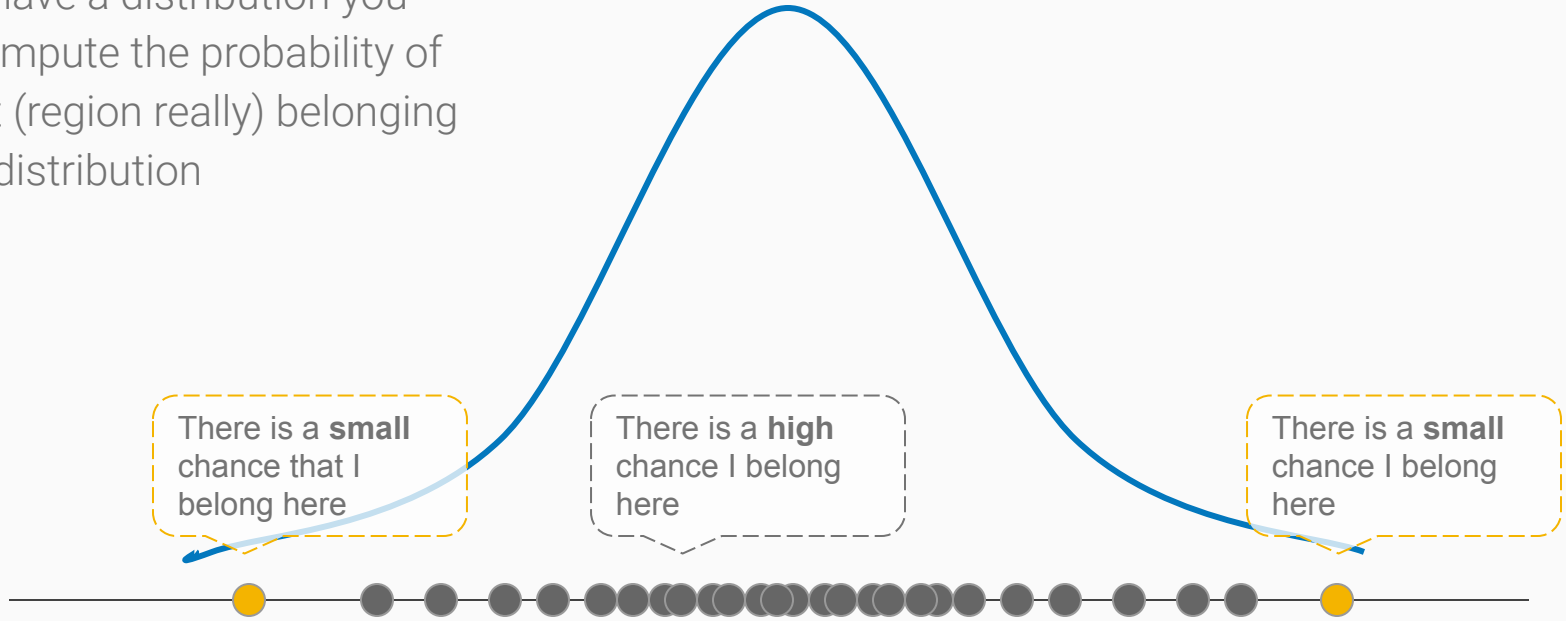
Distribution Based

If you have a distribution you can compute the probability of a point (region really) belonging to the distribution



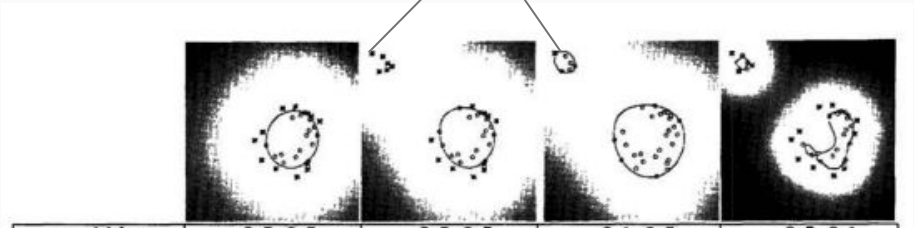
Distribution Based

If you have a distribution you can compute the probability of a point (region really) belonging to the distribution



Distribution Based

The regions are encompassed based on the hyperparameters



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|\|w\|\|$	0.84	0.70	0.62	0.48

Similar to SVM, but the support vectors are defined so that they **encompass** a region, instead of splitting regions.

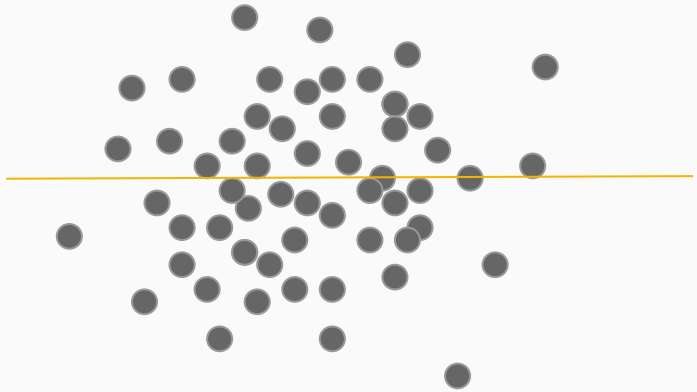
Read more if interested:
- Original paper:
https://papers.nips.cc/paper/1999/file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf
- One blog:
http://rvlasveld.github.io/blog/2013/07/12/introduction-to-one-class-support-vector-machines/
- Wikipedia:
https://en.wikipedia.org/wiki/One-class_classification#Introduction

One-Class SVM

Imagine a decision tree in which you make the partitions based on a random feature and value

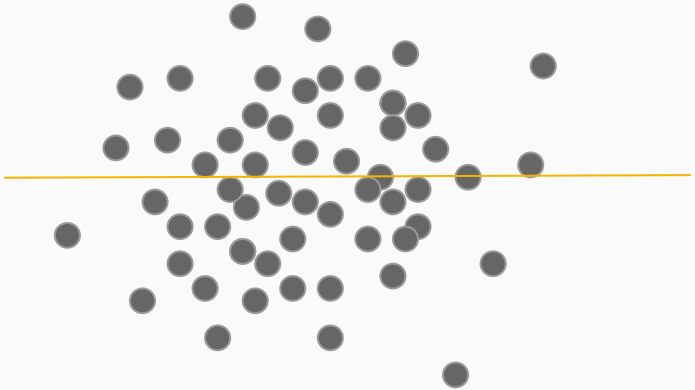
Isolation Forest

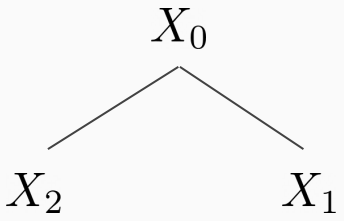
Imagine a decision tree in which you make the partitions based on a random feature and value



Isolation Forest

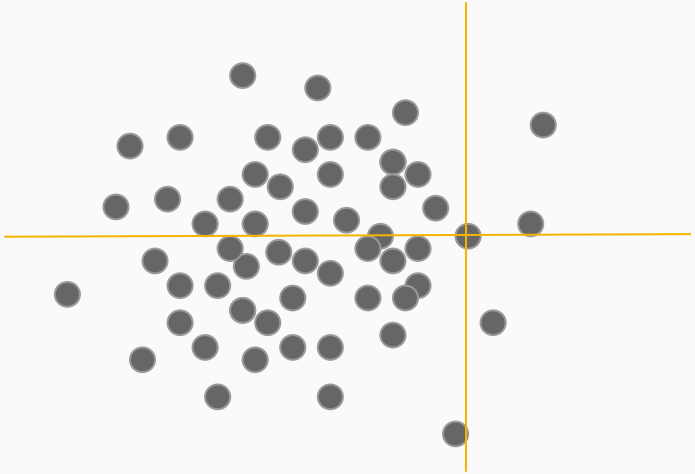
Imagine a decision tree in which you make the partitions based on a random feature and value

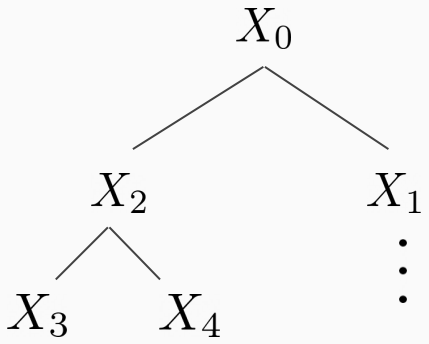




Isolation Forest

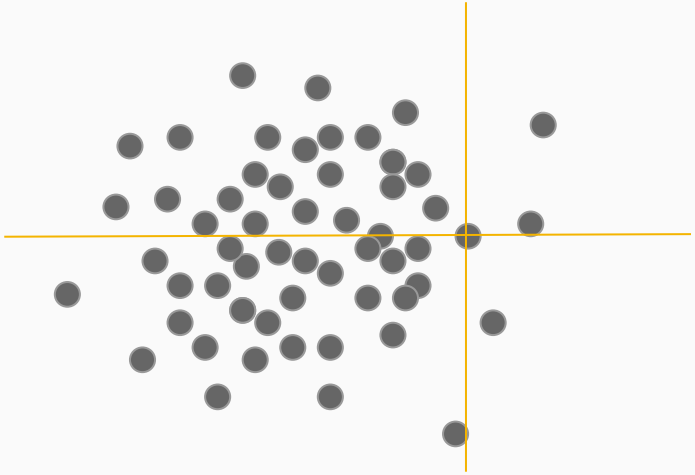
Imagine a decision tree in which you make the partitions based on a random feature and value

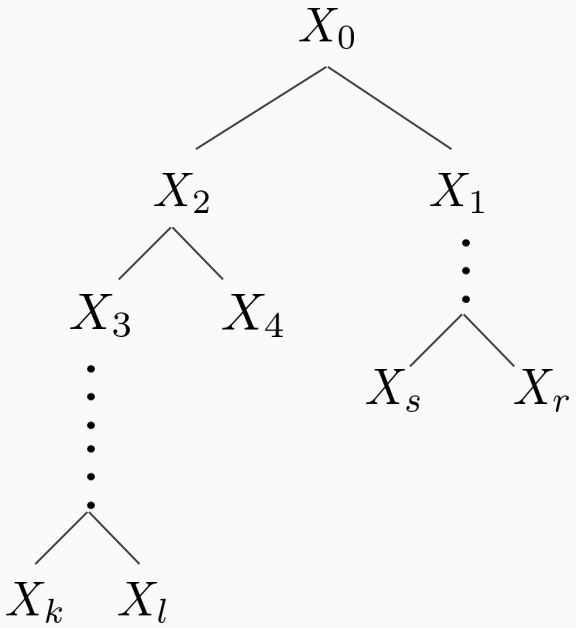




Isolation Forest

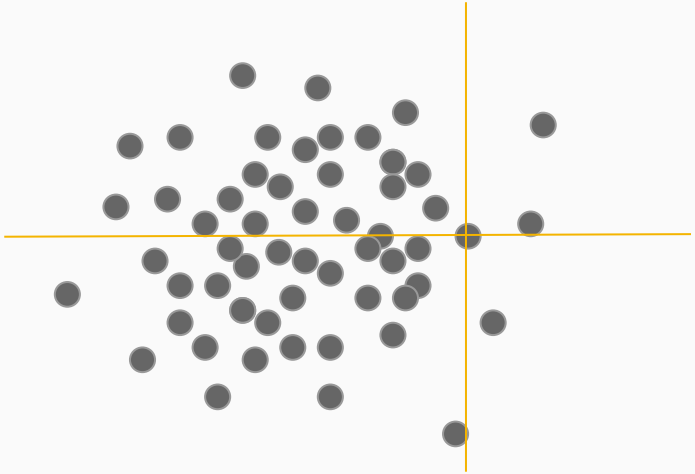
Imagine a decision tree in which you make the partitions based on a random feature and value

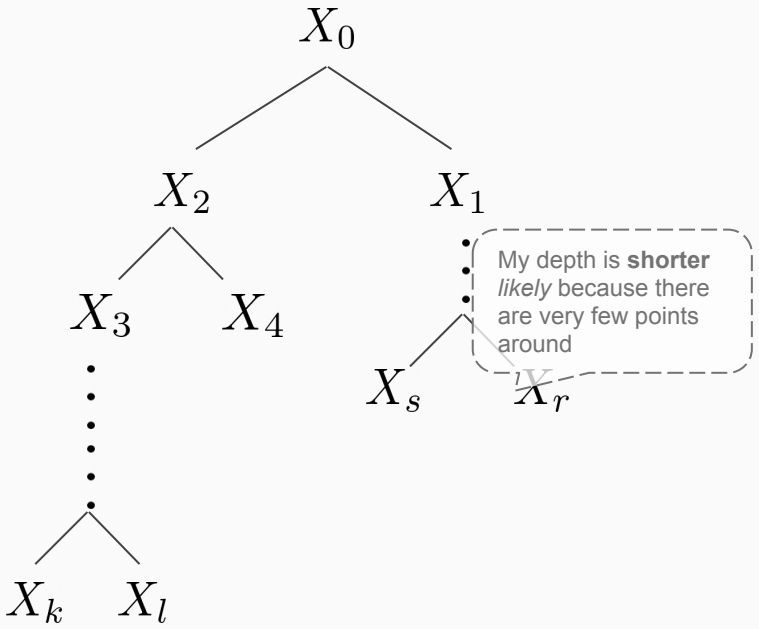




Isolation Forest

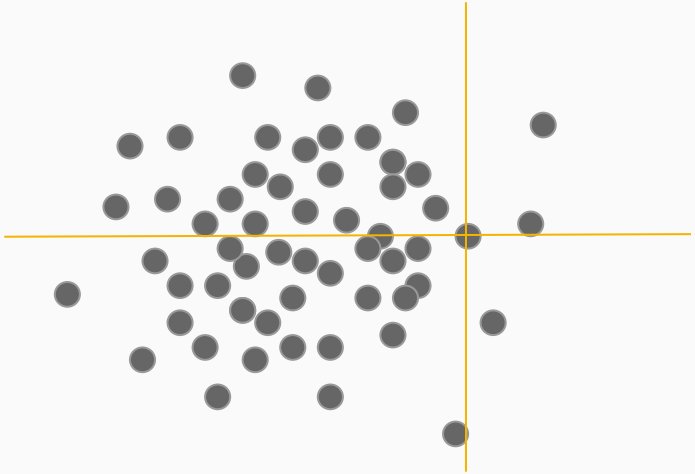
Imagine a decision tree in which you make the partitions based on a random feature and value

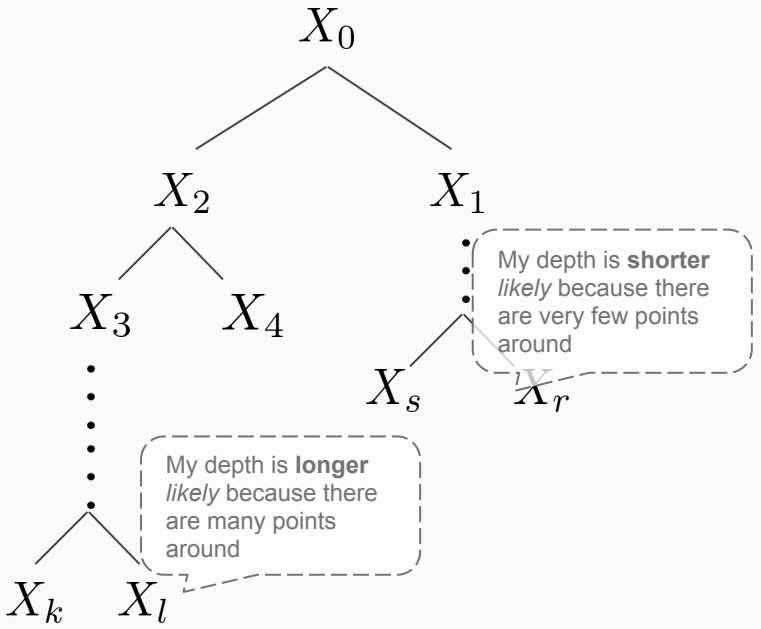




Isolation Forest

Imagine a decision tree in which you make the partitions based on a random feature and value





Isolation Forest

Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

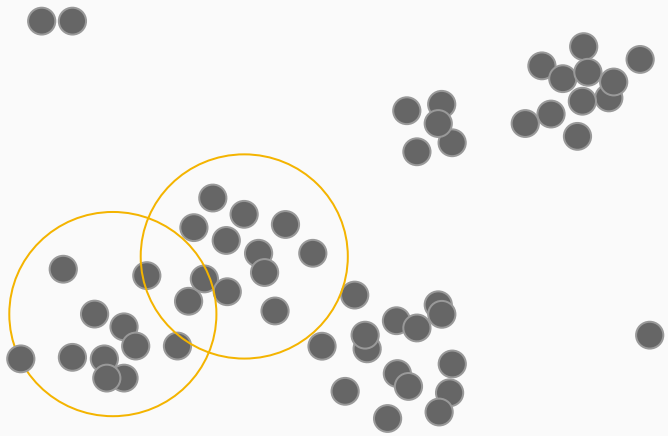
Local Outlier Factor

Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor

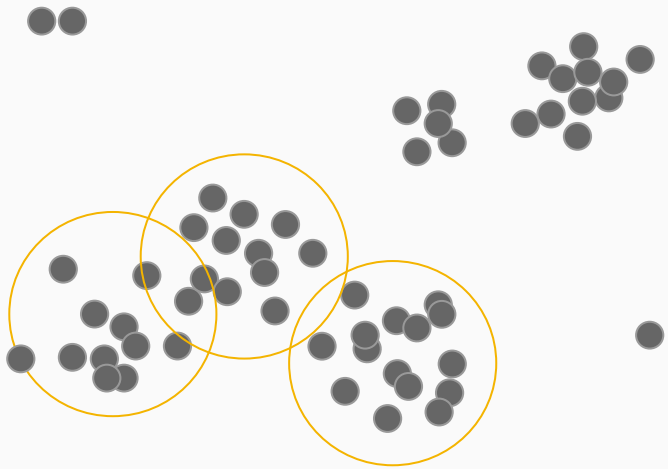
Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor



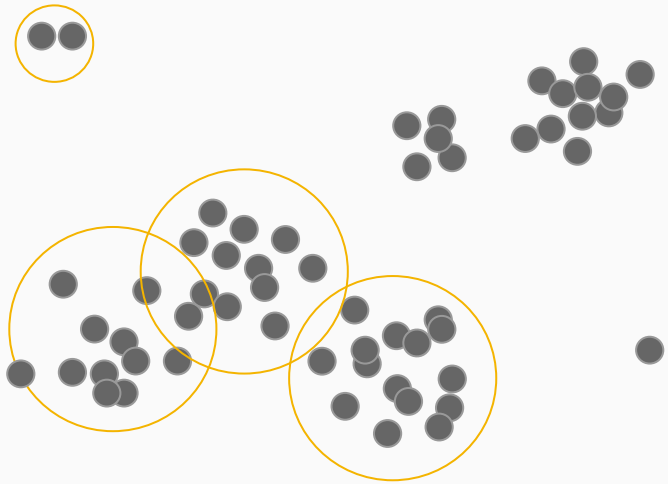
Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor



Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor



Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor

Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

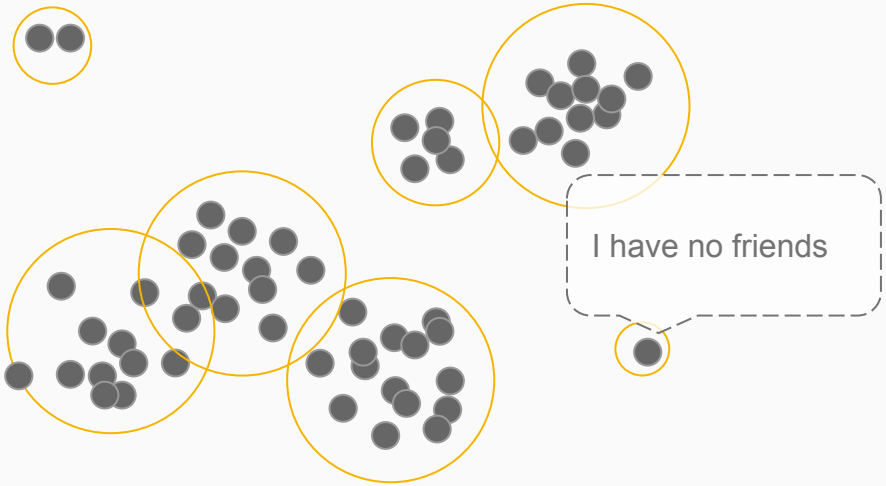
Local Outlier Factor

Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor

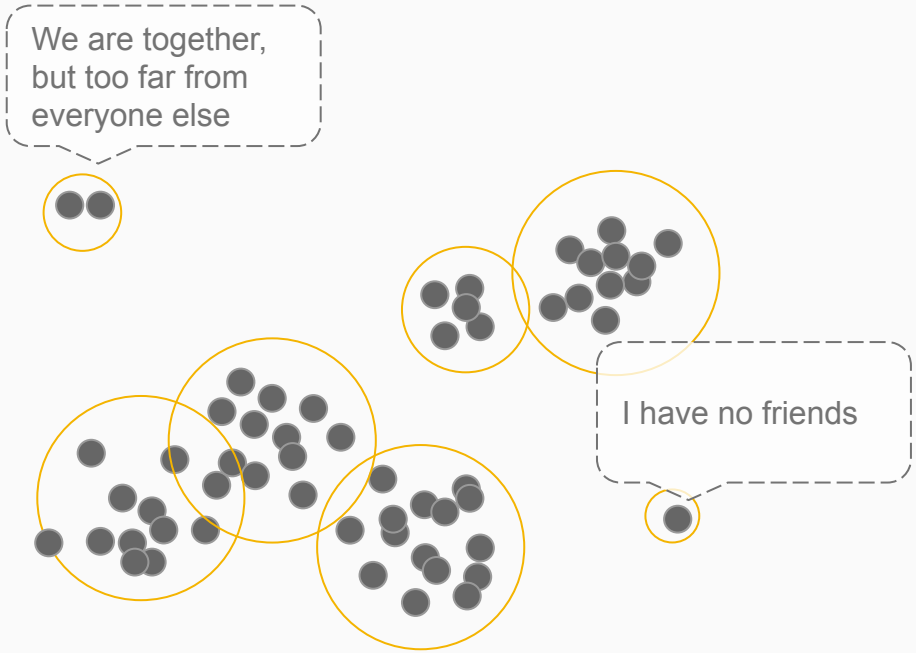
Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor



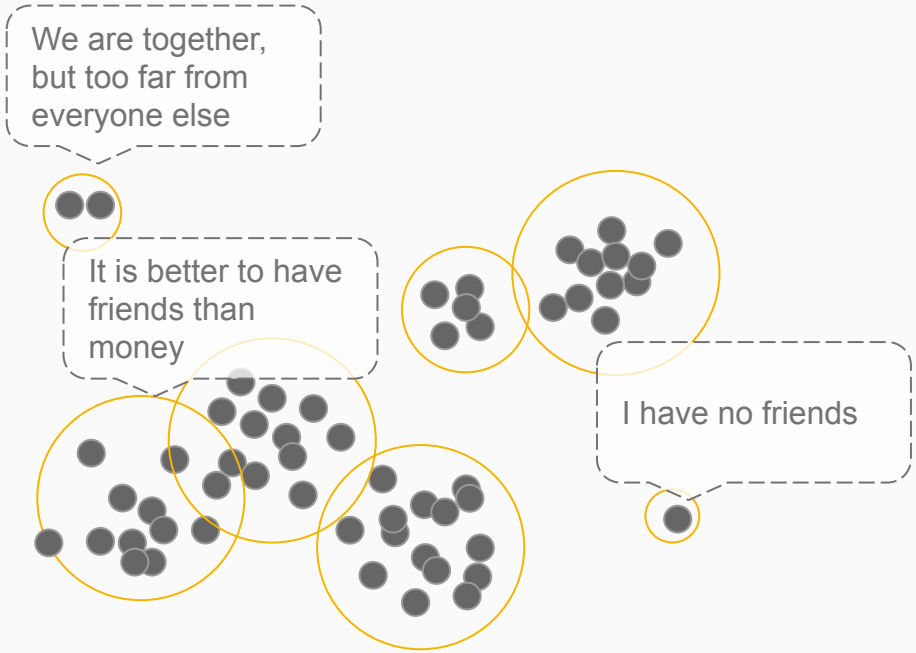
Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor



Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor

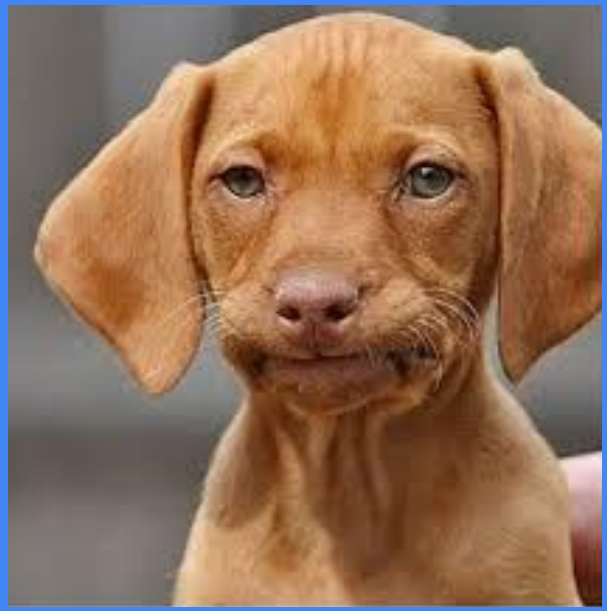


Cluster based on density and then treat as outliers the clusters that are too far from other clusters.

Local Outlier Factor

Metrics

You can’t really tell...



Maybe because you create them artificially

But...if you know some observations are outliers, you can evaluate as if it was supervised classification

- That is it

You could also define some theoretical bounds, look for *internal evaluation of anomaly detection*

- That is it

Note that you have an anomaly score from the model. You need to tune the threshold to define a sample as an anomaly, higher values give higher precision but lower recall.