

第1章 機械学習の概要

機械学習が活躍している場面

機械学習はIT、マーケティング、産業、医療など様々な分野で利用されています。具体例を挙げてみましょう。

- 医療診断：患者の検査の情報が与えられたとき、病気の有無の判定を出力する。
- ニュース記事のタグ付け：ニュース記事の文章が与えられたとき、その記事に対して適切なタグを出力する。
- 機械翻訳：日本語の文が与えられたとき、元の文と同じ意味の英語の文を出力する。
- 商品のレコメンド：顧客の情報が与えられたとき、その顧客と似た顧客が購入した商品を出力する。
- 異音検知：機械の動作音が与えられたとき、異常があるかどうかの判定を出力する。

ここで挙げた例は、ある入力を与えられたとき、適切な出力を返すことを目標にしているという共通点があります。

問 この目標を達成するための手段として人間による判断、ルールベースシステム、機械学習が考えられます。それぞれの違いを説明しましょう。

|
|
|
|
|
|

問 他に、「顧客のサービス解約予測」や「スパムメールの検出」なども機械学習が利用されているタスクの例です。入力と出力はどうなるか考えてみましょう。

|
|
|
|

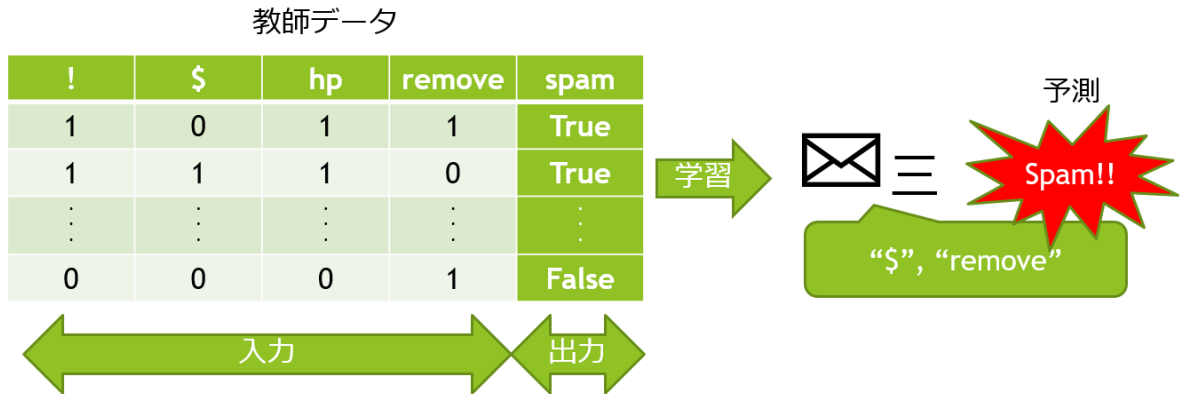
機械学習のフレームワーク

教師あり学習

入力と、入力に対する望ましい出力からなるデータが用意されているとします。教師あり学習とはこのようなデータに基き入力と出力の傾向を学習し、未知のデータを予測できるようにする機械学習の問題設定のことをいいます。

出力が質的変数のとき分類問題、出力が量的変数のとき回帰問題とといいます。

たとえばメールに現れる単語のパターンからそのメールがスパムメールかどうかを予測する問題は分類問題です。



問 ワンルームマンションの賃料を予測するモデルを構築したいとしましょう。

1. この問題は分類問題でしょうか回帰問題でしょうか。
2. 入力変数・出力変数として適切なものを考えてみてください。

|
|
|
|
|
|
|
|
|

教師なし学習

教師なし学習はデータの悪れた構造を見つけ、まとめるような手法のことです。教師あり学習とは異なり、データの中に出力はありません。例えば次のような手法は教師なし学習です。

- 次元削減

複数の変数からなるデータを何らかの目的で少ない変数で表現しなおす機械学習アルゴリズムの総称です。例えば、似ている変数をひとまとめにしてノイズの影響を減らしたり、データの解釈性を向上することを目指す主成分分析があります。

	温度1	温度2	内圧1	内圧2		温度	内圧	
4/1	64	66	12	8	➔	4/1	65	10
4/2	69	71	10	10		4/2	70	10
4/3	72	68	9	11		4/3	70	10
4/4	42	38	24	16		4/4	40	20
4/5	43	47	18	22		4/5	45	20
4/6	44	46	19	21		4/6	45	20

- クラスタリング

類似したデータポイントをまとめる機械学習アルゴリズムの総称です。例えば、k-means法や階層クラスタリングといった手法が基本的です。

	温度1	温度2	内圧1	内圧2	
4/1	64	66	12	8	} クラスタ1：高温低圧
4/2	69	71	10	10	
4/3	72	68	9	11	
4/4	42	38	24	16	} クラスタ2：低温高圧
4/5	43	47	18	22	
4/6	44	46	19	21	

他にも典型的な問題設定として強化学習があります。目的や得られるデータの形式によって手法を使い分ける必要があります。