ゼロから学ぶAI #2

データの”いろは”

2022/11/28

CTUグループ

発表概要

・2022/11/07の発表 『ゼロから学ぶAI』では、**AI / 機械学習**といったトピックを扱い、
出来ることや計算方法について簡単に紹介させていただきました。

・今回の発表では機械学習に用いる「データ」に焦点をあててお話します。

※ 分かりやすさを重視するため、厳密でない表現が含まれます

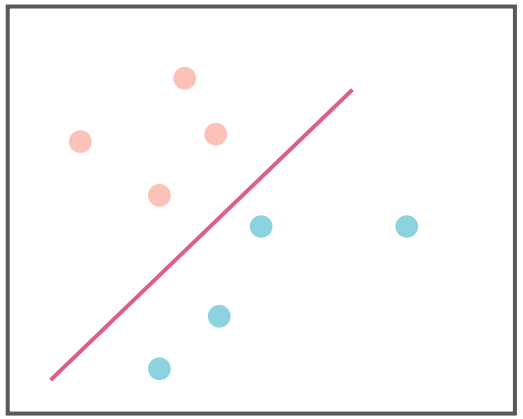
おさらい

AI / 機械学習でやっていること

- 「問題と答えのセット（**教師データ**）を与える」

  ⇒　両者の**対応関係 / 特徴を見出す（モデル化）**　⇒　問題（答え不明）をモデルに掛ける

- 良い結果を得るためには、教師データの「**量**」と「**質**」が重要



分類

築年	徒歩	面積	家賃
1	2	20	13
5	10	30	10
15	10	25	8
3	6	80	?

回帰

データの量

例 : 消費者行動のデータ収集

・レシート買取アプリ『ONE』（2018.6〜）

・アカウントの情報（性別 / 年齢 / 職業 など）と
レシートの情報（買い物内容 / 日時 / 場所 など）を
組み合わせることで、分析用のデータとなる

・データの「量」が重要視されるため、価値が発生する

（引用 : https://apps.apple.com/jp/app/one-ワン-お買い物でお金がもらえるアプリ/id1373644984）

データの質

質の高いデータとは

・（機械学習 / データ分析用として）データの質を高める要素として、

「データの性質を考慮して**適切な処理がされている**」 「余計な変数を含まない」ことが挙げられる

・以下の買い物に関するデータの中で、上記の2要素に該当する箇所を探してみて下さい

（後ほど答え合わせをします）

性別	年齢 [歳]	属性	気温 [℃]	地域	時刻	風速 [m/s]	値段 [円]
1 (男性)	21	2 (大学生)	10	1 (北海道)	18:15	18	350
2 (女性)	18	1 (高校生)	21	13 (東京)	12:10	2	1000
1 (男性)	51	3 (社会人)	8	14 (神奈川)	9:45	9	800

コーヒー1杯に払ったお金

データの種類・性質

データには大きく分類して4つの種類がある

質的データ

名義尺度

・順序がないデータ
・単なるカテゴリ分け

ex)　性別, 血液型

順序尺度

・順序があるデータ
・数値っぽくても間隔が等しくない ⇒ 足し算できない

ex)　順位, 階級

量的データ

間隔尺度

・数値で間隔が等しい ⇒ 足し算できる
・”0” が ”無” ではない ⇒ 掛け算できない

ex)　セ氏温度, 時刻

比例尺度

・”0” が ”無” である ⇒ 掛け算できる

ex)　絶対温度, 値段, 時間

データの性質によって処理してみる①

性別, 地域 : 名義尺度なので×

・カテゴリの数だけ**ダミー変数**を作ることで対処する

性別
1 (男性)
2 (女性)
1 (男性)

男性	女性
1	0
0	1
1	0

属性：順序尺度として扱うなら△

・性別, 地域と同様にダミー変数を作るのが基本

・1 高校生 ⇒ 2 大学生 ⇒ 3 社会人 と所得が増える傾向にあるため、そのまま使う方法もある

※ 余談 「地域」についても最低賃金順に並べることで順序尺度として扱う方法があります

データの性質によって処理してみる②

気温：間隔尺度なので△

・比例尺度である絶対温度に変換するのが基本

気温 [℃]
10
21
8

気温 [K]
283
294
281

時刻：そもそも扱うのが難しい

・ある時点からの経過時間 [秒] にすることで、比例尺度には変換可能

・周期的なデータのため、扱いが難しい（より複雑な分析方法をとる必要がある）

例) 0時0分からの経過時間とすると

0時1分 ⇒ 60秒, 23時59分 ⇒ 86340秒 と扱うことになるが、実際には2分差しかない

過学習

余計な変数を含んだまま学習した場合...

- 「意味のない変数だと分析し、考慮しなくなる」場合は問題無いが、
  「頑張って考慮してしまう」場合（**過学習**）がある
- 与えた教師データの範囲に限って正答率が良いモデルを構築してしまう

風速を含めた結果, 精度が落ちることも...
（風が弱い日は高いコーヒーを買う など）

性別	年齢 [歳]	属性	気温 [℃]	地域	時刻	最大風速	値段 [円]
1 (男性)	21	2 (大学生)	10	1 (北海道)	18:15	18	350
2 (女性)	18	1 (高校生)	21	13 (東京)	12:10	2	1000
1 (男性)	51	3 (社会人)	8	14 (神奈川)	9:45	9	800

まとめ

- 機械学習 / データ分析 の際にはデータを適切に前処理しましょう

- 機械学習 / データ分析をしない方も
  データの性質を理解し、変な計算をしないようにしましょう

- 良いデータには価値があります

※商用・営利目的の資料ではなく、社内発表用の資料です。

※個人的な見解や解釈を含んでいる場合もございますがご容赦ください。