半教師あり学習

■ **半教師あり学習**とは
- 教師あり学習と教師なし学習の間の存在
- 一部のデータのみに正解ラベルを付与して、それを用いて残りの正解無しのデータを活用するように学習を行う

■ メリット
- 学習データに正解ラベルを付けるコストを低減できる

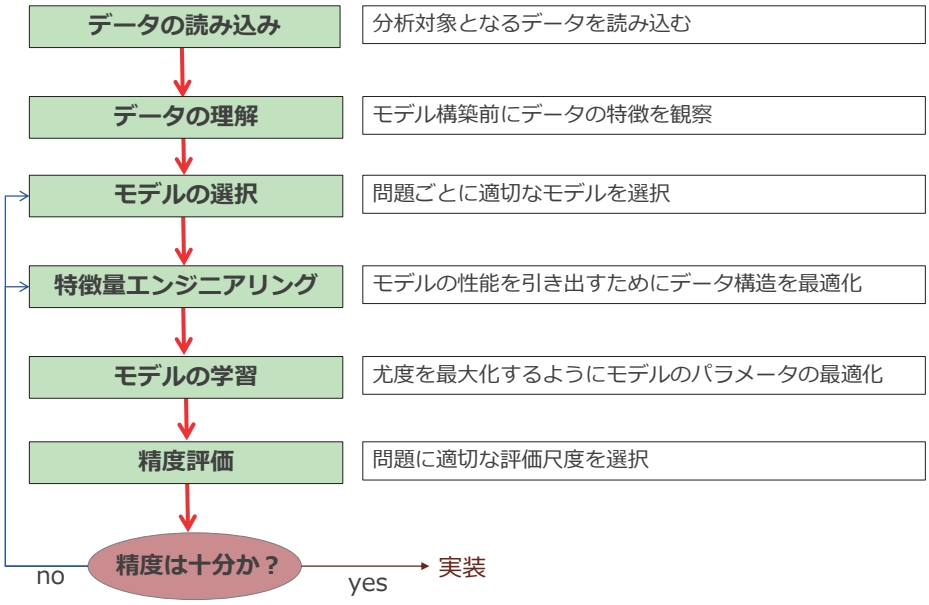
■ デメリット
- 教師あり学習に比べて精度が低いことがある
- 正解ラベルをつけたデータに偏りがあると精度低下の原因になる

次に、AI・機械学習の実装にとって、重要な概念である
「特徴量エンジニアリング」について説明する

【キーワード】

特徴量、特徴量設計、特徴量エンジニアリング
構造化データ、非構造化データ
データの前処理
カテゴリカル変数
ラベル・エンコーディング、ダミー変数
One-hot エンコーディング
欠損値

教師あり学習モデル構築の基本的なフロー



学習データの質 と **特徴量設計** は、
機械学習モデルの精度、つまり 機械学習プロジェクトの成果を大きく左右する

構造化データと非構造化データ

構造化データ（定型データ）

- **CSVファイル**
- **Excelファイル**
- **SQL型データベースのテーブル**

…

非構造化データ

- **テキストファイル**
- **XMLファイル**
- **画像ファイル**
- **音声ファイル**

…

- 使いやすいので分析に頻繁に登場
- 明確な「列」と「行」の概念
- 列構造で「どこに何があるか」が決まっているため、集計、演算、比較などがしやすい

- 統一的な列と行で整理されていない
- 数値データではないものもある

※「規則性のある非構造化データ」もある（HTMLのように情報を登録する方法が決まっており、原理的に100%取得可能）

インターネット・SNS等の普及により文章、音声、画像が大量に発信
➡ 非構造化データを上手く利用する技術も重要になった

（参考）取得可能なデータの例

- **自社のデータ**
  - 受注・発注・在庫データ
  - 人事データ
  - 営業活動データ
  - 顧客管理データ
  - メールやチャットのログ
- **コールセンター**
  - 電話対応の音声データ
  - FAQ参照ログ
- **ECシステム**
  - 商品説明・画像
  - 口コミ
  - 閲覧・購買履歴
- **データ販売企業**
  - 各種アンケート
  - 消費者パネルデータ
  - 小売店販売データ
  - ポイント利用データ
- **センター系**
  - 機器稼働センサーログ
  - 防犯カメラのデータ
  - 人流データ
  - 位置情報データ
  - その他、画像・動画・音声など
- **オープンデータ**
  - 気象データ
  - 経済統計データ
  - 人口動態データ
  - 映画データベース
  - 研究者が公開するデータ
- **Web API 利用**
  - 国立国会図書館
  - 楽天
  - ぐるなび
  - リクルート
  - Google
  - Facebook
- **ウェブコンテンツ**
  - SNS（Twitter、Instagramなど）
  - スクレイピング制限のない一般的なサイト

ここで、特に理解を深めたい概念は

特徴量

「特徴量」(英：feature) とは
分析対象の測定可能なプロパティ
予測の手掛りとなる変数

列の項目に相当

年齢	性別	年収	家族人数	勤務年数	…
45	1	700	2	2	…
32	0	450	4	3	…
:	:	:	:	:	:

【例①】物件条件から家賃の予測

面積	最寄駅(分)	築年数	管理会社Tel	正解
45	10	20	xxx	15
32	7	45	xxx	10

予測
14
9

面積,最寄駅,築年数
は特徴量として
効きそう

管理会社Tel
は効かない

【例②】個人データから保険契約するかを予測

年収	既婚	身長	生年月日	正解
450	0	176	YYMMDD	0
600	1	155	YYMMDD	1

予測
0
1

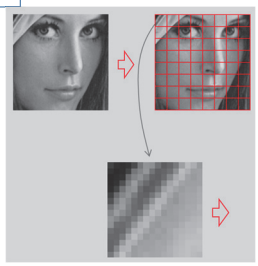
年収,既婚
は効きそう

身長は効かない

生年月日は現日付との差をとって**年齢**に変換すれば効きそう

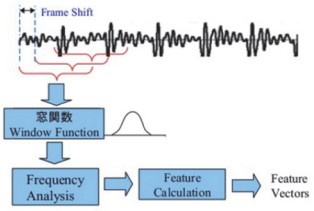
非構造化データの場合の特徴量

画像認識　特徴量 ＝ 画像のピクセル



音声認識

特徴量 ＝ 音波の波形を処理した結果



自然言語処理　特徴量 ＝ 注目文言の出現頻度

	単語１	単語２	単語３	…	単語M
文章１	4	8	0	…	2
文章２	2	0	1	…	6
文章３	7	0	8	…	4
文章４	3	4	3	…	1
⋮	⋮	⋮	⋮	…	⋮
文章N	0	2	5	…	6

生データのままでは、必ずしも理想な特徴量を得られない

予測に影響を及ぼす因子を過不足なく含むデータを作りたい

特徴量設計（特徴量エンジニアリング）のプロセスが重要

- ✓ 予測変数として採用する列を選別
- ✓ 元データに**前処理**を施す

前処理が必要なケース

カテゴリカルデータの処理

コンピュータは数値しか処理できない為、
文字列データを数値に変換してから機械学習モデルに入力

初回/リピート
初回
3回目
5回以上
…

「初回」なら1
それ以外は0に変換

初回フラグ
1
0
0
…

欠損値処理

欠損値(歯抜け)の多い場合は良い精度を期待できない ➡ 適切な値で補充

性別	年齢	身長	体重
男	35	170	
男		165	60
		159	
男	12	155	40
男		165	62
女		145	35

特徴量の変換・追加

予測に効きそうな情報をより効果的な形に編集し、新しい価値を持たせる

- 集計、カウント、データを結合・分割など

会員ID	年月日	購入個数
1	2019/10/11	2
2	2019/10/12	1
1	2019/10/12	1
3	2019/10/12	2
2	2019/10/13	1

集計

会員ID	総数
1	3
2	2
3	2

特徴量エンジニアリングの例① 数値へ変換

数値ではない**カテゴリカルデータ**を、コンピュータが処理できる形に変える

生データ（イベント参加の履歴）				
年月日	初回/リピート	評価	誰と？	…
2019/10/01	初回	○	友人	…
2019/10/12	3回目	△	一人	…
2019/10/13	5回以上	×	母、従姉妹	…
…	…	…	…	…

「初回」なら1
それ以外は0に変換

○なら1、
△なら2、
×なら3に変換

複数人参加なら1、
1人参加なら0に変換

変換後のデータ				
年月日	初回フラグ	評価	仲間フラグ	…
2019/10/01	**1**	**1**	**1**	…
2019/10/12	**0**	**2**	**0**	…
2019/10/13	**0**	**3**	**1**	…
…	…	…	…	…

カテゴリカルデータを数値化する上で、代表的なのは以下の２つ

One-hot encoding

- カテゴリ毎に列を作り、１行毎に、**１項目だけを1、それ以外を0にする**
- 新しい列がどんどん作られていくスパースな行列になるので、メモリはさほど消費しないが、列があまりにも増えると扱いにくい場合も

季節
春
秋
秋
夏
冬

⇒

春_FLG	夏_FLG	秋_FLG	冬_FLG
1	0	0	0
0	0	1	0
0	0	1	0
0	1	0	0
0	0	0	1

Label encoding

- １つのカテゴリが１つの数値に対応するように、数字に置き換える
- 列が増えない
- 手法は「マッピング」とも呼ぶ

季節
春
秋
秋
夏
冬

⇒

季節
1
3
3
2
4

特徴量エンジニアリングの例② 欠損処理

欠損値をどう処理すべきかは、欠損値の割合とデータ全体の状態によって決める

リストワイズ法

- 欠損値があるサンプルをそのまま削除してしまう方法
- 総データ量が多い場合、あるいは欠損が少ない場合に便利
- ただ、欠損に偏りがあると、削除した場合、データ全体の傾向を変えてしまう

回帰補完

- 欠損しているある特徴量と他の特徴量の間に相関が強い場合、回帰を利用して欠損値を埋める
- つまり、非欠損部分を利用して予測モデルを作り、欠損値を補充する値を推測
  具体例は次ページ

統計量で補完

- 欠損がある列について、欠損担っていないサンプルの平均値、中央値、最頻値などを算出し、欠損部分をその値で一律に埋める

回帰補完の例

元のデータの「年齢」
列に歯抜けが多い

元の会員属性データ

ID	性別	年齢	身長	体重	独身フラグ	子供の数	…
1	男	35	170	65	0	2	…
2	男		165	60	1	0	…
3	女		159	50	1	0	…
4	男	12	155	40	1	0	…
5	男		165	62	0	3	
6	女		145	35	0	1	
7	女	32	157	42	0	0	
…		…	…				…

歯抜け以外の部分の属性を学習データにする

身長**x**cm体重**y**kgの子供**z**人を持つ男性は**w**歳と思われる

性別	年齢	身長	体重	独身フラグ	子供の数	…
男	35	170	65	0	2	…
男	12	155	40	1	0	…
女	32	157	42	0	0	
	…	…				…

機械学習モデル

歯抜け部分に予測モデルを適用後

ID	性別	年齢	身長	体重	独身フラグ	子供の数	…
2	男	25	165	60	1	0	…
3	女	30	159	50	1	0	…
5	男	45	165	62	0	3	
6	女	35	145	35	0	1	

予測結果

機械学習モデルの具体的な手法と応用例