Importancia de características

Aprendizaje automático

Juan David Martínez
jdmartinev@eafit.edu.co

2023

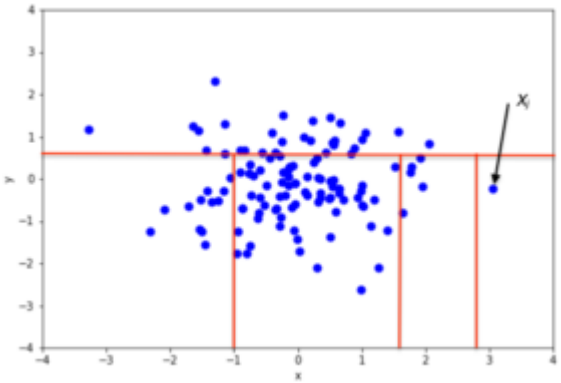
UNIVERSIDAD EAFIT

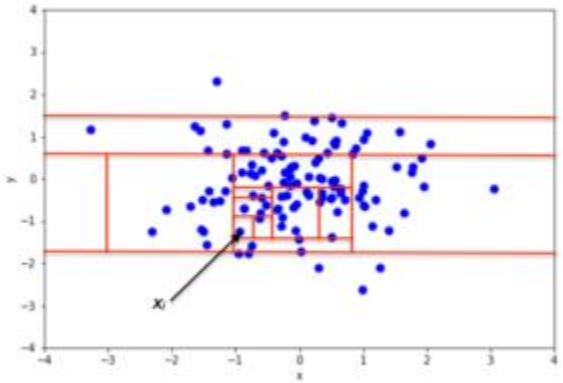
Agenda

- Isolation tres
- Isolation forests

Isolation trees

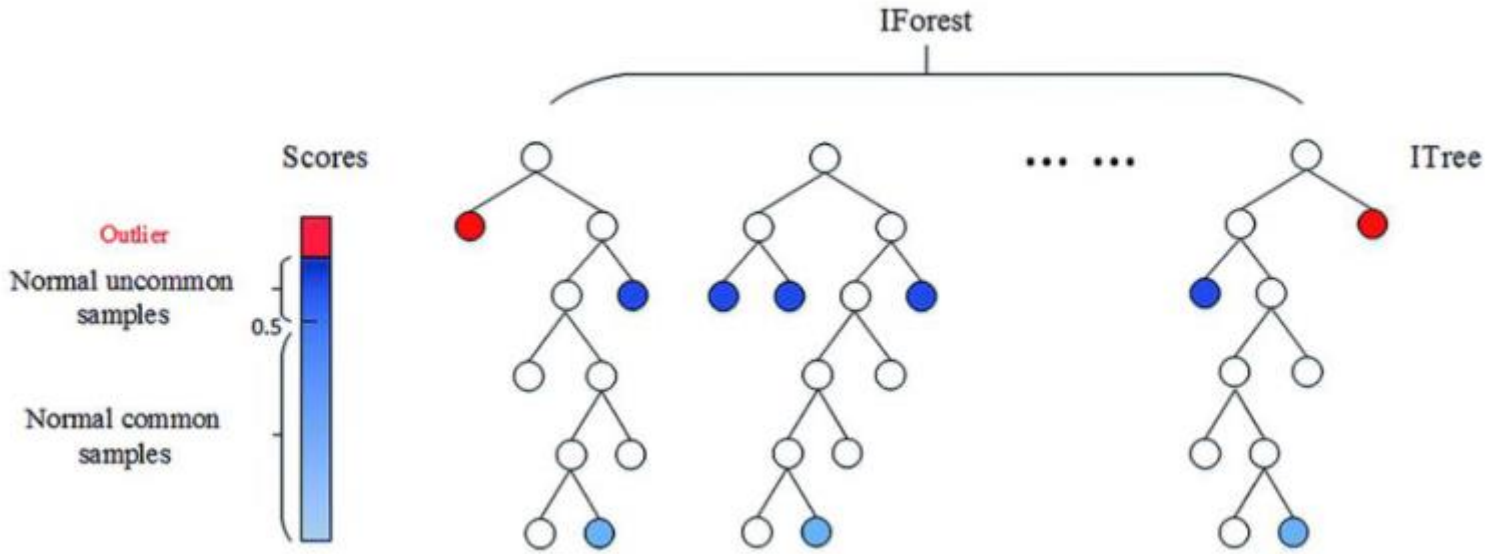
- Motivación:
  - Las anomalías generalmente son pocas muestras con características diferentes de las muestras “normales”
  - Podemos construir un árbol para aislar cada muestra
  - Las anomalías se agrupan cerca del nodo raíz
  - Las muestras normales se agrupan al final del árbol





Isolation forests

- La distancia promedio al nodo terminal se puede ver como un score de novedad de una muestra



Isolation forest

- **Isolation tree:** Dado con conjunto de muestras $\boldsymbol{X}$ de n muestras, este conjunto de datos se divide recursivamente seleccionando aleatoriamente una característica q con un valor de partición p, hasta que:
  - El árbol alcance una profundidad establecida
  - $|\boldsymbol{X}| = 1$
  - Todas las instancias de $\boldsymbol{X}$ tienen el mismo valor
- **Definición:** Longitud de trayectoria
  - La longitud de trayectoria $h(\boldsymbol{x}^{(i)})$ de una muestra $\boldsymbol{x}^{(i)}$ se mide por el número de nodos que se atraviesan en un árbol desde el nodo raíz al nodo terminal en el que se encuentra la muestra i
  - $h(\boldsymbol{x}^{(i)})$ se normaliza con la longitud de trayectoria promedio de $h(\boldsymbol{x}^{(i)})$ dado n

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}$$

$$H(i) = In(i) + 0{,}57721 \text{ (constante de Euler)}$$

Isolation forest

- **Definición:** score de novedad
  - El score de novedad s de una muestra $\boldsymbol{x}^{(i)}$ se define como:

$$\mathrm{s}\left(\boldsymbol{x}^{(i)}, \mathrm{n}\right) = 2^{-\frac{E(h(\boldsymbol{x}^{(i)}))}{c(n)}}$$

  - $E\left(h\left(\boldsymbol{x}^{(i)}\right)\right) \rightarrow c(n), s \rightarrow 0{,}5$
  - $E\left(h\left(\boldsymbol{x}^{(i)}\right)\right) \rightarrow 0, s \rightarrow 1$
  - $E\left(h\left(\boldsymbol{x}^{(i)}\right)\right) \rightarrow n-1, s \rightarrow 0$

Isolation forest

https://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html