

(YouTube ピクアカ)

【はじめてのデータサイエンス PART2】

特徴量とは？

機械学習で大切な**特徴量設計**と**ディープラーニング**を知ろう

機械学習モデルの精度
= 機械学習プロジェクトの成果

決め手は

データの質、特徴量設計

データの重要性

データの入手方法は様々

- ・ 自社既存システム
- ・ オープンデータ
- ・ 販売データ
- ・ IoT、ウェブスクレイピングで新たに取得

構造化データ、非構造化データ

構造化データ：定型データ テーブル（行と列）

非構造化データ：形は不統一 数値以外のデータ

⇒ 非構造化データをうまく利用する技術が重要

（ネット、SNSで集まるデータは非構造化データ）

教師あり学習

学習データの収集以外に大変なことは？

特徴量

特徴量 feature とは

分析対象の**測定可能**なプロパティ
予測の手掛かりとなりうる変数

特徴量 feature とは

分析対象の**測定可能**なプロパティ
予測の手掛かりとなりうる変数

例①：物件条件から家賃の予測

面積、最寄り駅、築年数 → 手掛かりに**なる**

管理会社の連絡先 → 手掛かりに**ならない**

特徴量 feature とは

分析対象の**測定可能**なプロパティ
予測の手掛かりとなりうる変数

例②：個人データから保険契約する・しない

年収、既婚・未婚 → 手掛かりに**なる**

生年月日 → 年齢に変換すれば手掛かりに**なる**

身長 → 手掛かりに**ならない**

非構造化データの特徴量

画像認識：

特徴量 = 画像のピクセル

音声認識：

特徴量 = 音の波形の処理結果

自然言語処理：

特徴量 = 注目キーワードの出現頻度 等

未加工のデータから、理想的な特徴量は得にくい

インサイトをデータから
引き出すためには

予測に影響を及ぼす因子を
過不足なく含むデータセットを作ることが必要

特徴量設計のプロセスが重要

予測変数の**選別**、元データの**前処理**

データの前処理が必要なケース

カテゴリデータの処理

数値化（**コンピュータは数値しか処理できない**）

欠損値処理（**適切な値で補充**）

特徴量の変換・追加（**集計、カウント、結合、分割 等**）

特徴量設計は、職人技



そこで・・・

特徴量もコンピュータに抽出させる



ディープラーニング

AI、ML、NN、DLの関係

AI（人工知能）

ML（機械学習）

NN（ニューラルネットワーク）

DL
（ディープラーニング）

画像認識、音声認識、自然言語処理 など

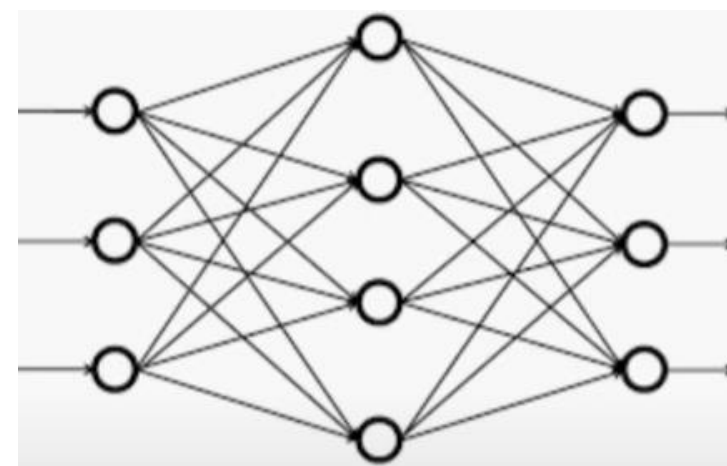
データ構造が複雑

特徴量設計が困難



ディープラーニングを活用

特徴量を自動抽出



図はニューラルネットワーク

ディープラーニングにおいても

データの前処理

(特徴量を見つけやすくする工夫) が**必要**

DLが**最強のML**ではないこと

課題、データの性質を理解した上で

分析手法を選択

汎用的なモデルは存在しない

NFLT ノーフリーランチ定理

no-free-lunch theorem

あらゆる問題に対応できる

汎用的モデルは存在しない

DLが有効なケース



直感的に捉えにくい非構造化データ

画像、音声、自然言語

十分なデータ量

予測精度・自動化のみ重視

※予測結果の理解は困難

シンプルですむ場合、あえてDLは使う意味はない

決定木を採用する場合

変数が明確な構造化データ（購買履歴 等）

判断結果の解釈が重要な場合

画像認識の最強手法

畳み込みニューラルネットワーク CNN

前段で学習した結果を下流に入力
層が進むに従い

より高度な特徴を学習可能

層が進むにつれ・・・

深い層：複雑・抽象的な特徴を得る
顔、目、鼻 など



中段層：やや高次な特徴を抽出 **エッジ、輪郭、形**



浅い層：単純、具体的な特徴を抽出
小領域の**明暗** 程度

<ここまで>