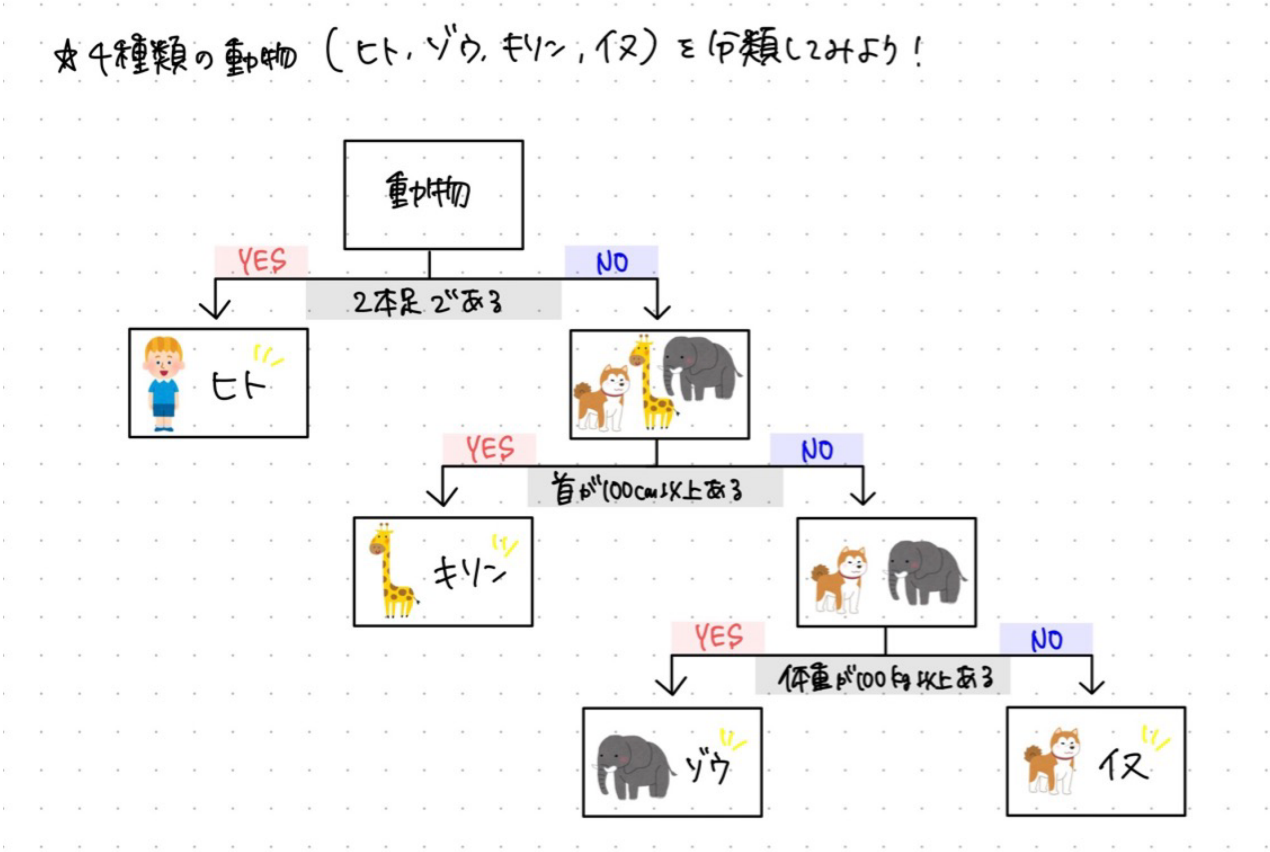
3章-2 決定木分析

決定木分析とは

決定木分析 は、データを分割するルールを次々と作成して分類を行うアルゴリズムです。

イメージ図です。



従属変数 に影響する 説明変数 を見つけ、樹木状のモデルを作成する分析方法となります。

従属変数：ヒト、ゾウ、キリン、イヌ

説明変数：2本足である、首が100cm以上ある、体重が100kg以上ある

決定木分析の特徴

1. 解釈がカンタン

   決定木分析 は上のイメージ図の通り、 処理過程が解釈しやすく、分析の妥当性を判断しやすいモデルです。

   このため、決定木分析は データマイニング でよく用いられます。
   データマイニング とはデータ解析の技法を大量のデータに網羅的に適用して傾向やパターンを分析し、新たな知識を取り出す技術のことです。

2. 分類木と回帰木がある

   決定木分析には、分類を行うモデルと回帰を行うモデルがあります。
   分類モデルでは DecisionTreeClassifier、回帰モデルでは DecisionTreeRegressor、といったクラスを使用します。
   今回は分類木のコードを解説していきます。

3. ハイパーパラメーターがある

   決定木のモデルはどうしても利用するデータに過剰適合してしまいがちなので、実践では汎化性能を向上させるために ハイパーパラメーター を設定したり他の種類のモデルと合わせて分析する必要があります。今回は決定木モデルのみを使って分析を行い、 ハイパーパラメーターとして木の深さを3に設定 しています。（このように、機械学習を行う前に人間があらかじめ設定しておくパラメーターのことをハイパーパラメーターと呼びます。）

4. 情報利得とジニ不純度

   実践において分類処理を厳密に行うときに重要なところですが、今回はコードの雰囲気を掴むことを優先して省略します。

   簡単に説明すると、決定木分類は最初いろんなものが混ざった ごちゃごちゃした集団 を条件ごとに分類していくわけですが、もし変な条件（本質とは関係ないような条件）で分類してしまうと 分類した後の方がごちゃごちゃでカオス になってしまうことがあります。

   それを防ぐために、 情報利得 と 不純度 という概念を使って 条件の妥当性を定量的に評価 します。