Random Tree Based Anomaly Detection

국가수리과학연구소 산업수학혁신센터
최 동 헌 (dhchoe@nims.re.kr)

2020.11.06

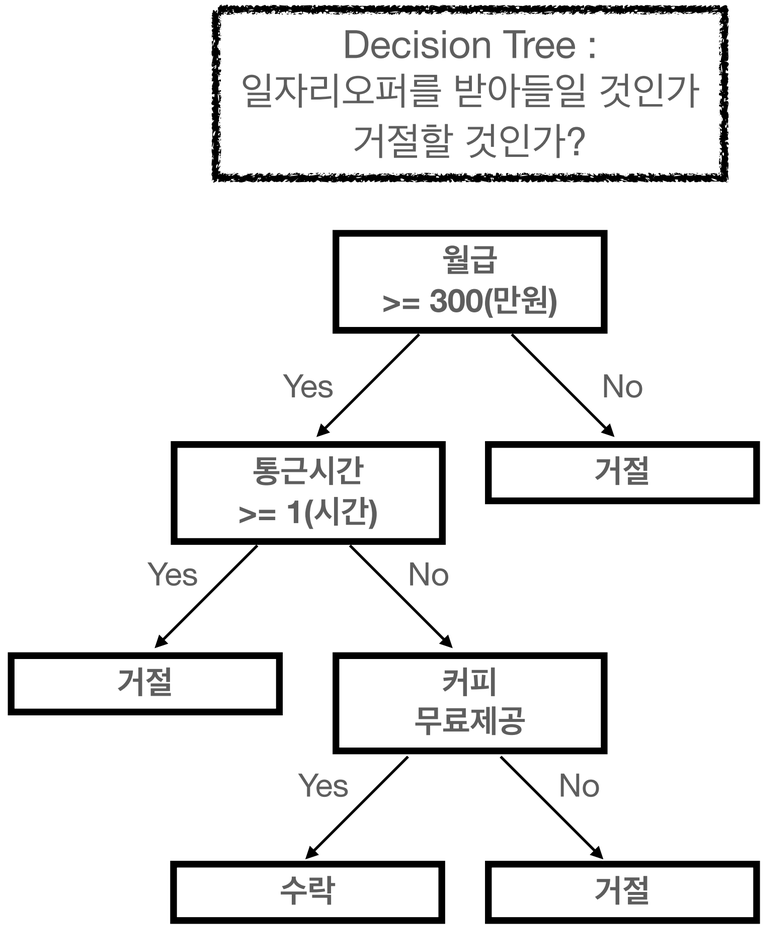
Contents

- Decision Tree Model(Classification, Regression)
- Bagging and Boosting
- Isolation based anomaly detection
  - 2-dim’l data figure and describe random tree generating process
- Isolation Forest 실습

Review

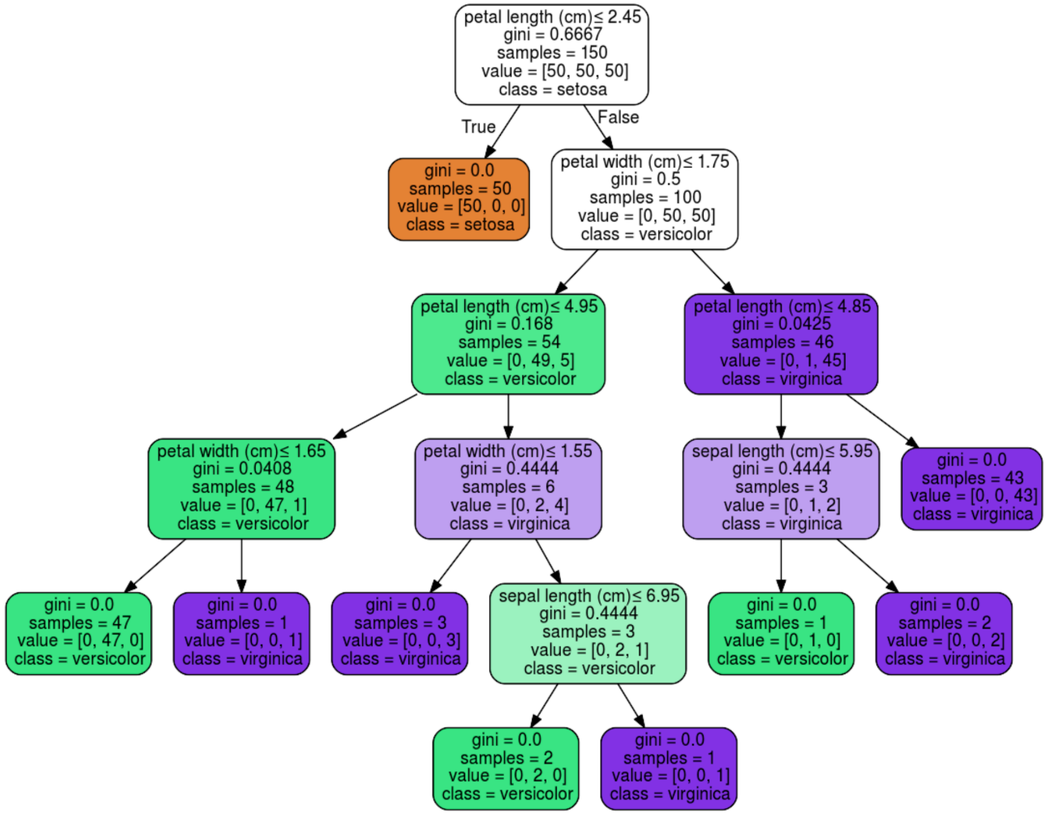
- NN-based models (Classification, Regression)
- Weakness of DB-outlier criterion
- Local Outlier Factor
- Parameters of LOF

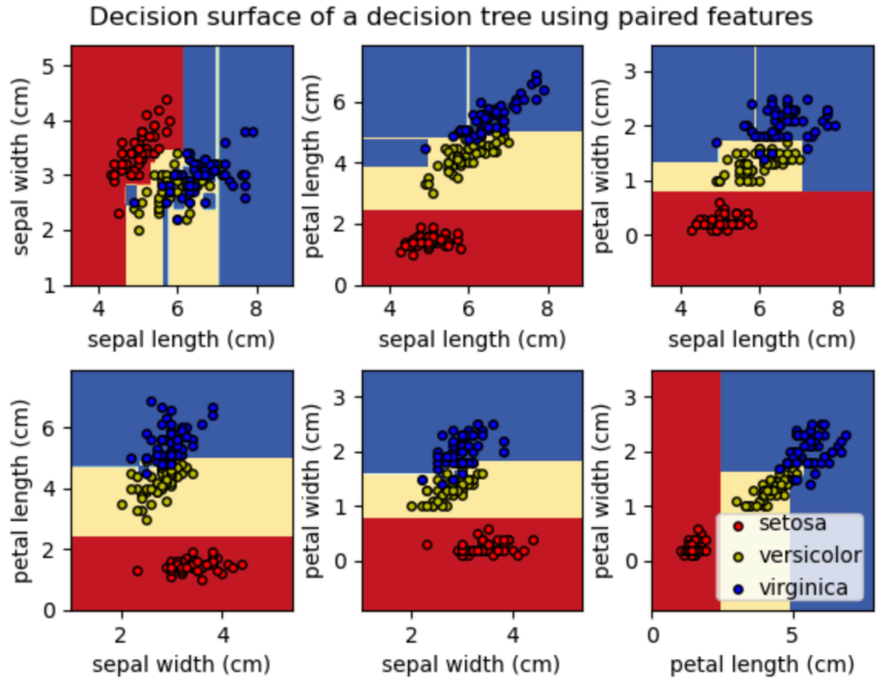
Decision Tree



- Terminology
  - Root node
  - Leaf node
  - Parent node
  - Child node

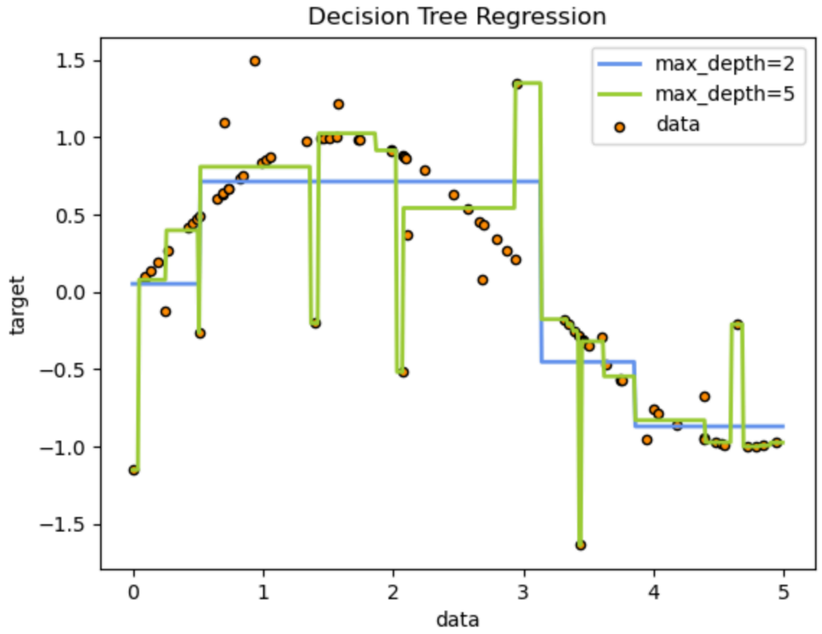
Decision Tree (Classification)

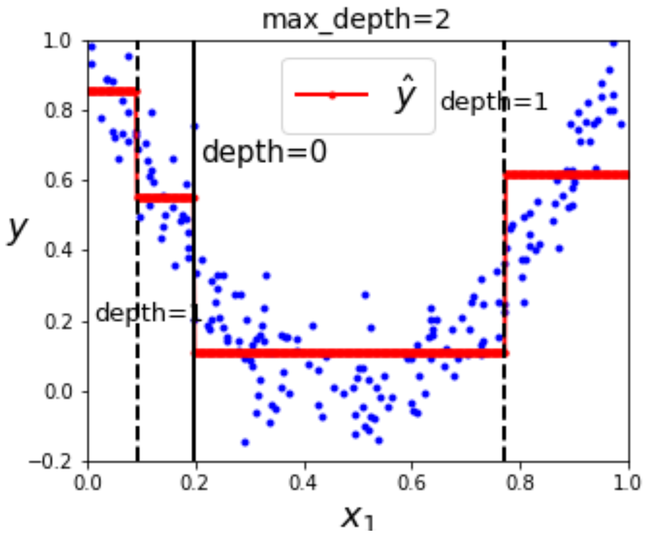


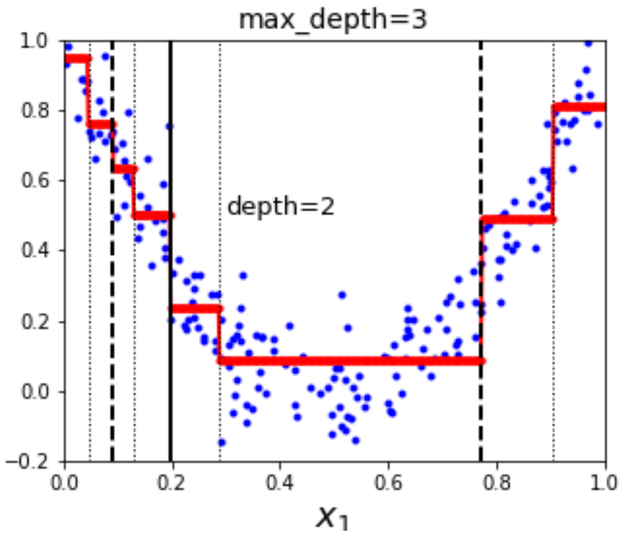


https://scikit-learn.org/stable/modules/tree.html

Decision Tree (Regression)







https://scikit-learn.org/stable/modules/tree.html

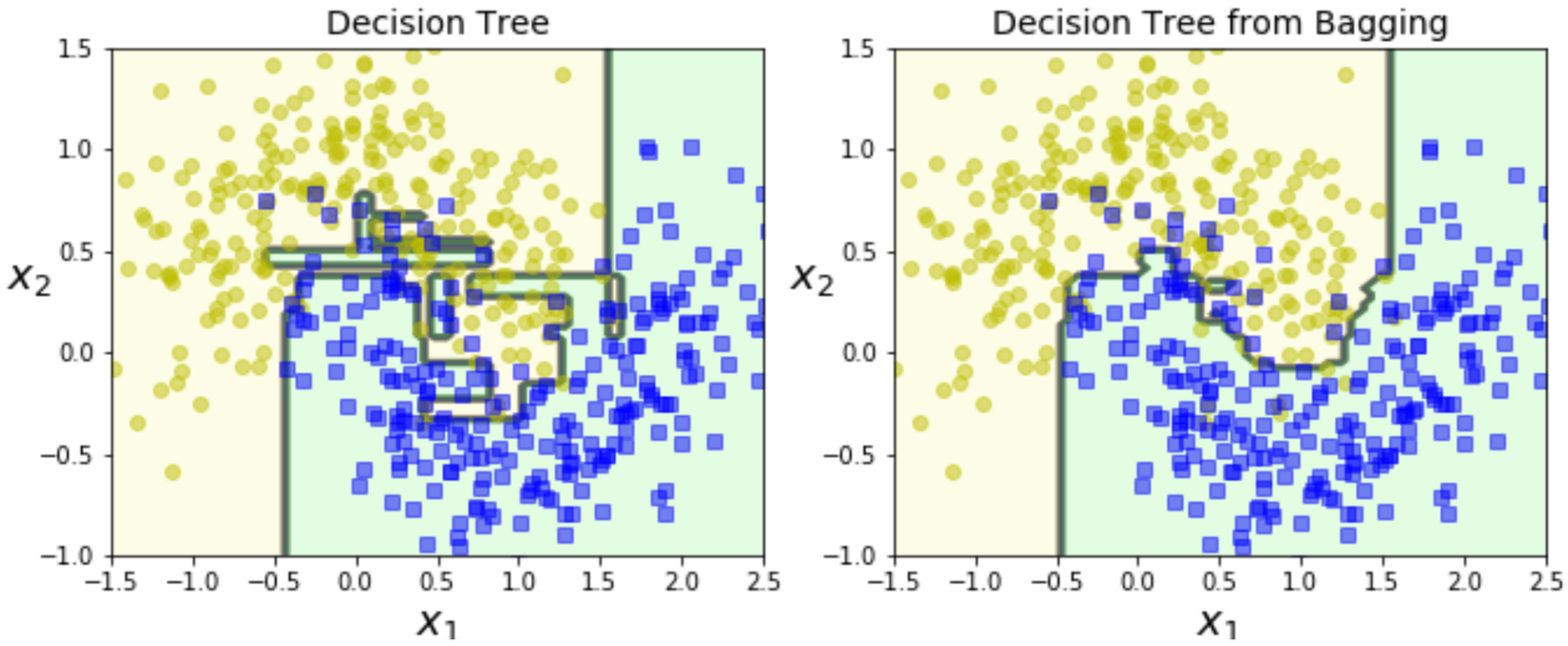
Bagging and Boosting

- Bootstrap Aggregating ensemble
- Combining weak learners in parallel
- Reduce variance (overfitting issue)
- Random sampling with replacement

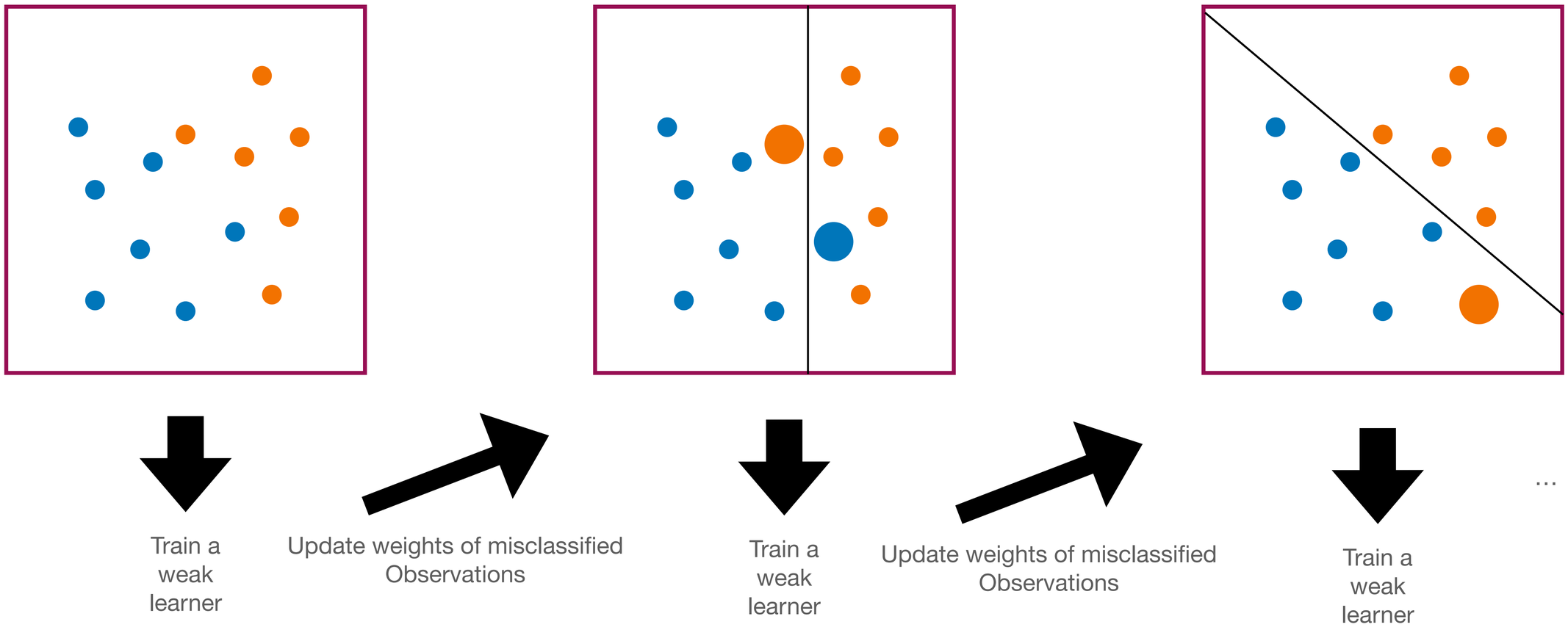
- Boosting ensemble
- Fit a sequence of weak learners
- Reduce bias (underfitting issue)
- Weights of incorrectly classified instances

Bagging

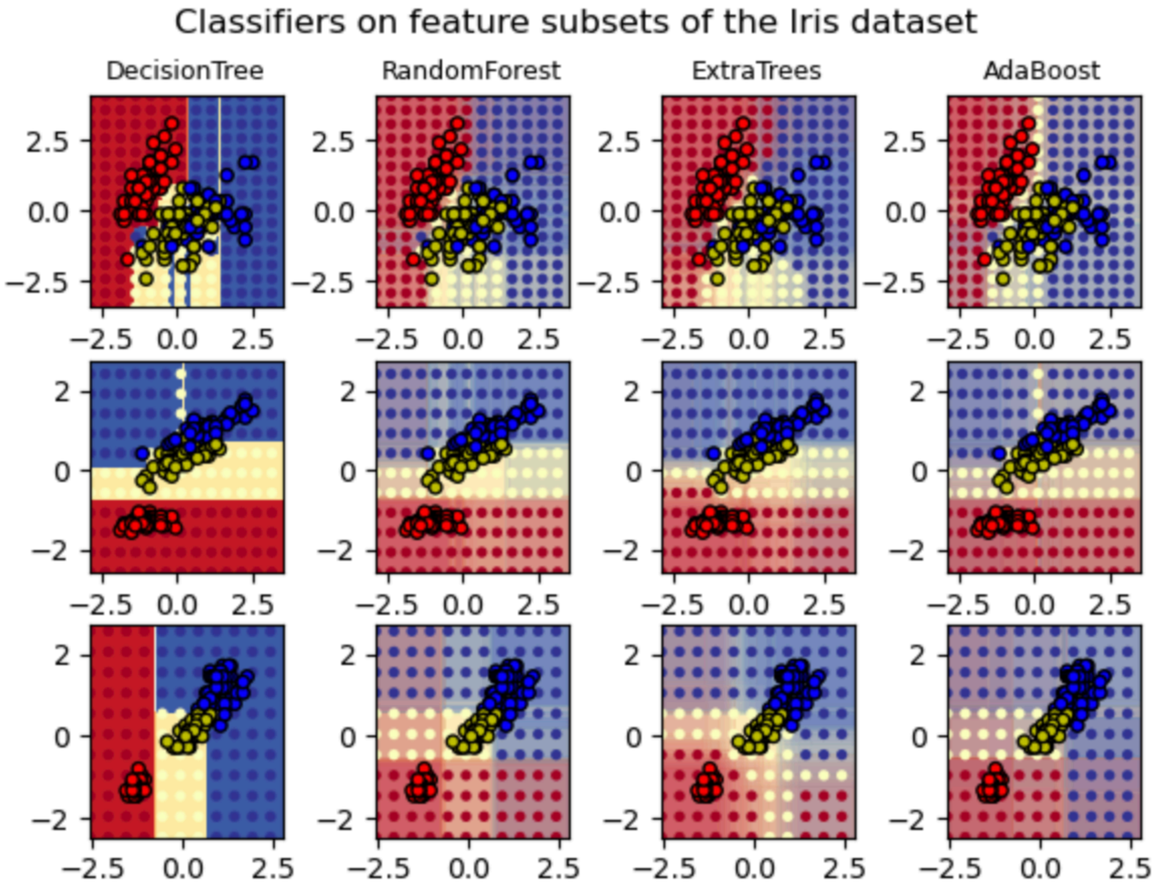
Bagging


Decision Tree
Decision Tree from Bagging
x_1
x_2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
2.0
2.5

Boosting


Train a weak learner
Update weights of misclassified Observations
Train a weak learner
Update weights of misclassified Observations
Train a weak learner
...

Bagging and Boosting



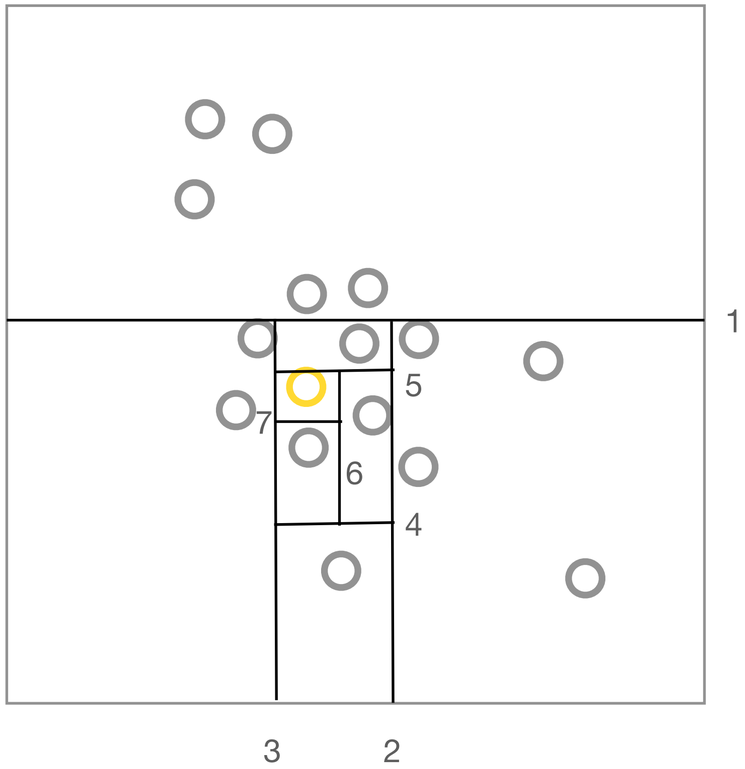
https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_iris.html#sphx-glr-auto-examples-ensemble-plot-forest-iris-py

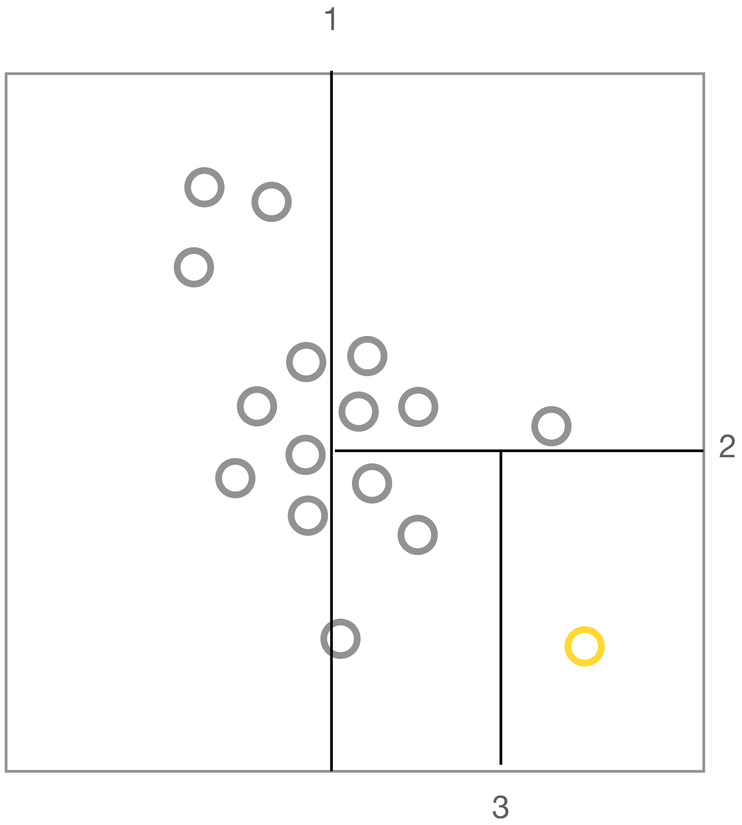
Isolation Forest

- Liu, Fei Tony, Kai Ming and Zhou, Zhi-Hua. *Isolation forest*, Data Mining, 2008. ICDM'08, Eighth IEEE International Conference on.
- Liu, Fei Tony, Kai Ming and Zhou, Zhi-Hua. *Isolation-based anomaly detection*, ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012): 3.

- Random tree based anomaly detection
- Bagging ensemble methods
- High performance than traditional methods based on classification and clustering
- Scale up to handle high dimensional problems with a large number of irrelevant attributes

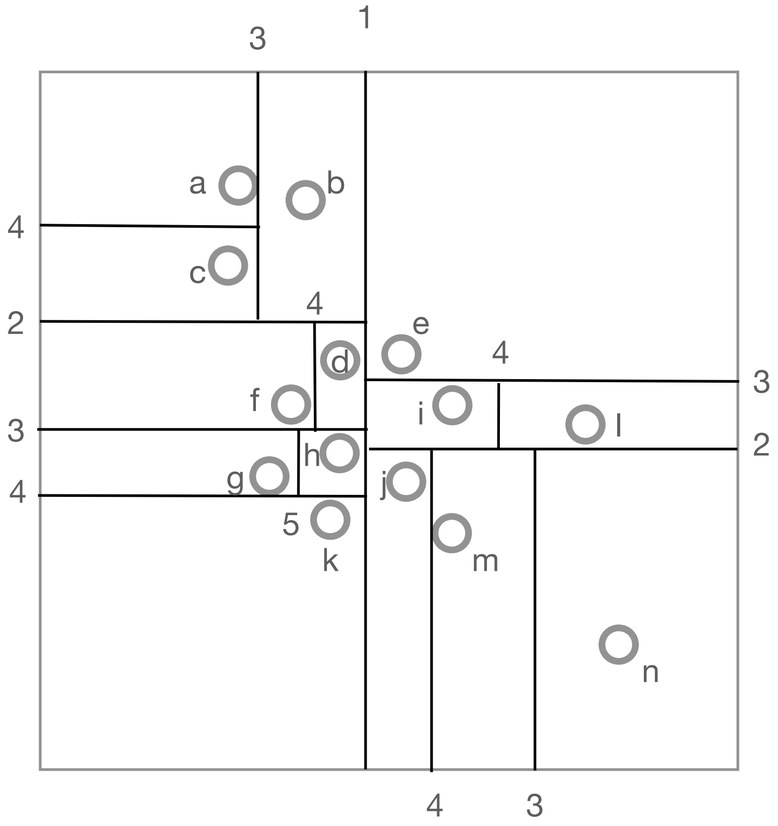
Isolation Forest (idea)

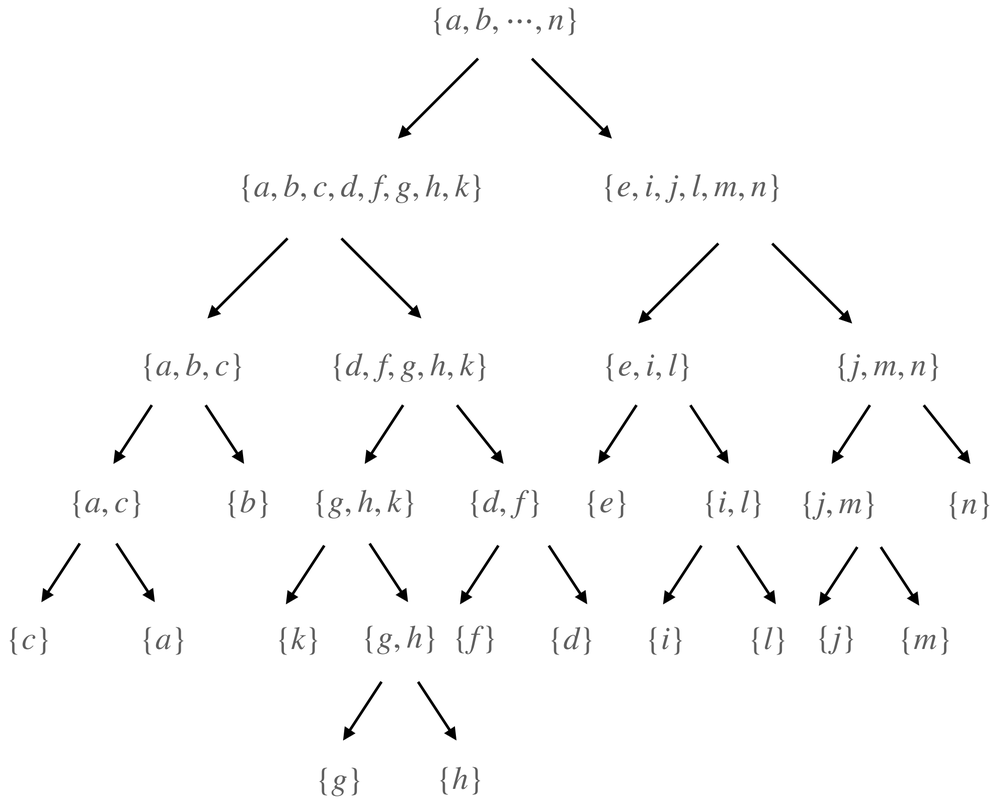




How many splits are needed to isolate the yellow points?

Isolation Forest (idea)





Notations

Goal : Detecting anomalous points

x : a data point

X : a data set of N instances

ψ : a subsampling size

$h(x)$: the path length of x

s : an anomaly score

Isolation Tree

Definition : Isolation Tree. Let T be an isolation tree. T is either an external-node with no child, or an internal node with one test and exactly two daughter nodes (T_l, T_r). A test consists of an attribute q and a split value p such that the test $q < p$ determines the traversal of a data point to either T_l or T_r

Path Length

Definition : Path Length $h(x)$ of a point x is measured by the number of edges x traverse an iTrees, from the root node until the traversal is terminated at an external node.

- Short path length means high susceptibility to isolation
- Long path length means low susceptibility to isolation

Anomaly Score

- Average path length of unsuccessful searches in BST

$$c(\psi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/\psi & \text{for } \psi > 2 \\ 1 & \text{for } \psi = 2 \\ 0 & \text{otherwise.} \end{cases}$$

- The anomaly score s of an instance x is defined as:

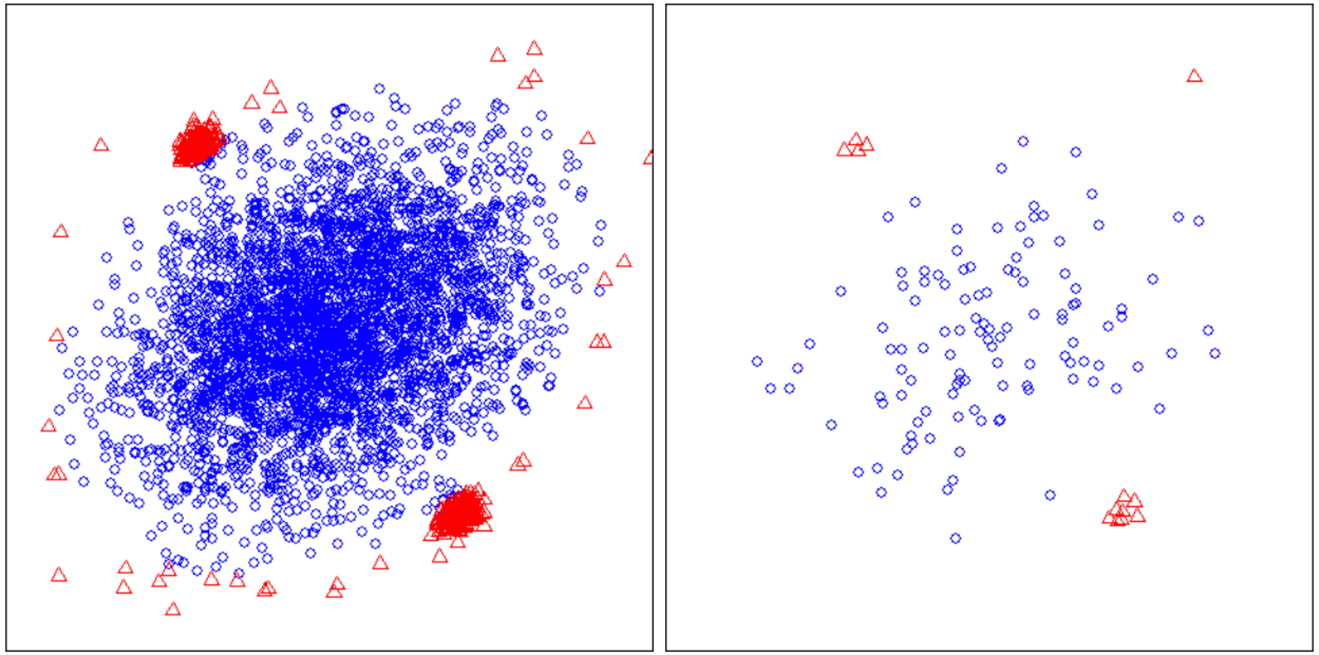
$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}}$$

(a) when $E(h(x)) \to 0,\ s \to 1;$

(b) when $E(h(x)) \to \psi - 1,\ s \to 0;$ and

(c) when $E(h(x)) \to c(\psi),\ s \to 0.5;$

Subsampling Effect



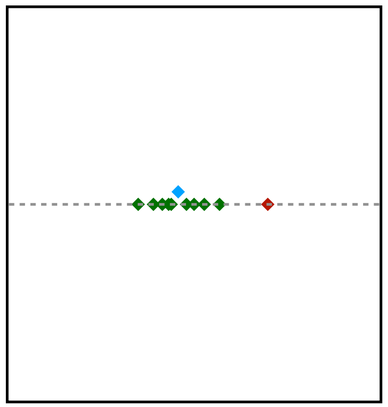
(a) Original sample
(4096 instances)

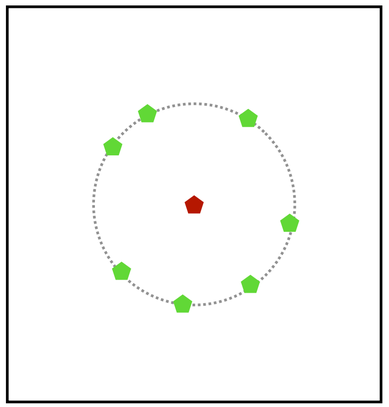
(b) Sub-sample
(128 instances)

- Small subsampling reduces the effect of masking and swamping

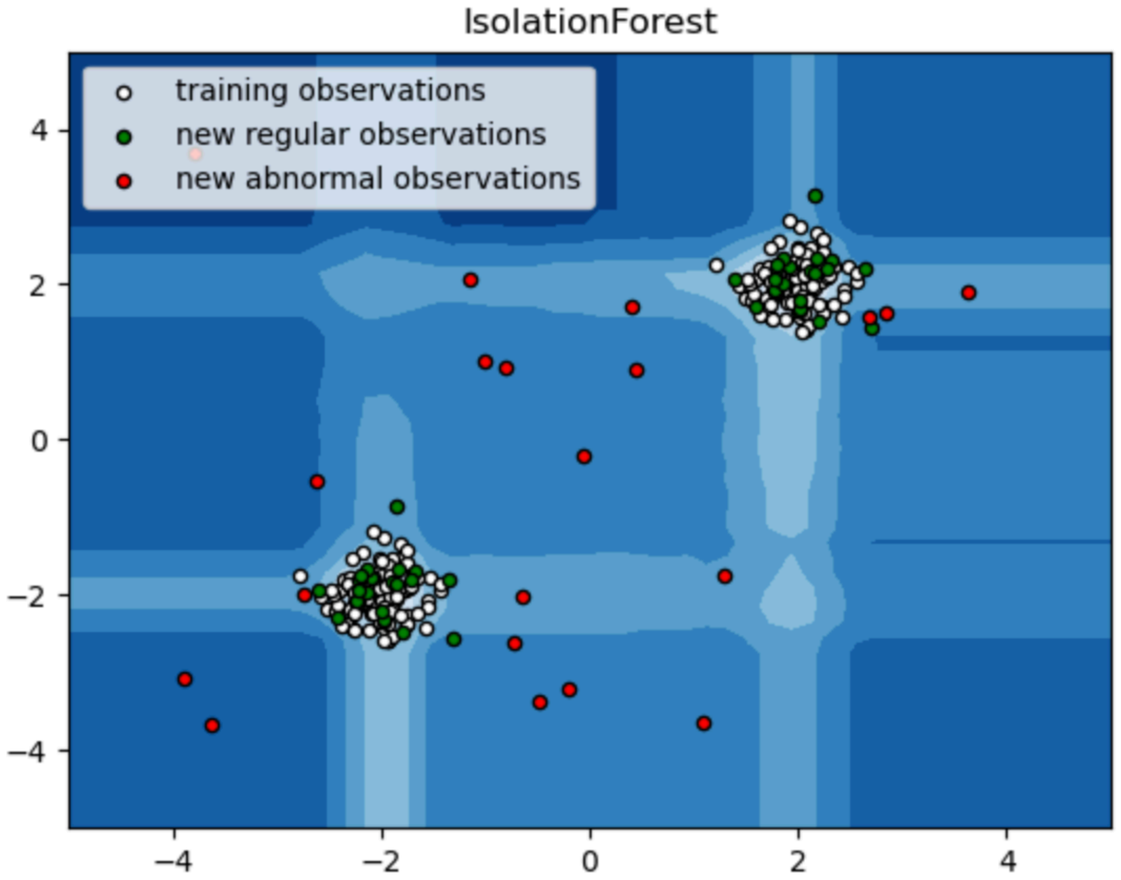
Weakness of Isolation Forest

1. For a fixed attribute, the majority of values are constant but only a few instances have slightly different values.
2. When a point surrounded by a closed hyper-sphere
3. More suitable for outlier detection than novelty detection





Weakness of Isolation Forest



https://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html#sphx-glr-auto-examples-ensemble-plot-isolation-forest-py