2.2　汎化、過剰適応、適合不足

15t4034s
荘司　響之介

汎化とは

訓練データに基づいて構築したモデルを用いて、（訓練データと同じ性質を持つ）未見のデータに対して正確な予想ができるようにすること。

汎化がうまくいかない例

年齢	自動車保有数	持ち家か	子供の数	婚姻状況	犬を飼っている	ボートを購入
66	1	yes	2	未亡人	no	yes
52	2	yes	3	結婚	no	yes
22	0	no	0	結婚	yes	no
25	1	no	1	独身	no	no
44	0	no	2	離婚	yes	no
39	1	yes	2	結婚	yes	no
26	1	no	2	独身	no	no
40	3	yes	1	結婚	yes	no
53	2	yes	2	離婚	no	yes
64	2	yes	3	離婚	no	no
58	2	yes	2	結婚	yes	yes
33	1	no	1	独身	no	no

上のデータを用いて、顧客がボートを購入するかどうかを予測

「45歳より年上で、子供の数が3人より少ないか、もしくは離婚していない顧客はボートを買う」
⇒複雑すぎるかつデータ数が少ないため、新しい顧客に対して適応できるとは考えにくい。

「家を持っている顧客はボートを買う」
⇒単純すぎるため、左のデータに対してすらうまく機能していない。（家を持っているがボートを購入していない顧客もいる）

過剰適合

複雑にしすぎたことが原因で、訓練データに対してはうまく機能するが、新しいデータに対しては汎化できないモデルを作ってしまうこと。

例）「45歳より年上で、子供の数が3人より少ないか、もしくは離婚していない顧客はボートを買う」

適合不足

単純にしすぎたことが原因で、訓練データに対してすらうまく機能しないモデルを作ってしまうこと。

例）「家を持っている顧客はボートを買う」

2.2.1 モデルの複雑さとデータセットの大きさ

訓練データがバリエーションに富んでいれば、過剰適合を起こすことなく、より複雑なモデルを利用することができる。

例）

12人分の顧客データに対して

「45歳より年上で、子供の数が3人より少ないか、もしくは離婚していない顧客はボートを買う」が適用できる

⇒過剰適合

12人分の顧客データ＋10,000人分の顧客データに対して

「45歳より年上で、子供の数が3人より少ないか、もしくは離婚していない顧客はボートを買う」が適用できる

⇒信用できる良ルール

まとめ

- モデルを構築する場合、複雑にしすぎても、単純にしすぎてもいけない
- より多くのデータを用いて、適度に複雑なモデルを作成することでうまくいく。